J-035

Web ニュース記事閲覧時の視線情報を利用した興味語の抽出
Interest extraction using gaze behavior during web news article reading

庄林 祐太† 藤江 真也‡
Yuta Shobayashi Shinya Fujie

1 はじめに

Web ニュース記事の閲覧行動，特に記事を読む際の視線の動きから暗黙的にユーザの興味語を抽出する手法を検討する．

現在，PC やスマートフォンで用いて，インターネット上から情報収集を行うことができる．そして，インターネット上には多くの情報がアップロードされている．しかし，膨大な情報の中からユーザにとって興味を持つ情報が見つけるのは困難である．そこで，通販サイト等でよく見る「おすすめの商品」のような情報を提示する情報推薦システムがある．ユーザにとって有用な情報を推薦するためには，ユーザの興味を収集する必要がある．ユーザの興味を収集する方法の 1 つとして，コンテンツを長く閲覧するなどのユーザの行動からコンテンツへの評価を予測する暗黙的な方法がある．

ユーザの検索意図を推定するために，ユーザの行動の 1 つである Web サイト閲覧時の視線情報を用いて注目語を抽出する手法が提案されている [1]．単語に対する注目頻度と出現頻度を組み合わせることでユーザの意図に沿った単語を抽出している．視線は，ユーザが気になったものや対象となるものを視覚的に得るために向けているものであり，視線を用いることで記事中のどこを見ているかを収集することができる．本研究では，Web ニュース記事を対象に，視線情報から記事内の興味語の抽出を目的とする．記事内の言語情報に加え，記事閲覧時の視線の動きや記事中の各単語への注視を利用し，機械学習を用いて，記事中の各単語への興味の有無を判定する．

2 興味語抽出

2.1 興味語抽出器

閲覧したニュース記事内に含まれる単語の言語情報と，その単語に対する閲覧時の視線情報に関する特徴量を抽出し，系列特徴量として扱った興味語抽出手法について述べる．本研究では，ニュース記事閲覧時の視線行動から，ユーザが興味を持つ（持った）単語を興味語と定義し，抽出することを目的としている．ニュース記事内に含まれる単語毎に，その単語への興味の有無を判定し，興味があると判定した単語を興味語として抽出する．抽出器としては，ニューラル・ネットワーク，特に，系列情報の前後関係を扱うことができる Bidirectional LSTM を用いる．言語情報の特徴量と視線情報の特徴量を入力層，800 素子の Bidirectional LSTM を中間層とし，各単語への興味の有無を判定する 2 素子の出力層を持つネットワークになっている．

2.2 データ収集

興味語抽出器に入力するデータを収集する．閲覧したニュース記事内に含まれる単語の言語情報と，その単語に対する閲覧時の視線情報を取得する．視線情報の検出に Tobii Technology 社が開発した Tobii Eye Tracker 4C[2] という視線検出装置を使用する．本装置は，PC 画面のフレームの下側につけて使用する．視線情報として，画面上の視線位置，装置に対する眼球位置を取得することができる．記事中の単語に視線情報を対応づけるために，特殊な Web ブラウザを作成した．ユーザに対しては通常の Web ブラウザと同様の操作で Web ページを閲覧できる機能を与えたうえで，記事に対する形態素解析を前処理として行うことで，単語ごとに HTML 要素を作成した．これにより，記事に対する視線位置に応じた記事中の単語を収集できる．

収集では，1 日 10 記事閲覧することとし，被験者 11 人から 219 記事 (複数ページの記事も 1 記事とカウント) を収集した．ニュースサイトから，気になった記事を選択し，自由に閲覧させることとした．各記事閲覧後にアンケートを実施し，興味を持った（持っている）単語を興味語として収集した．収集したデータを，図 1〜3 で示す．記事に沿った単語位置とその記事を閲覧した際の視線行動を示した．記事を左から右へ，上から下へ文章を読んでいる様子がわかる．また，図 1 から興味語，非興味語のそれぞれの周辺に拡大して図 2,3 で示す．単語周辺の視線行動の様子を確認できる．

2.3 言語情報

ニュース記事には，多くの単語が含まれている．そのため，多くの単語を網羅している日本語版 Wikipedia データベース [3] を用いて，言語情報に関する特徴量を算出する．単語の重要度を評価する手法として，tf-idf 法が知られている．記事中の単語列において，tf-idf 値が高いほど，その記事においてその単語が重要な単語である．興味のある記事を閲覧している上では，tf-idf 値が高い単語に興味を持つ可能性がある．一般的に，動詞や助詞，副詞が文章中に多く含まれており，tf-idf 値が高くなる傾向にある．また，助詞や動詞が興味語となることはない．このことから，名詞のみを対象に，記事内の単語の重要度を算出した．そのため，tf-idf 値が算出されていない助詞や動詞，副詞の tf-idf 値は，「0」としている．Wikipedia データベースに収録されていないが品詞が名詞である単語に関しても，一般的ではない点では，重要な単語の可能性があるが，出現頻度が低いことが予想されるため，tf-idf 値を「0」とした．また，記事内の文章に含まれる単語周辺には，その単語にまつわる説明がなされているため，単語の意味を考慮する必要がある．そこで，単語自体は，Wikipedia データベースから学習させた Word2vec モデルを利用することで，単語の意味情報を持つ 200 次元の単語ベクトルで表現した．単語ベクトルに関しても，Wikipedia データベースに収録されていない単語に対する単語ベクトルは，ゼロベクトルで表現した．

† 千葉工業大学工学研究科未来ロボティクス専攻
‡ 千葉工業大学先進工学部未来ロボティクス学科

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

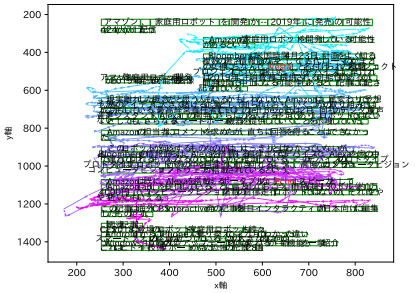

図 1　収録データ，記事中の単語位置と視線行動（閲覧時間につれて色を変化）

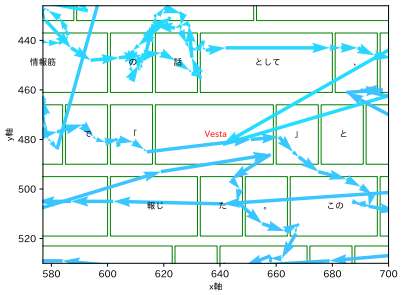

図 2　興味語周辺の視線行動，興味語を赤色で示す

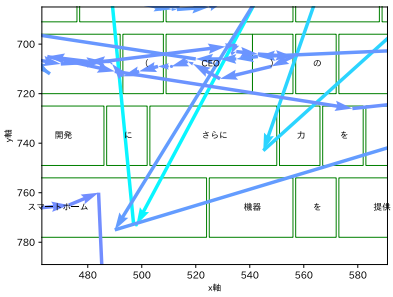

図 3　非興味語周辺の視線行動

2.4　視線情報

興味語の推定では，記事内の単語ごとに興味の有無を判定するので，記事内の各単語に対する閲覧時の視線情報に関する特徴量を抽出する必要がある．ユーザの意図に沿った単語抽出には，注目頻度が有効である [1]．そこで，単語に対して「どれくらいの時間見ているのか」という滞留時間，「何回見直しているのか」という注視回数を特徴量として加える．これによって，じっと見ている単語や滞留時間は短いが注視回数が多く，何回も見直している単語を考慮できる．また，記事によって文章の長さが違うことから，単に閲覧時間が長いだけでは，文章が長いもの記事の単語には，興味を持つことになる．そこで，滞留時間はその記事への閲覧時間で商をとり，注視回数はその記事における単語の注視回数の最大値で商をとる正規化を行った．

また，図 1〜3 より見直す動作が存在し，記事上の単語周辺に，その単語の説明や補足がある．また，興味語周辺では，興味語自体だけではなく，興味語に関係する説明や補足をよく見ていることが予想される．そのため，記事を「どのように見ているのか」を考慮する必要がある．そこで，各単語に対して，視線がどの方向からその単語に移動してきたのか，その単語から離れたのかを，回数で示す上下左右（4 方向）×入出力（ 2 方向）の 8 次元の視線方向ヒストグラムを算出し，「どのように見ているのか」を考慮する．視線方向ヒストグラムについても，その記事でのヒストグラム内の最大値で商をとる正規化を行った．

3　比較実験

3.1　実験条件

提案した興味語抽出手法の評価を，収録したデータに対して行う．

収録した全データをテストデータ，学習データそれぞれに用いた「データクローズ (Data Close; DC)」と，ある被験者が閲覧した記事のデータをテストデータとして，それ以外の人のデータを学習データにした「人オープン (Person Open; PO)」のそれぞれについて，視線情報に関する特徴量の有無を変化させて比較する．

3.2　結果と考察

DC 下での結果を被験者ごとに表したものを図 4 で示す．各被験者に対して，視線情報に関する特徴量の有無の 2 つの条件での Recall に加え，いずれの条件においても抽出できた興味語の割合を 3 つ目の条件「両方」として示した．

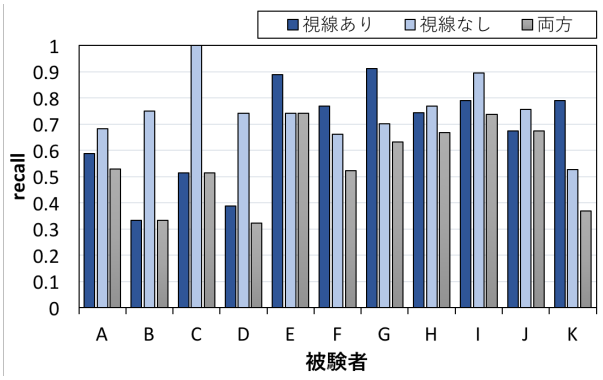

図 4　DC 下の被験者ごとの Recall

DC 下では，視線情報の有無によって，Recall に差が見られた．「両方」に対して，「視線情報あり」での Recall が高い値をとる被験者がいることから，視線情報を利用することで，言語情報のみでは抽出することができなかった興味語を抽出できることがわかった．このことから，興味語抽出に視線情報が有効だと考えられる．

PO 下の条件では，視線情報の有無に関わらず，適切に興味語を抽出することができなかった．視線情報を特徴量として用いることで，記事の内容によらず興味語の抽出ができることを期待したが，視線情報（すなわち記事の読み方）にもある程度の個人性があり，人オープンの条件では適切に学習が行うことが難しかったことが考えられる．また，今回利用した視線情報が興味語を抽出するのに必要な情報を捉えられていない可能性がある．特徴量の再検討は今後の課題としたい．

4　おわりに

情報推薦のための視線情報を用いたユーザの興味語の抽出手法について検討した．今後の課題としては，特徴量の検討と抽出器のチューニングが挙げられる．日常的な記事閲覧情報を蓄積して提案手法を評価し，個人の興味語推定をするために必要な学習データ量（どの程度の数の記事が必要なのか）を確認したい．

参考文献

[1]　梅本和俊, et al. 視線情報からの注目語抽出に基づく検索意図のリアルタイム推定. 情報処理学会論文誌データベース (TOD), 2013, 6.3: 120-131.

[2]　Tobii Technology Inc, Tobii Eye Tracker 4 http://tobiigaming.com/eye-tracker-4c/

[3]　日本語版 Wikipedia データベース https://dumps.wikimedia.org/jawiki/

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.