

対数正規分布を用いた英単語筆記時の確信度推定手法の提案

丸市 賢功^{1,a)} 黄瀬 浩一^{1,b)}

概要: 英単語の学習において、復習は重要なプロセスである。一般的に復習は不正解の英単語に対して行われる。しかし、この方法では確信なくたまたま正解してしまった英単語を復習することができない。そこで本研究では、このような英単語を復習に含めるため、解答に関する確信度を推定する手法を提案する。これにより、確信をもって不正解した英単語に特別の注意を与えることも可能となる。本研究では、確信度推定のために使用する学習者の振舞いとして、英単語問題解答時の手書きログから抽出した特徴量に着目する。これらの特徴量の分布は対数正規分布に従うことが確認されたため、対数正規分布を用いて確信度推定を行う手法を提案する。また、提案手法を推定精度を検証するために 21 名を対象に実験を行い、実験の精度について正規分布で仮定したものととの比較を行った。各手法のユーザ依存での平均推定精度は、正規分布を仮定した場合には 76%、対数正規分布を仮定した場合には 78%であった。

キーワード: 確信度, 手書きログ, 対数正規分布, 学習支援, Quantified learning

1. はじめに

スマートセンサから取得される視線や姿勢、反応速度などの生体情報は、人間の心的状態を知る上で大変有益な情報である。学習支援の分野では、これらの情報を用いて、集中度や疲労度、そして確信度などの心的状態を推定し、学習者の学習効率を向上させる取り組みは“Quantified learning”と呼ばれ、近年盛んに研究されている [1], [2], [3].

中でも、確信度は学習者のテストの成績と関連があることが知られている。Kleitmann らの研究 [4] では、小学生を対象に、解答に対する確信の有無とテストの成績との関係の調査がなされている。調査の結果、確信の有無と解答の正誤の一致率が高い学習者では、テストの成績が高い傾向にあることが明らかとなった。また、Stankov らは、確信度が英語と数学の理解に対する良い指標であると主張している [5]。確信をもって解答したにも関わらず、不正解となってしまった問題は、誤った知識に基づいて解答された可能性が高いことが示されている。

これらの研究から、英単語の学習において、確信度は復習対象を決定するための良い指標となりうることが考えられる。英単語の学習においては、テストを反復して行う必要がある。一般的な学習方法としては、英単語を暗記し、テストを行うことで解答の正誤をフィードバックし、その

中から、図 1 (a) のように不正解の英単語のみを復習対象として復習を行い、再度テストをして確認するというものがある。この一連のプロセスにおいて、復習は学習効率の向上に大きく寄与している。確信度を考慮することで、復習対象がより明確になると考える。なぜなら、テストで正解した問題の中にも、確信が持たず偶然正解した英単語が含まれているからである。これらの英単語は次のテストでは間違える可能性が高く、復習対象とすべきである。また、テストで不正解となってしまった英単語の中にも、解答に対する確信があった英単語も含まれている。これらの英単語は誤った暗記をしており重点的に復習する必要がある。すなわち、学習効率を上げる理想的な復習方法とは、図 1 (b) に示す通り、解答の正誤と確信度の両方をフィードバックすることで、復習対象に優先順位をつけ、優先度の高いものから重点的に復習する復習方法であると考えられる。

英単語に対する確信度を推定するには、確信度を逐次学習者が記録しておく方法 [6] が考えられるが、この方法は学習者の集中を阻害してしまう恐れがある。そこで我々はテストの際の学習者の振舞いをもとに確信度を機械学習によって自動的に推定する手法を提案する。

確信度推定のために使用する学習者の振舞いとして、英単語問題回答時の手書きログに着目する。確信度推定のための手法としては、視線、タイピングログ、手書きログにもとづくものが既に提案されている [7], [8]。しかし、手書きログの手法については、ブロック体での筆記のみに特化した手法であり、筆記体での検証はなされていない。本研

¹ 大阪府立大学
〒 599-8531 大阪府堺市中央区学園町 1-1
a) maruichi@m.cs.osakafu-u.ac.jp
b) kise@cs.osakafu-u.ac.jp

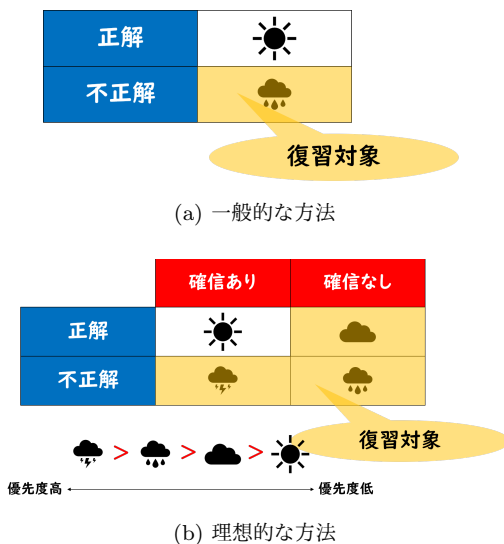


図 1: 復習方法

究では、筆記体を交えた自由筆記形式においても、有効な手法を提案する。

提案手法の既存手法との最も大きな差異は、対数正規分布を用いたことにある。先行研究において、手書きログから抽出した筆記スピードや解答時間などの特徴量の分布は対数正規分布に従うことが確認されている [9], [10]。本研究で提案する手法においても、特徴量の分布が対数正規分布に従うと仮定して確信度推定を行う。

したがって、本研究では、英単語問題回答時の手書きログから抽出した特徴量の分布を対数正規分布に従うと仮定して確信度推定を行う手法を提案する。提案手法を有効性を検証するために 21 名を対象に実験を行い、実験の精度について正規分布で仮定した既存手法 [8] との比較を行った。実験の結果、各手法のユーザ依存での平均推定精度は、正規分布を仮定した場合では 76%、対数正規分布を仮定した場合では 78%であった。

なお、本研究は、大阪府立大学工学研究科倫理委員会の承認を得ている事を付記しておく。

2. 関連研究

確信度と成績の関連についての調査研究としては以下のものが挙げられる。Kleitmann ら [4] は、小学生を対象に、解答に対する確信の有無とテストの成績との関係の調査を行い、確信の有無と解答の正誤の一致率が高い学習者では、テストの成績が高い傾向にあることが明らかにしている。確信の有無に関する自覚の高い学習者は不正解の問題だけでなく、偶然正解した問題についても復習を行うことができるため、復習漏れをなくすことができ、それが成績の向上に影響していることが考えられる。

Stankov らの研究では、確信度が英語と数学の理解に対する良い指標であることが示されており [5]、確信をもって解答したにも関わらず、不正解となってしまった問題は、

誤った知識に基づいて解答された可能性が高いことが示されている。これらの問題を重点的に復習することで、誤った知識を正し、次の正解につなげることが成績向上につながると述べられている。

テスト時の振舞いと心的状態との関連についても既に以下の研究で示されている。Wall らは、分数大小比較テスト解答時のユーザの確信度と目の注視点のデータとの関連について調査している [11]。調査の結果、目の動きはテストの難易度によって大きく差がでることが示されたが確信度推定には至っていない。Yamada らは、多肢選択式問題解答時の視点情報を用いて、解答している問題に対する確信度を推定する手法を提案している [7]。実験の結果、実験環境で 90.1%の識別率を達成しているが、検証は多肢選択式問題に限定されており、記述式問題における検証はなされていない。

また、記述式問題については、Thomas らが、タイピング中の反応時間とプログラミングテストのスコアとの間に負の相関があることを確認している。加えて、浅井らが、タブレットで数学の問題に解答している際の筆圧や筆記速度、筆記間隔などの情報を用いて、つまずいた箇所や忘れやすい単語を検出する手法を提案している [12], [13], [14]。しかし、これらの研究も確信度推定には至っていない。

丸市らは、英単語筆記問題解答時のペンの動作を確信推定に用い、実験参加者 11 名を対象とした実験環境で 80%の識別率を達成している [8]。しかし、ここで提案されている手法はブロック体での筆記のみに特化した手法であり、筆記体での検証はなされていない。本研究ではブロック体・筆記体双方に有効な手法を提案する。

さらに、手書きログから抽出した筆記スピードや解答時間などの特徴量の分布が対数正規分布に従うことが既に確認されている [9], [10]。そこで本研究においても、手書きログから抽出される特徴量の分布が対数正規分布に従うと仮定して確信度推定を行う。

3. 提案手法

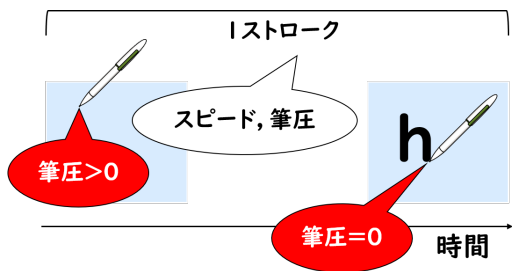
3.1 データ取得

本手法では、ペンがディスプレイに触れる動作、ペンがディスプレイから離れる動作、そしてペンを動かす動作の 3 種類をペンイベントとして定義する。ペンイベントの種類、タイムスタンプ、ペンイベントが起こった点のディスプレイ上の $x-y$ 座標、筆圧の 4 種類の情報を取得する。

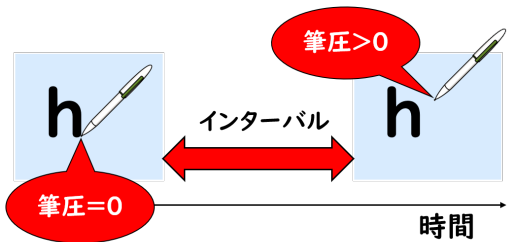
3.2 特徴量抽出

取得したデータから、特徴量を抽出する。本手法では、図 2 (a) のようにペンがディスプレイに触れてから離れるまでの一連の動作を 1 ストロークと定義する。その定義をもとに以下の 5 種類の特徴量を抽出する。

- (1) インターバル：図 2 (b) のようにストロークとスト



(a) ストローク



(b) インターバル

図 2: 本研究におけるストロークとインターバルの定義

ロークの間にかかった時間をインターバルと定義する。すなわち、ユーザがディスプレイからペンを放してから、次にディスプレイに触れるまでの時間間隔のことである。取得されたインターバルから問題ごとの平均を求めて用いる。

(2) **スピード**: 1ストロークでのペンの総移動距離を1ストロークにかかった時間で割った値をスピードと定義する。これらの1ストロークごとスピードから問題ごとの平均を計算し用いる。

(3) **筆圧**: 取得した筆圧情報から問題ごとの平均を計算する。

(4) **解答開始前/終了後の時間間隔**: 解答開始前の時間間隔とは、問題が提示されてからユーザが解答を開始するまでの時間間隔のこと、解答終了後の時間間隔とは、ユーザが最後のストロークを終えてから解答の決定ボタンを押すまでの時間間隔のことである。

3.3 特徴量の対数化

これまで行われてきた筆記ログや解答時間などの解析研究において、筆記スピードや解答時間は対数正規分布に従うことが確認されている [9], [10]。対数正規分布は以下の式で表される:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma x}} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}$$

本研究においても、図 3 に示すインターバルの平均、スピードの平均、筆圧の平均、解答開始前の時間間隔、解答終了後の時間間隔の 5 種類全ての特徴量の分布について、対数正規分布に従うことが Kolmogorov-Smirnov 検定 [15] によって確認できた ($p < 0.01$)。したがって、これらの特徴量については対数化を行うこととする。最終的に得られ

表 1: 特徴量の一覧

No	特徴量
f1	インターバルの平均を対数化したもの
f2	スピードの平均を対数化したもの
f3	筆圧の平均を対数化したもの
f4	解答開始前の時間間隔を対数化したもの
f5	解答終了後の時間間隔を対数化したもの

た特徴量を表 1 に示す。

3.4 識別器の学習

SVM を用いて学習を行い、確信あり (ラベルは 1 とする) または確信なし (ラベルは 0 とする) の 2 クラスに分類する。

4. 実験

提案手法の有効性を示すための実験を行った。本実験では先行研究 [8] で取得されたデータセット (以下データセット A と呼称する) の他に新たに取得したデータセット (以下データセット B と呼称する) も使用する。

4.1 実験設定

4.1.1 使用機器

データセット A, B ともに Surface Studio を用いた。データのサンプリング周波数は 100Hz であり、筆圧は 1024 段階で取得した。

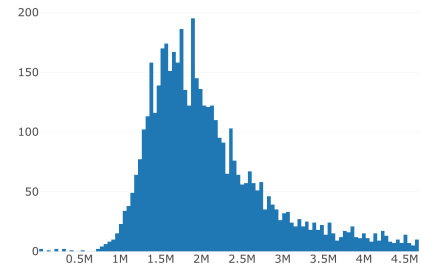
4.1.2 データセット A

データセット A では 11 名 (男性 9 名, 女性 2 名) のドイツの大学に通う日本人または中国人大学生に、図 4 (a) に示すような形式の問題を解くよう依頼した。問題には TOEIC テストでよく問われる英単語を出題した。

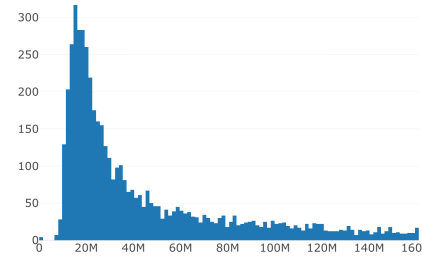
実験手順について述べる。まず、10 問の英単語問題を解答してもらい、1 問の解答終了ごとに確信度を (1) 意味・スペルともに確信あり (2) 意味のみ確信あり (3) 意味・スペル共に確信なしの 3 通りで報告するよう指示した。しかし、(2) に対応するデータが少なすぎたため、推定時には (2) と (3) を統合し、(1) を確信あり、(2) または (3) を確信なしとした 2 値分類を行うこととした。解答は小文字ブロック体で、インターバルがとりやすいように 1 マスに 1 文字を入れるような形で解答するよう指示した。解答が分からない場合には問題をスキップすることも可能であり、その問題については推定には用いないこととした。この手順を 1 時間で出来る限り繰り返し、解答する際のペンの振舞いを記録した。

4.1.3 データセット B

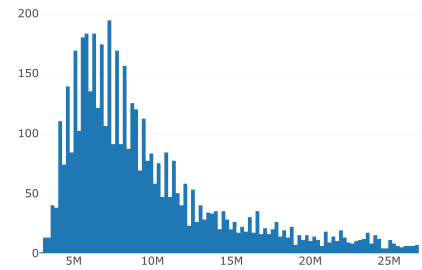
データセット B では 10 名 (男性 5 名, 女性 5 名) の日本人大学生に授業の 2 コマ (3 時間) を使って、図 4 (b) に示すような形式の問題を解くよう指示した。問題には大学



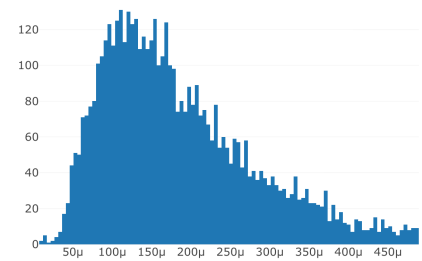
(a) インターバルの平均 ($\mu = 15, \sigma = 0.52$)



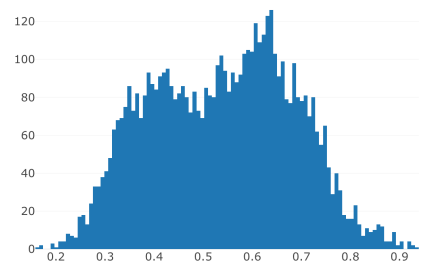
(b) 解答開始前の時間間隔 ($\mu = 18, \sigma = 0.87$)



(c) 解答終了後の時間間隔 ($\mu = 17, \sigma = 0.71$)



(d) スピードの平均 ($\mu = -8.7, \sigma = 0.51$)

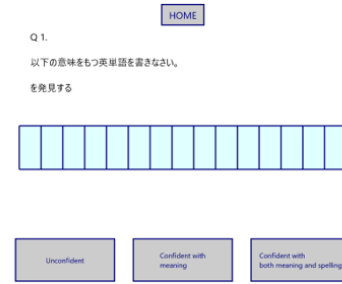


(e) 筆圧の平均 ($\mu = -1.0, \sigma = 0.13$)

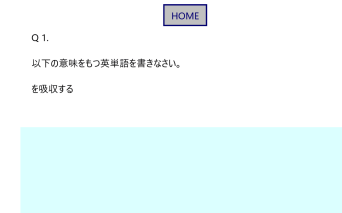
図 3: 各特微量のヒストグラム

入試でよく問われる英単語を出題した。

実験手順について述べる。まず、10 問の英単語問題を解



(a) データセット A



(b) データセット B

図 4: 解答フォーム

答してもらい、1 問の解答終了ごとに確信度を (1) 確信あり (2) 確信なしの 2 通りで回答するよう指示した。解答は小文字で、ブロック体または筆記体のどちらか好きな方で解答するよう指示した。解答が分からない場合には問題をスキップすることも可能であり、その問題については推定には用いないこととした。この手順を 3 時間で出来る限り繰り返し、解答する際のペンの振舞いを記録した。

4.2 検証方法

識別器の学習方法としては、実験参加者依存と実験参加者非依存の 2 通りが考えられるが、本実験で取り扱った手書きの振舞いは実験参加者ごとに大きく異なっていたため、実験参加者依存の場合のみについて検証を行った。

実験参加者依存とは、識別器の学習に際して、特定の実験参加者のデータのみを用い、同一実験参加者に対して確信推定を行う検証法をいう。実験参加者依存での学習を行うことで、その実験参加者が解答する際に確信がある場合あるいは確信がない場合にどのような振舞いを行うのかを学習することができる。その振舞いの差が識別できれば、確信推定がより容易になる。実験参加者依存で学習された識別器を他の実験参加者に用いることはできない。本実験では、各実験参加者について、データを 10 のセクションに分割し、1 セクション分をテストデータ、残りのセクションを学習データとして分割し、全てのセクションについて評価を行う 10-fold cross-validation を用いた。

表 2: 実験参加者ごとの問題数

実験参加者	問題数	実験参加者	問題数
A01	210	B01	176
A02	146	B02	234
A03	174	B03	210
A04	208	B04	592
A05	70	B05	546
A06	180	B06	396
A07	182	B07	950
A08	340	B08	988
A09	380	B09	580
A10	224	B10	438
A11	394		
合計	2508	合計	5110

また、本実験で使用したデータセットは実験参加者が確信なしと回答した問題数が少なかったため、Random-over-sampler を用いて確信なしの問題数をオーバーサンプリングして用いた。その結果得られた実験参加者ごとの問題数を表 2 に示す。問題数のうち、半数は確信あり、半数は確信なしである。

5. 結果・考察

5.1 識別率

実験から得られた識別率を表 3 に示す。比較手法として、表 1 の特徴量を対数化せずにそのまま適用した手法を用いた。提案手法の識別率が比較手法と比較して統計的に有意である ($p < 0.01$) ことが符号検定 [16] によって確認できた。

各実験参加者ごとの識別率は図 5 に示す通りである。ここで、実験参加者 A05 の識別率が他の実験参加者に比べて著しく低くなっている。この原因としては問題数が少ないことが挙げられる。表 2 から分かる通り、この実験参加者は他の実験参加者に比べて、1 問ごとの解答時間が長く、十分なサンプル数を取得することができなかった。そのため、振舞いのパターンを十分に学習しきれなかったと考えられる。実験参加者 B04 と B06 については、一部の問題について、推定がうまくいった実験参加者 A07 や B07 に比べて確信がない場合も素早く解答する傾向が見られた。顕著な例として、図 6 に各実験参加者の解答開始前の時間間隔の分布を示す。図 6 (a) および図 6 (b) が推定がうまくいった例、図 6 (c) および図 6 (d) がうまくいかなかった例である。このような振舞いをする実験参加者の識別率を上げるためにも、特徴量にさらなる工夫が必要であると考える。

5.2 Precision-Recall 曲線

識別結果から得られた 11 点 Precision-Recall 曲線 [17] を確信あり、確信なし双方の場合について描画した (図 7 参照)。凡例の数字は AUC (Area Under the Curve) を示している。提案手法・比較手法の双方において確信なしの

表 3: 実験結果

手法	データセット	識別率
比較手法	A+B	0.76 ± 0.06
	A	0.76 ± 0.07
	B	0.74 ± 0.05
提案手法	A+B	0.78 ± 0.06
	A	0.78 ± 0.07
	B	0.77 ± 0.05

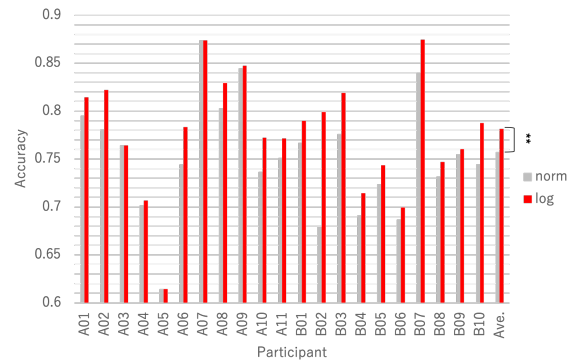


図 5: 各実験参加者ごとの識別率：**印は提案手法での識別率が比較手法と比較して統計的に有意であることを示している ($p < 0.01$).

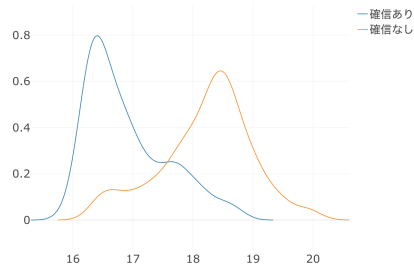
表 4: データセット全体での各特徴量と確信の相関係数

No	特徴量	相関係数
f3	筆圧の平均を対数化したもの	0.10
f2	スピードの平均を対数化したもの	0.02
f5	解答終了後の時間間隔を対数化したもの	-0.14
f1	インターバルの平均を対数化したもの	-0.19
f4	解答開始前の時間間隔を対数化したもの	-0.25

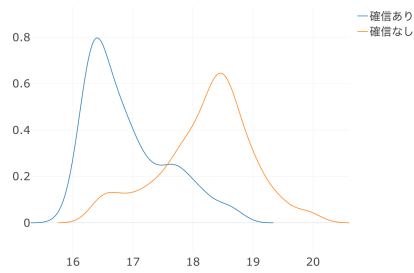
場合の AUC が確信ありの場合よりも低いことが確認された。この理由としては、どちらのデータセットにおいても確信なしと回答された問題の割合が低く、オーバーサンプリングしたからであると考えられる。これにより、確信なしの場合の振舞いが十分に学習されず、推定が難しくなっていると考えられる。今後行う実験では、この点を考慮してバランスの良いデータセットが取得できるよう難易度を調整するなどの対策を講じる予定である。

5.3 特徴量との相関

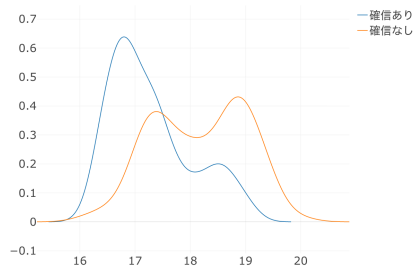
各特徴量と確信の有無の相関係数を表 4 に示す。表 4 から、インターバルの平均、スピードの平均、解答開始前の時間間隔、解答終了後の時間間隔は確信の有無と負の相関があることが分かる。すなわち、確信がない場合は、思考時間が長くなるということである。したがって、これらの特徴量は確信推定に有効であるといえる。また、確信の有無と強い相関を持った特徴量は確認できなかった。そのため、選択された特徴量一つ一つの組み合わせが確信推定に寄与している可能性もあると考える。



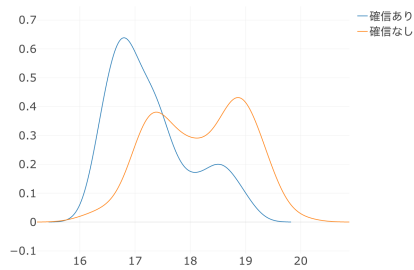
(a) 実験参加者 A07



(b) 実験参加者 B07



(c) 実験参加者 B04

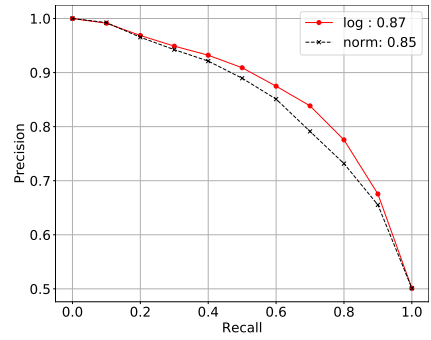


(d) 実験参加者 B06

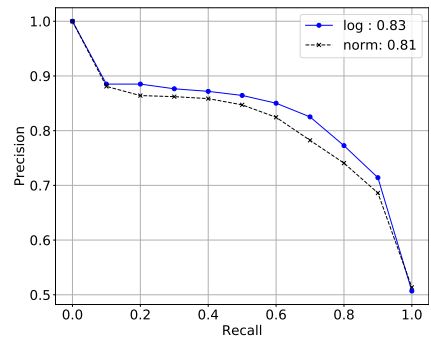
図 6: 解答開始前の時間間隔の分布: 図 6 (a) および図 6 (b) が推定がうまくいった例, 図 6 (c) および図 6 (d) がうまくいかなかった例である。

6. まとめ・今後の課題

本稿では, より効果的な英単語の復習を実現するために確信を手書きログによって推定する手法を提案した。実験の結果, 各手法のユーザ依存での平均推定精度は, 正規分布を仮定した場合では 76%, 対数正規分布を仮定した場合では 78%であり, 対数正規分布を用いた手法の方が良い手法であることが統計的に確認された。また, 提案手法は確信がある場合もじっくり考えて解答するタイプのユーザの



(a) 確信あり



(b) 確信なし

図 7: Precision-Recall 曲線: log. が提案手法, norm. が比較手法, 数字が AUC を示している。

識別を苦手とすることが分かった。今後の課題としては以下の 3 点があげられる:

- (1) 確信がない場合も素早く解答する傾向をもつユーザに対処する手法を考案する。
- (2) 学習効果検証のための実験を実施し確信フィードバックが本当に復習に有効であるかを確認する。
- (3) 確信のフィードバックによって学習効果が現れる人と現れない人を比較し, その振舞いについて分析する。

謝辞

本研究の一部は, JST CREST (Grant No. JP-MJCR16E1), ならびに大阪府立大学キーププロジェクトの補助による。

参考文献

- [1] Dengel, A.: Digital co-creation and augmented learning, *Proceedings of the The 11th International Knowledge Management in Organizations Conference on The changing face of Knowledge Management Impacting Society*, ACM, p. 3 (2016).
- [2] Ishimaru, S., Jacob, S., Roy, A., Bukhari, S. S., Heisel, C., Großmann, N., Thees, M., Kuhn, J. and Dengel, A.: Cognitive State Measurement on Learning Materials by Utilizing Eye Tracker and Thermal Camera, *2017 14th IAPR International Conference on Document Analysis and Recognition*, Vol. 8, IEEE, pp. 32-36 (2017).
- [3] Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T. and Graesser, A. C.: Better to be frustrated than bored:

- The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments, *International Journal of Human-Computer Studies*, Vol. 68, No. 4, pp. 223–241 (2010).
- [4] Kleitman, S., Stankov, L., Allwood, C. M., Young, S. and Mak, K. K. L.: Metacognitive self-confidence in school-aged children, *Self-directed learning oriented assessments in the Asia-Pacific*, Springer, pp. 139–153 (2012).
- [5] Stankov, L., Lee, J., Luo, W. and Hogan, D. J.: Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety?, *Learning and Individual Differences*, Vol. 22, No. 6, pp. 747–758 (2012).
- [6] Stankov, L., Kleitman, S. and Jackson, S. A.: Measures of the trait of confidence, *Measures of personality and social psychological constructs*, Elsevier, pp. 158–189 (2015).
- [7] Yamada, K., Kise, K. and Augereau, O.: Estimation of confidence based on eye gaze: an application to multiple-choice questions, *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, ACM, pp. 217–220 (2017).
- [8] 丸市賢功, 石丸翔也, Augereau, O., 黄瀬浩一: 英単語問題解答時の確信判定システムの精度検証, *SIG-ALST*, Vol. B5, No. 01, pp. 44–49 (2019).
- [9] Kucherbaev, P., Daniel, F., Tranquillini, S. and Marchese, M.: ReLauncher: crowdsourcing micro-tasks runtime controller, *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, ACM, pp. 1609–1614 (2016).
- [10] Plamondon, R., O'Reilly, C., Rémi, C. and Duval, T.: The lognormal handwriter: learning, performing, and declining, *Frontiers in psychology*, Vol. 4, p. 945 (2013).
- [11] Wall, J., Thompson, C. and Morris, B. J.: Confidence Judgments and Eye Fixations Reveal Adults' Fractions Knowledge., *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, pp. 2571–2576 (2015).
- [12] Asai, H. and Yamana, H.: Detecting student frustration based on handwriting behavior, *Proceedings of the Adjunct Publication of the 26th Annual ACM Symposium on User Interface Software and Technology*, ACM, pp. 77–78 (2013).
- [13] 浅井洋樹, 野澤明里, 苑田翔吾, 山名早人: オンライン手書きデータを用いた学習者のつまづき検出, *DEIM Forum*, Vol. 2012 (2012).
- [14] Asai, H. and Yamana, H.: Detecting Learner's To-Be-Forgotten Items using Online Handwritten Data, *Proceedings of the 15th New Zealand Conference on Human-Computer Interaction*, ACM, pp. 17–20 (2015).
- [15] Massey Jr, F. J.: The Kolmogorov-Smirnov test for goodness of fit, *Journal of the American statistical Association*, Vol. 46, No. 253, pp. 68–78 (1951).
- [16] 高木英行: 使える!統計検定・機械学習-I: 2群間の有意差検定, システム/制御/情報, Vol. 58, No. 8, pp. 345–351 (2014).
- [17] Baeza-Yates, Ricardo, Ribeiro, de Araújo Neto, B. et al.: *Modern information retrieval*, New York: ACM Press; Harlow, England: Addison-Wesley, (2011).