

注視特徴を用いたSVMによる非母語読解時の未知語検出

平岡 類 田中 宏季 サクティ サクリアニ 吉野 幸一郎 グラム ニュービッグ
中村 哲

† 奈良先端科学技術大学院大学 情報科学研究科

E-mail: †{rui-hi,hiroki-tan,ssakti,koichiro,neubig,s-nakamura}@is.naist.jp

あらまし 本稿では、アイトラッカから取得する視点情報を用いて、非母語読解時に文書内の未知語を検出する手法を提案する。従来手法では、眼球運動の一種である fixation (固視) に関連する特徴量と、大規模語彙コーパスから計算した語彙の希少度を用いて未知語の検出を行っているが、自然読解中の未知語検出においては検出精度が十分ではないという課題がある。そこで本稿では、自然読解中での未知語検出精度向上のために、(1) 機械学習の一手法、SVM [4] の適用、(2) 眼球運動に関する新たな特徴量の使用を提案する。実験的な評価により、従来の識別法と比較し、F 値において未知語検出精度の改善効果が得られることを示した。

キーワード アイトラッカ, 未知語検出, 眼球運動, 自然読解, 自動検出

Unknown Word Detection in Foreign Language based on Eye Gaze Features by SVM

Rui HIRAOKA, Hiroki TANAKA, Sakriani SAKTI, Koichiro YOSHINO, Graham NEUBIG, and Satoshi NAKAMURA

† Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)

E-mail: †{rui-hi,hiroki-tan,ssakti,koichiro,neubig,s-nakamura}@is.naist.jp

Abstract This paper proposes a method to detect unknown words during natural reading by using eye-tracking technique. A previous approach utilizes fixation and word rarity features to detect unknown words in foreign language. However, the performance of conventional approach is not sufficient during natural reading. To improve detection performance, we try to 1) apply support vector machine (SVM) [4], and 2) use new eye movement features that were not considered in the previous work. The experimental results demonstrate that SVM and proposed eye movement features are capable of improving detection performance as measured by F-measure.

Key words Eye tracking, Unknown word detection, Eye movement, Natural reading, Automatic detection

1. はじめに

国際化に伴って、非母語を使用する機会が多くなっている。例えば日本人の場合、共通言語として英語の普及が進み、最先端の学術論文は英語で書かれ、国際会議や論文誌で発表されることが多くなった。また書籍も英語で書かれることが一般的である。オンライン百科事典の Wikipedia の記事数も、英語は 500 万件あるのに対し、日本語では 100 万件に留まる [16]。非母語文献には母語文献よりも未知語の割合は多い傾向にあり、そのため非母語文献の読解時は翻訳ソフトウェアや辞書を駆使して読み進めていく必要がある。

しかし、翻訳ソフトウェアの精度は実務利用において満足度

があまり高くない [12]。また部分的に辞書を用いる作業にも時間が掛かってしまうという問題点があり、円滑な読解の妨げとなっている。そこで非母語習熟度の低いユーザーの非母語読解を補助するため、素早くかつ自動で未知語を検出し、翻訳や注釈をつけるシステムが考えられる。これにより、辞書を用いる際の手間が軽減され、円滑な非母語文献の読解が可能となる。このようなシステムの一例として、Hyrskykari の眼球運動に基づく未知語検出が挙げられる [1]。Hyrskykari の未知語検出法は視線の総滞留時間と語彙の希少度の 2 つの特徴量を用いている。さらにユーザーに対して訓練を行うことで、高い識別率 (再現率: 91%, 偽陽性: 2.4%) を達成している。しかし、ユーザーの訓練がない場合の検出精度は、再現率: 28.8%, 偽陽性: 0.9%に

留まる。本研究では、Hyrskykari の閾値による識別法 [1] を発展させ、SVM の適用と、用いていない眼球運動特徴量を用いることで、自然読解に最適化することを提案する。実験的評価によって、F 値での検出精度において、従来法と比較し改善効果が得られることを示す。

2. 関連研究

本研究は、ユーザーの英文読解時の眼球運動データを用いて、ユーザーごとの未知語を検出することを目標とする。未知語の検出に関連する研究として、江原らのクリックログから未知語を検出するシステムがある [9] [10]。この手法は過去のクリックログから自動で未知語を予測する。眼球運動と未知語の関係を示した研究として、ユーザー眼球運動と文章読解中の読み行動に関する関係を報告した Rayner の研究がある [7]。Rayner の実験結果は低頻出度語に対しては、高頻出度語よりも注視時間が長くなる傾向を示している。また視線の逆行数も多くなることが示されている。Kunze らは英文読解時の視線の動きから、ユーザーの英語力を推定する手法を提案している [6]。実験では、難易度が高い語彙に固視が多く発生しやすいことを示している。また Hyrskykari は、総注視時間からの未知語検出法を提案し、その手法を実装した視線から iDict を開発している [1]。iDict は上記の検出法を用いることで自動的に辞書にアクセスし、語彙に注釈をつけるシステムである。

本研究では、総注視時間以外の眼球運動特徴量の使用の検討と、識別に SVM を用いることで、未知語検出精度の向上を図る。

3. 総注視時間と語彙の希少度を用いた未知語検出法

本節で総注視時間と語彙の希少度を用いた、Hyrskykari の未知語検出法を説明する。一般に読解時の眼球運動の主な特徴は saccade (跳躍), fixation (固視), regression (逆行) の 3 つとされている [7]。fixation は定点を見るために眼球の動きを固定する眼球運動のことである。gaze (注視) は fixation よりもさらに意識的に視点を対象物に向け続ける眼球運動を表す。saccade は fixation 間に発生する跳躍運動であり、この間は認知活動がほとんど行われぬ。regression は、逆行運動を順方向とものと区別するために定義される。文書中の N 個の単語の群を順に $W = \{w_1, w_2, \dots, w_N\}$ と定義する。同じ単語が複数ある場合でも、異なる単語と見なす。同様に計測から得られる M 個の fixation の群を順に $F = \{f_1, f_2, \dots, f_M\}$ と定義する f_k は表示画面上での fixation の二次元座標を表す。

3.1 識別特徴量

本節で識別に使用された特徴量をまとめる。

gaze duration

単語 w_x 上で連続した fixation の滞留時間の和 (総注視時間) $g(w_x)$:

$$g(w_x) = \sum_{k=a}^{a+b} d(k). \quad (1)$$

such $w(a) = w(a+1) = \dots w(a+b) = x$

and $w(a-1)$ or $w(a+b) \neq x$.

ここで $d(k)$ は fixation f_k の時間長さである。 w_x 上での L 個の gaze の集合を $G(w_x) = \{g_1(w_x), g_2(w_x), \dots, g_L(w_x)\}$ と定義する。

total gaze duration

単語 w_x 上の fixation の滞留時間の総和 $g_{\text{total}}(w_x)$:

$$g_{\text{total}}(w_x) = \sum_{g_k(w_x) \in G(w_x)} g_k(w_x). \quad (2)$$

word rarity

単語 w_x のコーパス内の希少度 $Rare(w_x)$:

視覚刺激となる文章に対し、各単語の BNC コーパスにおける出現頻度の順位 (降順) を抽出する。この順位を語彙の希少度とするが、ここで単語出現頻度総数の寄与率を考慮する。上位 100 位の単語は、寄与率が 50% を超え、上位 6000 位までで 85% を超える。したがって頻度の順位を制限するために、上位 100 位以上の単語はすべて 100 とし、下位 6000 位以下のものについては、すべて 6000 とする。たとえば the ならばコーパス内の頻度が 1 位であるので 100 とし、incredibly であれば 6312 位なので 6000 とする。

$$Rare(w_x) = \begin{cases} 100 & (Rank(freq(w_k)) < 100) \\ 6000 & (Rank(freq(w_k)) > 6000) \\ Rank(freq(w_k)) & (\text{else}) \end{cases} \quad (3)$$

3.2 識別方法

本節で識別方法を示す。本手法では以下に示す閾値関数を用いて未知語検出を行っている。

$$th(w_x) = th_h - \frac{th_h - th_l}{100 - 6000} (Rare(w_x) - 100) \quad (4)$$

ここで $th(w_x)$ は単語 w_x が未知か既知かを判別する閾値である。 th_h は BNC コーパス内の頻出語上位 100 位以上の単語の未知と既知を識別する total gaze duration の閾値であり、 th_l は頻出語下位 6000 位以下の total gaze duration の単語の未知と既知を識別する total gaze duration の閾値である。 th_h, th_l は以下のように定義されている。

$$th_h = \mu_h + a_h \sigma_h, \text{ and } th_l = \mu_l + a_l \sigma_l$$

ここで μ_h は BNC コーパス内の頻出語上位 100 位以上の total gaze duration の平均であり、 σ_h は標準偏差を表す。 μ_l は頻出語下位 6000 位以下の total gaze duration の平均、 σ_l は標準偏差を表す。 a_h, a_l は、得られたデータの識別率が高くなるようにする学習パラメータである ($a_h > 0, a_l > 0$)。出力予測は以下ようになる。

$$\text{Pred}(w_x) = \begin{cases} \text{Known} & g_{\text{total}}(w_x) < th(w_x) \\ \text{Unknown} & g_{\text{total}}(w_x) \geq th(w_x) \end{cases} \quad (5)$$

本手法をベースラインとし、提案手法を用いた場合と比較する。

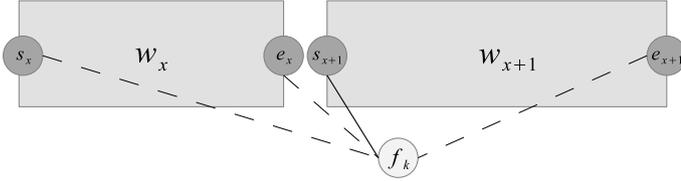


図1 fixationと単語の関連付け。 f_k との最短距離 (図中の実線) を与える点が s_{x+1} なので、 f_k が属する単語は w_{x+1} となる。したがってこの例では、 $w(k) = x + 1$ である。

4. 提案手法

4.1 提案手法

本節で提案手法の流れを示す。英文読解時の眼球運動データを用いて、未知語検出を行う手法を提案する。まずアイトラッカを用いて眼球運動データを計測する。計測器が出力した fixation-saccade のデータを元に、fixation を単語に関連付ける。単語ごとに言語特徴量と、関連付けた fixation を元にして眼球運動特徴量を抽出する。この特徴量を利用して未知語推定を行う。

4.2 眼球運動データの取得

ディスプレイ上に視覚刺激となる英語文献を被験者に読んでもらい、その間の眼球運動を計測する。計測には、非装着型のアイトラッカを使用する。非装着型を用いる理由としては、スクリーン上での文章読解となるので、頭部の動きがあまり大きくないことが挙げられる。また単語レベルでの解析をするため、装着型よりも高い精度を必要とするためである。

4.3 単語と fixation の関連付け

取得した眼球運動データと単語の座標を元に、注視時間や瞳孔の大きさの変化を未知語検出の特徴量として取得する。アイトラッカの計測データは時間当たりの視点を記録したものであり、fixation と saccade をそこから抽出する必要があるが、本研究ではアイトラッカの付属ソフトを使用することで、fixation と saccade を抽出することが出来る。

単語 w_x ごとの fixation の判定に対しては、単語 w_x の始点座標 s_x と終点座標 e_x を取得し、fixation f_k がどの単語に与えられたかは以下の式で定義する。

$$w(k) = \arg \min_x (D(s_x, f_k), D(e_x, f_k)) \quad (6)$$

ここで $w(k)$ は k 番目の fixation が何番目の単語に属するのかを定義する関数である。 $w(k) = x$ のとき k 番目の fixation は x 番目の単語に属することを指す。 $D(s_x, f_k)$ は fixation f_k と x 番目の単語 w_x の始点 s_x 、 $D(e_x, f_k)$ は終点 e_x のユークリッド距離を表す関数である。このいずれか一方の距離が最も小さくなる w_x に f_k は属する。図1に fixation 所属単語の判定の一例を示す。

4.4 特徴量抽出

本節では、読解課題における眼球運動特徴量および、言語特徴量の抽出法を示す。

4.4.1 眼球運動特徴量抽出

本節で眼球運動に関連する特徴量の抽出法を示す。瞬きや saccade と識別されるものを計測データから削除し、fixation を元に特徴量を作成する。fixation の座標と時間長さを抽出した後、fixation はどの単語に属するのかを判定する必要がある。本研究では単語ごとの fixation を元に、total gaze duration を含め、7つの特徴量を作成する。特徴量の定義は[1]を参考にした。max gaze duration と pupil size variation に関しては[1]で考慮されていないため、独自に定義した。

first gaze duration

単語 w_x 上の最初の gaze $g_{\text{first}}(w_x)$:

$$g_{\text{first}}(w_x) = g_1(w_x) \in G(w_x). \quad (7)$$

number of fixations

単語 w_x 上の fixation の総数 $N(w_x)$:

$$N(w_x) = |\{f_k \in F \mid w(k) = x\}|. \quad (8)$$

regression

単語 w_x 上での視線の逆行回数 $Reg(w_x)$: 単語 w_x に fixation が戻ってきた回数で定義する。

$$Reg(w_x) = |\{k \in [2, m] \mid w(k-1) > x, w(k) = x\}|. \quad (9)$$

mean fixation duration

単語 w_x 上の fixation の平均時間 $f_{\text{mean}}(w_x)$:

$$f_{\text{mean}}(w_x) = \frac{1}{N(w_x)} \sum_{w(k)=x} d(k). \quad (10)$$

$F(w_x)$ は単語 w_x に関連付けられた fixation の群を表す。

max gaze duration

単語 w_x 上の gaze duration の中で最大のもの $g_{\text{max}}(w_x)$:

$$g_{\text{max}}(w_x) = \max \{G(w_x)\}. \quad (11)$$

pupil size variation

単語 w_x 上の瞳孔の大きさの変化量 $P_{\text{variation}}(w_x)$:

$$P_{\text{variation}}(w_x) = P_{\text{max}}(w_x) - P_{\text{min}}(w_x). \quad (12)$$

$P_{\text{max}}(w_x)$ は単語上 w_x での瞳の直径の最大値、 $P_{\text{min}}(w_x)$ は最小値を表す。

4.4.2 言語特徴量抽出

この節では言語特徴量の抽出法を示す。

word rarity

前処理として、刺激文の単語に対し、TreeTagger [14] を用いて形態素解析を行い、動詞の活用形や名詞の複数形に対して原形を抽出する。抽出した原形に対して、語彙の頻度の順位 (降順) を BNC コーパスの語彙頻度リスト [2] から抽出し、3章の (3) 式に基づいて $Rare(w_x)$ を決定する。

word length

1600x900 の画面上での単語 w_x の物理的長さ $Len(w_x)$:

$$Len(w_x) = D(s_x, e_x). \quad (13)$$

4.5 特徴量選別

定義にしたがって特徴量を抽出し、未知語と既知語の間で、平均に差がないという帰無仮説のもとで非等分散 t 検定を行い、 p 値の低いものを優先して特徴量として用いる。さらに未知語ラベルを 1, 既知語ラベルを 0 とした場合に、これらのラベルと各特徴量の相関係数を計算し、相関の高いものを優先して特徴量として用いる。

4.6 未知語検出

未知語の検出にあたっては、提案手法として非線形 SVM (RBF カーネル) を用いることを提案する。評価は 10 分割の交差検定で行う。評価尺度については未知語の割合は、全文章に対して少ないデータであるため、通常の正解率での精度を用いずに F 値を用いる。使用する特徴量によって最適パラメータが変わることを考慮する。よって各特徴量ごとに grid search を行い、テストセットの F 値に対して、計算機的にハイパーパラメータの最適化を行う。

5. 実験的評価

本節では提案手法が、未知語検出の精度向上に影響を与えるかを実験的に検証する。

5.1 実験条件

実験では 1600×900 の Eizo のスクリーン上に視覚刺激として 3 つの文章セットを被験者に提示し、読解させた。計測には非装着型である Tobii Pro x2-30 を用いた。Tobii Pro x2-30 のサンプリング周波数は、30Hz であるがこれは fixation の検出に十分と考える。計測は密閉された同一の部屋で行い、照明は一定であった。fixation-saccade の判定は Tobii studio の IV-filter の基本設定を使用した。

文章セットには 283 語の速読英単語 [11], 285 語の TOEFL の読解課題 [15], 381 語の Rotter の心理学論文 [5] を用いた。刺激は PDF 形式で verdana font の 14pt で表示し、垂直方向の計測誤差を減らすために 6 行の改行を入れた。図 2 に使用した視覚刺激の一部を示す。実験で用いる文章の量が多いことを踏まえ、3 回に分けて計測は各文章ごとに行った。内容を理解しながら読解を進めるため、一つの文章刺激終了後、内容についての要約課題を被験者に課した。要約課題が終了次第、該当する刺激文章を被験者に再度見せ、未知語のアノテーションを行った。

実験は 20 代の日本人大学生および研究員、計 5 名 (女性 3 名, 男性 2 名) のユーザーを対象に行った。その内 1 名は眼鏡を着用しており、さらに 1 名はコンタクトレンズを着用していた。被験者をスクリーンから 60cm 程離し、頭部が極力動かないように肘をつけて読解をしてもらった。被験者は各自のマウス操作により、任意のタイミングで、文章を切り替えるようにした。計測中は画面以外を見ないように伝え、マウスカーソルも画面上には表示されないようにした。表 1 に被験者ごとの、英語処理能力 (WPM, TOEIC) と読解時間、未知語の割合の関係を示す。P4 に関しては、文章 1 [11] での測定中にアイトラッカの動作不備があった。そのため文章 1 を除き、残りの 666 語を用いて分析を行った。

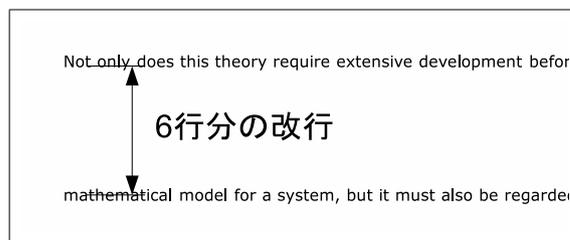


図 2 視覚刺激用文章の提示例。Rotter の心理学論文 [5] の概要を複数枚の PDF にし、十分な改行を加えて、視覚刺激としたもの。1 ページあたり 4 行以内とした。

表 1 被験者の英語処理力と未知語の割合。表中の P は被験者の番号を表す。未知語は被験者ごとの未知語総数を表す。読語は被験者が実験で読解した単語の総数を表す。表中の * は文章 1 を除いた 2 つの刺激文章での読解所要時間を表す。

被験者	未知語	読語	割合	所要時間 [s]	WPM	TOEIC
P1	31	949	0.033	460.8	124	790
P2	75	949	0.079	551.8	103	685
P3	60	949	0.063	564.7	101	680
P4	41	666	0.062	768.0*	52	480
P5	72	949	0.076	1909.7	30	475
合計	279	4462	0.063	4255.0		

5.2 特徴量の分布

被験者全員の要約内容を主観評価した結果、全員が文章と関連した内容を記したため、実験データとして用いることにした。

図 3 に被験者ごとの特徴量の分布を示す。全体的にどの特徴量に関しても未知語の平均のほうが既知語の平均よりも大きくなっていることがわかる。また、眼球運動特徴量に関して、mean fixation duration 以外はいずれも個人差が大きくなっていくことが分かる。特に total gaze duration, max gaze duration については、既知語の標準偏差での分布範囲が未知語の平均を上回ることはなく、特徴量としての有効性を示している。言語特徴量に関して、word length や word rarity は個人差が出にくい特徴量となった。具体的には、標準偏差を分布範囲と考えれば、既知語の分布と未知語の分布を分けやすい特徴量となった。

5.3 特徴量選別結果

3.4 節の定義にしたがって特徴量を抽出し、未知語と既知語の間で、平均に差がないという帰無仮説のもとで非等分散 t 検定を行った。さらに未知語ラベルを 1, 既知語ラベルを 0 とした場合に、これらのラベルと各特徴量の相関係数を計算し、識別に用いる特徴量を選別した。表 2 に特徴量の検証結果を示す。word rarity が最も相関が強く、 p 値も小さいことがわかる。逆に regression が最も相関が低く、 p 値も大きくなっている。表 2 より、未知語と既知語のラベルとの相関は強くはないことが分かるが、いずれも無相関検定で、無相関ではないことが示された ($p < 0.01$)。また t 検定はいずれも平均に有意な差があると言えた ($p < 0.01$)。word rarity が最も相関が強く、 p 値も小さかった。これは難しい語彙ほど未知語になりやすいことを意味する。また max gaze duration が次に相関が高く、 p 値も 3 番目に低い。これはユーザーが未知語の単語に関しては、一定の時間同じ単語を注視したことを意味する。また word length も相

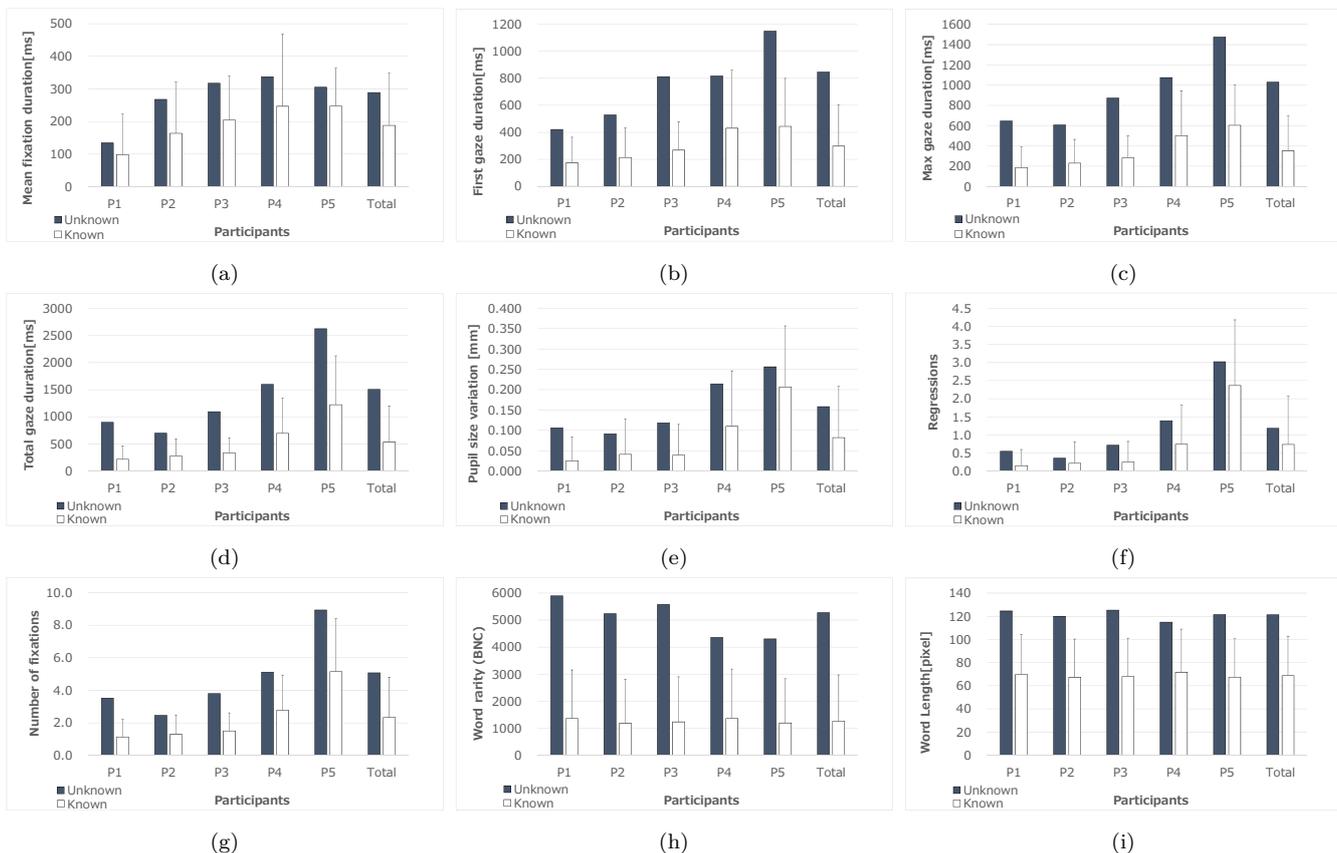


図3 各特微量の被験者ごとの分布。縦軸は各特微量ごとに取り得る値であり、横軸は被験者を表し、左から TOEIC スコア順 (降順) にソートした (表 1)。図中の色付きの値が未知語の平均を表す。図中の白の値が既知語の平均を表す。またエラーバーは既知語の標準偏差を表している。

表 2 特微量選別。ラベルの相関係数で特微量をソートした。太字は特微量として識別に使用したものを表す。 t 検定の p 値は指数のみを記載してある。

特微量	相関係数	t 検定の p 値
word rarity : $Rare(w)$	0.499	$\cdot x10^{-151}$
max gaze duration : $g_{\max}(w)$	0.375	$\cdot x10^{-28}$
word Length : $Len(w)$	0.353	$\cdot x10^{-94}$
first gaze duration : $g_{\text{first}}(w)$	0.336	$\cdot x10^{-20}$
total gaze duration : $g_{\text{total}}(w)$	0.305	$\cdot x10^{-24}$
number of fixation : $N(w)$	0.244	$\cdot x10^{-21}$
mean fixation : $f_{\text{mean}}(w)$	0.148	$\cdot x10^{-15}$
pupil variation: $P_{\text{variation}}(w)$	0.144	$\cdot x10^{-17}$
regression : $Reg(w)$	0.079	$\cdot x10^{-6}$

関と p 値において、未知語との関連が強いことがわかる。これは長い語ほど未知語になりやすいということを意味する。反対に regression は相関でも p 値でも、未知語との関連性は低かった。これは未知語であることが直接逆行回数に反映されなかったことを意味する。同様に pupil size variation も関連性が低く、瞳孔の大きさの変異は、未知語との関連性が低いことを意味する。

5.4 識別結果

本研究では、従来手法として (4) 式をベースラインとした。尚、 $a_h\sigma_h$, $a_l\sigma_l$ をそれぞれ一つのパラメータとして扱い、テストセットに対して F 値が最大となる $a_h\sigma_h$, $a_l\sigma_l$ を用いた。提案

手法として、相関係数が上位 5 つの特微量の組み合わせを用いて識別を行った (提案手法 1 は相関係数の高い特微量 $g_{\max}(w)$ での閾値関数を用いた結果である)。

表 3 より、自然読解の未知語検出において、SVM は従来手法と比較し、F 値において精度向上に有効であることが確認された。F 値での精度の向上の有意性を確認するために、Kohen の精度評価 [8] を参考にし、従来手法と提案手法での Paired Bootstrap 検定を行った。検定は提案手法と従来手法の出力値に対して行う。サンプリングサイズはデータの半分とし、リサンプリング回数は 10000 とした。結果、従来手法と表 3 中の SVM を用いた全ての場合との間で、有意な精度の改善が確認された ($p < 0.01$)。また max gaze duration を使った場合の精度向上の有意性を確かめるため、提案手法 2,3 間と、提案手法 5,6 間で同様の検定を行ったが、いずれも有意な精度改善とはならなかった ($p > 0.1$)。

6. 考察

提案した注視特徴である max gaze duration は total gaze duration と比較して識別の精度向上は有意差ではなかったものの、閾値関数での total gaze duration との比較、SVM での識別結果を踏まえると、誤検出を減らす役割があると考えられる。これは max gaze duration は gaze の最大値であるため、ユーザーが意図せずに固視した場合でも、値が更新されにくい

表 3 識別結果. 識別関数は閾値関数 (4) か, SVM (RBF カーネル) のいずれかを表す. 適合率, 再現率, 偽陽性, F 値はいずれも未知語検出についての精度指標である. 特徴量は識別に使用した特徴量を意味する. 手法は左から 1 番目に従来手法と名付け, 2 番目から 8 番目に提案手法 1-7 とラベルを振った. 太字は精度評価では F 値に注目したこと, 最大の F 値を表している.

手法	従来手法	提案手法 1	提案手法 2	提案手法 3	提案手法 4	提案手法 5	提案手法 6	提案手法 7
識別関数	閾値関数	閾値関数	SVM	SVM	SVM	SVM	SVM	SVM
適合率	0.612	0.670	0.485	0.520	0.490	0.512	0.523	0.506
再現率	0.412	0.365	0.763	0.703	0.756	0.756	0.742	0.756
偽陽性	0.016	0.011	0.050	0.041	0.049	0.045	0.042	0.049
F 値	0.493	0.472	0.593	0.598	0.594	0.610	0.613	0.606
特徴量	<i>Rare(w)</i>	<i>Rare(w)</i>	<i>Rare(w)</i>	<i>Rare(w)</i>	<i>Rare(w)</i>	<i>Rare(w)</i>	<i>Rare(w)</i>	<i>Rare(w)</i>
	<i>g_{total}(w)</i>	<i>g_{max}(w)</i>	<i>g_{total}(w)</i>	<i>g_{max}(w)</i>	<i>g_{total}(w)</i>	<i>g_{total}(w)</i>	<i>g_{max}(w)</i>	<i>g_{total}(w)</i>
					<i>g_{max}(w)</i>	<i>Len(w)</i>	<i>Len(w)</i>	<i>g_{max}(w)</i>
								<i>Len(w)</i>
								<i>g_{first}(w)</i>

だと考えられる. すべての特徴量を利用した提案手法 7 は F 値で 0.606 に留まり, 3 つの特徴量を用いた手法 6 が最も F 値が高かった. total gaze duration, max gaze duration, first gaze duration の間には以下の関係が成り立つ.

$$g_{\text{first}}(w_k) \leq g_{\text{max}}(w_k) \leq g_{\text{total}}(w_k). \quad (14)$$

よって max gaze duration を使用することで, total gaze duration と比較し, 検出速度の改善となる可能性が示された. first gaze duration は検出速度改善に働く可能性があるが, 一度記録されてしまうと更新されないというデメリットがある.

本研究の限界を考察する. 初めに pupil size に対しては, 変化までに 200-300ms のレイテンシーがあることが知られている [3]. 特徴量を抽出する上で, これが考慮されていなかったことが問題点として挙げられる. 次に, fixation-saccade の判定を Tobii studio の IV-filter の基本設定で行っているため, 問題設定に適した fixation-saccade 判定を行うことも検討していく.

7. まとめ

本稿では, 非母語文書の読解支援をするために, 新たな注視特徴を用いて SVM での検出精度の向上を目指した. これにより非装着型アイトラッカから取得した眼球運動データから, 自然読解中の未知語検出において, 有意な F 値での精度改善を達成することができた. また新たな特徴量 max gaze duration を用いることで誤検出を減らせることを確認した.

今後の課題として, ユーザーごとに WPM や知識量の違いがあるため, 個人適応をする必要がある. saccade も重要な眼球運動の特徴量であり, 理解度判定や読み判定に有効である [13] ので, 今後考慮することでシステムパフォーマンスの改善が期待できる. 文章の難易度による被験者の読解行動の違いも考慮する必要がある. さらに未知語に対しては fixation の時間は長くなる傾向があるが, 文章自体が理解しにくかったり, 混乱を招いたりする場合にも高くなることが知られている [7]. したがって, 精度改善のためには未知語遭遇以外で, fixation が長くなる現象を考慮する必要がある.

謝辞 本研究は, JSPS 科研費 26540117 の助成を受けて行われたものである.

文 献

- [1] A. Hyrskykari. Eyes in attentive interfaces: Experiences from creating iDict, a gaze-aware reading aid. Tampereen yliopisto, 2006.
- [2] A. Kilgarriff's webpage, BNC database and word frequency lists, <http://www.kilgarriff.co.uk/bnc-readme.html>, 1996.
- [3] B. Gagl, S. Hawelka, F. Hutzler, Systematic influence of gaze position on pupil size measurement: analysis and correction, Behav Res 43, pp.1171-1181, 2011.
- [4] C.-C. Chang, C.-J. Lin LIBSVM, A Library for Support Vector Machines, <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>, 2015
- [5] J. Rotter, Social learning and clinical psychology. Englewood Cliffs, NJ, US: Prentice-Hall, Inc. 1954.
- [6] K. Kunze, H. Kawaiichi, K. Yoshimura, K. Kise, Towards inferring language expertise using eye tracking, CHI'13 Extended Abstracts on Human Factors in Computing Systems, ACM, pp.217-222, 2013.
- [7] K. Rayner, Eye movements in reading and information processing: 20 years of research. Psychological bulletin, 124(3), pp.372-422, 1998.
- [8] P. Koehn, Statistical Significance Tests for Machine Translation Evaluation, EMNLP, pp.388-395, 2004
- [9] Y. Ehara, N. Shimizu, T. Ninomiya, H. Nakagawa. Personalized Reading Support for Second-Language Web Documents by Collective Intelligence. In the Proceedings of the 2010 International Conference on Intelligent User Interfaces (IUI 2010), pp. 51-60, Hong Kong, China, February 2010.
- [10] 江原遥, 二宮崇, 中川裕志. Web 文書中の単語クリックログの解析から未知単語を予測する語義注釈システム. 情報処理学会研究報告, pp.1-7, 2009.
- [11] 風早寛, “速読英単語必修編”[第四版], Z 会, 2004.
- [12] 長瀬友樹, 小谷克則, 工藤竜広, 佐久間みゆき, 秋葉泰弘, 実務翻訳における機械翻訳の利用に関する調査報告, 言語処理学会, 第 20 回年次大会発表論文集, pp.1-4, 2014.
- [13] 藤好宏樹, 吉村和代, Kai Kunze, 黄瀬浩一, 英文問題解答時の視点情報を用いた英語能力推定法, 電子情報通信学会技術研究報告 pp.49-54, 2015.
- [14] TreeTagger:<http://www.cis.uni-muenchen.de/~schmid/tools/>
- [15] <https://www.ets.org/Media/Tests/TOEFL/pdf/SampleQuestions.pdf>
- [16] <https://ja.wikipedia.org/wiki/Wikipedia:全言語版の統計>, 4/1, 2016.