CK-003

音読音声と視線情報を用いた英語漫画中の未知語を含んだ吹き出し推定

Estimation of Speech Balloons Including Unknown Words in English Comics Using Speech and Eye Gaze

高池 太郎[1)]　岩田 基[1)]　黄瀬 浩一[1)]
Taro Takaike　Motoi Iwata　koichi Kise

1　はじめに

近年のグローバル化に伴い，世界で幅広く使われている英語の習得は，非英語圏の人々にとって今や必須のものとなっている．英語の習得には，新たな語彙の獲得が非常に重要である．語彙学習においては，偶発的学習と意図的学習を組み合わせることが効果的であると言われている [1]．偶発的学習とは，何らかの行為の副産物として学習効果が得られるような行為のことである．英語で書かれたコンテンツを読んでいく過程で文章の中に含まれる単語の意味を理解していくことで新たな語彙を獲得する，という学習方法が偶発的学習に該当する．偶発的学習は単語の用法を学習できるというメリットがある一方で，時間効率が悪いというデメリットを持つ．意図的学習とは学習効果を得ることを目的として行う行為のことである．語彙学習ではフラッシュカードによる学習がこれに当たる．意図的学習は時間効率が良いというメリットを持つ一方，例文を目にすることで初めて理解できるような単語の用法やニュアンスなどを掴むのが困難であるというデメリットを持つ．したがって，最初に述べたように，2 つの学習をバランスよく行うことが必要である．

偶発的学習の教材としては記事や小説など様々なものを用いることができるが，本稿では漫画を用いた学習について注目する．漫画は人気の高いコンテンツであり，ストーリーを楽しみながら漫画を読む過程で同時に英語を学習できれば非常に有用である．また，漫画を読む習慣がある学習者にとって漫画を用いた英語学習は日常に取り入れやすく，英語学習を始める際に感じる心理的なハードルを下げてくれる．さらに，漫画はテキストとイラストが組み合わさった媒体である．イラストがあることでストーリーをイメージしやすく，テキストのみのコンテンツよりも内容やその中で触れた表現が記憶に残りやすいことが期待できる．

そこで，英語で書かれた漫画を読み進めていく中で出会った意味を知らない単語を記録しておき，後にそれらをフラッシュカードにして復習するという学習方法を考える．これは英語漫画を用いた偶発的学習と意図的学習を組み合わせた学習方法となっており，高い学習効果が期待できる．しかし，実際には漫画中に含まれる意味の知らない単語を逐次記録していくのは手間であり，漫画を読むという行為を阻害してしまう．そこで，学習者が英語漫画を読んでいるときに，コンピュータが自動で意味を知らない単語を記録してくれるシステムがあれば便利である．そこで本研究では，英語漫画を読んでいるときに漫画中から読者が意味を知らない単語を自動で推定する手法を検討する．また，以降は学習者が意味を知らない単語のことを「未知語」と呼ぶことにする．

本研究では，英語学習者に英語漫画を音読してもらい，そのときの学習者の振る舞いをセンシングして得られるセンサ情報を用いることによって漫画中に含まれる未知語を推定する．センサ情報として，音読音声と視線情報の 2 つを用いる．音読音声とは，学習者が英文を声に出して読み上げたときの音声のことである．音読音声に関しては，学習者は未知語を読み上げるときに読み方が不自然になることがあり，それを検知することで未知語が推定できる可能性がある．視線情報は PC を用いて学習しているとき学習者がディスプレイのどこを見ているのかという情報のことである．視線情報に関しては，Garain らの研究で英文の理解度と関係があることが示されており [2]，未知語推定に有効であると考えられる．

漫画はテキストが吹き出しという形で分散配置されている．そのため，吹き出しで括られた比較的小さいテキスト集合と対応する視線情報を取得することが可能である．したがって本研究では英語漫画音読時の音読音声と視線情報を用いて英語学習者の未知語を含んだ吹き出しを推定する手法を提案する．

本稿では，2 章で関連研究，3 章で提案手法，4 章でデータ計測実験，5 章で評価を述べる．

2　関連研究

未知語の推定は，理解の度合い (理解度) や自信の度合い (自信度) など学習者の内的状態と非常に関係が深い．ここではまず，音声を用いて人の内的状態を推定した研究を紹介する．Sabu らは，物語を音読しているときの子供の音読音声から，子供の自信度を推定した [3]．Zhang らは，レゴのタスクを行いながらチューターと話している子供の音声から，子供が自信があるのか，困惑しているのか，ためらっているのかを推定した [4]．このように音声は自信度が反映されるものであり，音声は未知語の推定に有効であると期待できる．

次に，視線情報を用いて人の内的状態を推定した研究を紹介する．Martínez-Gómez らは，読書時の視線情報から学習者の言語の習熟度を推定した [5]．Ishimaru らは，一度読んだ物理のテキストに関する問題を回答しているときの視線情報から，学習者の習熟度を分類した [6]．また，Garain らは視線情報は理解度と関係があることを示した [2]．

最後に，英語漫画を読んでいるときの理解度推定に関する研究を紹介する．大工らは，英語漫画を用いた多読学習時の理解度を視線情報，微表情特徴量，単語数に基づいて吹き出し単位で推定した [7]．視線情報はアイトラッカというデバイスを用いて計測され，微表情特徴量はハイスピードカメラを用いて記録した顔映像から抽出された．高橋らは，アイトラッカを用いて計測した視線情報を用いて英語漫画多読時の学習者の理解度推定をページ単位で行った [8]．これらの研究では視線情報と

1)　大阪公立大学大学院 情報学研究科
Graduate School of Informatics, Osaka Metropolitan University

して fixation と saccade が用いられ，視線情報は英語漫画を読んでいるときの理解度推定にも有効であることが示されている．ここで，fixation とは視線が一定時間停留する点のことであり，saccade とは fixation 間の素早い視線の動きのことである．

3 提案手法

提案手法は，英語漫画の音読学習時の音読音声と視線情報から，各吹き出しに対して未知語が含まれているか否かを推定する．提案手法は，以下の 4 つの過程から構成されている．まず，ヘッドセットとアイトラッカによって音読音声と視線情報を取得する．次に，得られたデータを切り分け，各吹き出しに対応する音読音声と視線情報を特定する．次に，切り分けた音読音声と視線情報から特徴量を抽出する．最後に，識別器を学習させ，各吹き出しに対して未知語が含まれているか否かを推定する．以降，それぞれの過程について詳細に述べる．

3.1 センサデータの取得

音読音声は，学習者が装着するヘッドセットに内蔵されているマイクを通して録音する．音読音声の録音にヘッドセットを用いるのは，マイクと口の間の距離を常時一定に保つことで録音の質に差異が生じないようにするためである．音読音声は非圧縮の音声ファイル形式である wav 形式で録音する．視線情報は，アイトラッカを PC のディスプレイ下部に設置して取得する．

3.2 データ切り分けと対応する吹き出しの特定

本節では，ヘッドセットとアイトラッカによって得られた音読音声と視線情報のデータを吹き出し単位に切り分け，対応する吹き出しを特定する方法について述べる．この方法は，英語漫画の吹き出し領域の座標が事前に得られていることを前提としている．

まず，音読音声に対する操作について述べる．最初に，音読音声を無音区間で分割することで有音区間を取得する．次に，分割された有音区間の開始時刻における視点の座標をアイトラッカから取得する．最後に，取得した視点の座標を吹き出し領域の座標と照らし合わせることで，音声区間と対応する吹き出しを特定する．1 つの吹き出しを音読した音声が複数の有音区間として切り分けられている場合はそれらを統合する．以上の操作により，吹き出し単位での音読音声を取得する．

次に，アイトラッカから得られたデータに対する操作について述べる．まずアイトラッカによって得られた生データを，Buscher らの手法 [9] を用いて fixation と saccade に変換する．次に，英語漫画の吹き出し領域の座標を参照し，fixation の座標がある吹き出しの領域内，あるいは吹き出しの周囲 15 ピクセル以内にある場合，その吹き出しを fixation に対応した吹き出しとして特定する．以上の操作により，吹き出し単位の fixation と saccade を取得する．

3.3 特徴量抽出

本節では，吹き出し単位で切り分けられた音読音声と視線情報から推定に用いる特徴量を抽出する方法について述べる．表 1 に，提案手法で用いる 47 個の特徴量を示す．

最初に，音読音声に関する特徴量を抽出する方法について述べる．まず，各吹き出しに対して周波数特徴量である Mel Frequency Cepstral Coefficients(MFCC) を計算する．MFCC は 20 次元で求め，各次元について時間軸方向に平均を取った値を特徴量とする．次に，IBM Watson™ Speech to Text[1)]（以降，IBM Watson と呼ぶことにする）を用いてテキスト情報を考慮した音声特徴量を計算する．IBM Watson とは，英語の音声の文字起こしを行う API である．IBM Watson に各吹き出しに対応する音読音声とテキスト情報を入力すると，文字起こしの結果とその信頼度の値と，文字起こしの結果が入力したテキスト通りになったと仮定した場合の結果の信頼度の値が出力される．文字起こしの結果に対する信頼度は吹き出し単位での値と単語単位での値が得られ，文字起こしの結果が入力したテキスト通りになったと仮定した場合の信頼度は単語単位の値が得られる．特徴量としては，吹き出し単位での文字起こしの信頼度，単語単位での文字起こしの信頼度の最大値，最小値，平均値，文字起こしの結果が入力したテキスト通りになったと仮定した場合の単語単位の信頼度の最大値，最小値，平均値を用いる．最後に，吹き出し単位での音読時間と，それを吹き出し内に含まれる単語数で割ることにより求めた 1 単語当たりの音読時間を特徴量とする．

次に，視線情報に関する特徴量を抽出する方法について述べる．まず，各吹き出しに対応する fixation の数を求める．次に，fixation の持続時間，saccade の長さ，saccade の速さそれぞれについて，最大値，最小値，平均値を吹き出しごとに計算し，特徴量とする．最後に，吹き出し単位の注視時間と，それを吹き出し内に含まれる単語数で割ることにより求めた 1 単語当たりの注視時間を特徴量とする．

次に，音読音声と視線情報の両方を用いた特徴量について述べる．音声の音読時間と視線の注視時間について，それらの開始時刻の時間差と終了時刻の時間差をそれぞれ求め，特徴量とする．

最後に，テキスト情報に関する特徴量について述べる．各吹き出しに含まれる単語数と，単語出現頻度の最大値，最小値，平均値を求め，特徴量とする．

3.4 識別

本節では，識別器の学習と推定の方法について述べる．まず，抽出した特徴量と，吹き出しに未知語が含まれているか否かの 2 値ラベルを識別器に入力し，識別器を学習させる．その後，各吹き出しに対して，その吹き出しに未知語が含まれているか否かの 2 値分類問題として推定を行う．

識別器には，サポートベクターマシン (SVM) を用いる．カーネルは線形カーネル，損失関数はヒンジ損失，正則化パラメータ C の値は 1.0 とする．各特徴量は，識別器に入力する前に標準化を施してスケーリングする．PR 曲線の AUC である AUPR の値が高くなるように Sequential Backward Floating Selection(SBFS) [10] での特徴量選択を行い，選択された特徴量のみを識別器の学習と推定両方に用いる．SBFS は，一般的な Backward Selection に一度削除した特徴量を復帰させる操作を組み入れた方式である．その操作を組み入れることにより，一般的な Backward Selection よりも多くの特徴量の組み合わせを検証できる．

以上の方法により吹き出しに未知語が含まれているか否かを，音読音声とテキスト情報を用いる場合，視線情

1) https://www.ibm.com/jp-ja/cloud/watson-speech-to-text

表 1　推定に用いる特徴量

用いたセンサ情報	特徴量の種類	特徴量
音読音声	周波数特徴量	MFCC(20 次元) の各次元における時間方向の平均
	IBM Watson を用いた特徴量	吹き出し単位での信頼度 単語単位での信頼度の最大，最小，平均 文字起こしがテキスト通りになった場合の信頼度の最大，最小，平均
	音読時間	1 吹き出し当たりの音読時間 1 単語当たりの音読時間
視線情報	fixation	fixation の数 持続時間の最大，最小，平均
	saccade	その長さの最大，最小，平均 その速さの最大，最小，平均
	注視時間	1 吹き出し当たりの注視時間 1 単語当たりの注視時間
音読音声と視線情報	音読時刻と注視時刻の時間差	開始時刻の時間差 終了時刻の時間差
テキスト情報	単語数	吹き出し内の単語数
	単語出現頻度	単語出現頻度の最大，最小，平均

図 1　データ計測実験で英語漫画を音読しているときの様子

報とテキスト情報を用いる場合，音読音声と視線情報とテキスト情報を用いる場合の 3 パターンで推定を行う．

4　データ計測実験

本章では，実施したデータ計測実験について詳しく述べる．ここではデータ計測の方法や実験で取得できたデータ数について述べることにし，取得したデータを用いた手法の評価は次の章で述べることにする．

4.1　データ計測実験の概要

実験参加者には，PC の漫画表示アプリケーションを用いて英語漫画の音読を行ってもらった．使用した漫画表示アプリケーションでは，英語漫画の見開き 2 ページ分を 1 枚の画像として表示した．表示した画像のサイズは 2732 × 2048 であった．音読学習時には，音読音声と視線情報をヘッドセットとアイトラッカを用いて記録した．ヘッドセットは Jabra Evolve 30 II UC Stereo，アイトラッカは Tobii pro 4C を用いた．そして，1 エピソード音読するごとにそのエピソードに含まれている未知語を記録してもらうことで，吹き出しごとに未知語が含まれているか否かの正解ラベルを得た．実験参加者数は日本人大学生 20 名で，それぞれの実験参加者について 1 回 45 分の計測を 2 回行った．ガイダンスや計測後のアンケートなどを含めた総実験時間は 2 時間で，実験参加者 1 人当たり 2,000 円の謝礼を支払った．

4.2　データ計測実験の流れ

データ計測前に，実験参加者には英語漫画中の文字が画面に近づかなくても読めるような位置にモニターと自身の姿勢を調整してもらい，安定した視線情報を計測できるように，データ計測時にはそこから動かないようにしてもらった．次に，ヘッドセットとアイトラッカの設定を行った．その後，データ計測を開始し，実験参加者に英語漫画を音読してもらった．1 エピソード音読するごとにそのエピソードの中に含まれている未知語を記録してもらうことで，吹き出しごとに未知語が含まれているか否かの 2 値ラベルを得た．データ計測は，45 分の計測を 2 回行った．図 1 はデータ計測実験で英語漫画を音読している時の様子を示した写真である．

4.3　データ計測実験で得られたデータ数

表 2 に，実験により得られた，対応する視線情報と音読音声が得られた吹き出しのデータ数を示す．表 2 では，未知語が含まれていた吹き出し数を「未知語有」，未知語が含まれていなかった吹き出し数を「未知語無」，データが得られた吹き出しの総数を「合計」，吹き出しの総数に占める未知語有の吹き出し数の占める割合を「未知語有の占める割合」と表記している．データが得られた吹き出し数は平均 334.9，標準偏差 93.9 で，実験参加者によって大きく異なった．また，データが得られた吹き出しの総数に占める未知語有の吹き出し数の占める割合は最小で 8.6%，最大で 27.6%であった．すなわち，全ての実験参加者において「未知語有」の吹き出しは 50%以下を大きく下回っており，「未知語有」と「未知語無」のデータ数に大きな偏りがあった．

5　評価

5.1　評価方法

ユーザ依存の交差検証を行うことで提案手法を評価した．交差検証の方法としては，1 エピソード分のデータをテストデータとし，残りを学習データとする leave-one-episode-out 交差検証を行った．ラベルに大きく偏りがあったため，学習データには SMOTE [11] を用いたオーバーサンプリングを行った．評価値としては，Precision-Recall 曲線（PR 曲線）の AUC である AUPR を用いた．ベースラインは表 1 に示したテキスト情報に関する特徴量のみを用いた推定とした．本稿ではこれ以降，音読音声とテキスト情報を用いた推定を「音声 + テキスト」，視線情報とテキスト情報を用いた推定を「視線 + テキスト」，音読音声と視線情報とテキスト情報を用いた推定を「音声 + 視線 + テキスト」と表記することにする．

表 2　実験で得られたデータ数

参加者番号	吹き出し数 (個) 未知語有	未知語無	合計	未知語有の占める割合
P01	40	243	283	14.1%
P02	33	277	310	10.6%
P03	67	247	314	21.3%
P04	81	366	447	18.1%
P05	22	235	257	8.6%
P06	64	348	412	15.5%
P07	53	222	275	19.3%
P08	35	323	358	9.8%
P09	87	228	315	27.6%
P10	48	316	364	13.2%
P11	36	223	259	13.9%
P12	49	196	245	20.0%
P13	24	99	123	19.5%
P14	43	485	528	8.1%
P15	25	230	255	9.8%
P16	37	414	451	8.2%
P17	62	290	352	17.6%
P18	30	305	335	9.0%
P19	35	282	317	11.0%
P20	51	446	497	10.3%
平均	46.1	288.6	334.9	

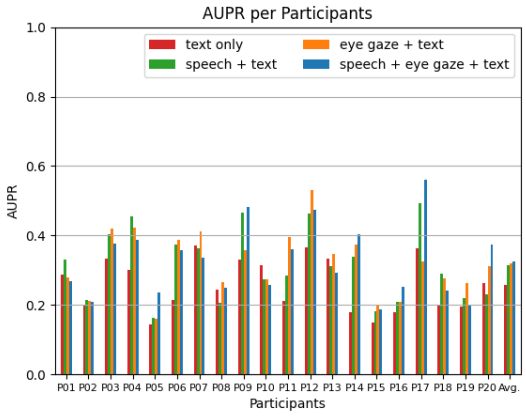

図 2　leave-one-episode-out 交差検証における実験参加者ごとの AUPR の値

5.2　結果

図 2 は，leave-one-episode-out 交差検証における提案手法の 3 種類の推定とベースラインの AUPR の値を実験参加者ごとに表したグラフである．実験参加者 20 名中，最も AUPR が高くなった推定が「音声 + 視線 + テキスト」であったのは 6 名，「視線 + テキスト」であったのは 9 名，「音声 + テキスト」であったのは 4 名，ベースラインであったのは 1 名であり，P10 を除いた全ての実験参加者において提案手法の 3 種類の推定方法のうちいずれかの AUPR 値がベースラインの AUPR 値を上回った．また，推定方法ごとの全実験参加者の AUPR の平均値は「音声 + 視線 + テキスト」が 0.325，「視線 + テキスト」が 0.321，「音声 + テキスト」が 0.314，ベースラインが 0.259 であり，提案手法の 3 種類の推定方法の全てがベースラインを上回った．

5.3　考察

P10 を除いた全ての実験参加者において提案手法の 3 種類の推定方法のうちいずれかの AUPR 値がベースラインの AUPR 値を上回ったことから，音読音声や視線情報といったセンサ情報を用いることで未知語を含んだ吹き出し推定の推定性能が向上することが分かった．また，AUPR の値が最も高かった推定方法は実験参加者によって異なっていたことから，提案手法の 3 種類の推定方法のうち最も推定性能が高い推定方法は学習者によって異なることが分かった．

6　おわりに

本稿では，英語漫画音読時の音読音声と視線情報を用いて英語漫画から英語学習者の未知語を含んだ吹き出しを推定する手法を提案した．日本人大学生 20 名分のデータを収集して手法の評価を行った結果，音読音声と視線情報を推定に用いた提案手法の方がテキスト情報のみを用いた推定よりも性能が良いことが確認でき，音読音声や視線情報は未知語を含んだ吹き出しの推定に有用であることが分かった．

参考文献

[1] Amirreza Karami and Freddie A Bowles. Which strategy promotes retention?: Intentional vocabulary learning, incidental vocabulary learning, or a mixture of both? *Australian Journal of Teacher Education (Online)*, Vol. 44, No. 9, pp. 25–43, 2019.

[2] Utpal Garain, Onkar Pandit, Olivier Augereau, Ayano Okoso, and Koichi Kise. Identification of reader specific difficult words by analyzing eye gaze and document content. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 1, pp. 1346–1351. IEEE, 2017.

[3] Kamini Sabu and Preeti Rao. Automatic prediction of confidence level from children's oral reading recordings. In *INTERSPEECH*, pp. 3141–3145, 2020.

[4] Tong Zhang, Mark Hasegawa-Johnson, and Stephen E Levinson. Children' s emotion recognition in an intelligent tutoring scenario. In *Proc. Eighth European Conf. Speech Comm. and Technology (INTERSPEECH)*, 2004.

[5] Pascual Martínez-Gómez and Akiko Aizawa. Recognition of understanding level and language skill using measurements of reading behavior. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pp. 95–104, 2014.

[6] Shoya Ishimaru, Syed Saqib Bukhari, Carina Heisel, Jochen Kuhn, and Andreas Dengel. Towards an intelligent textbook: eye gaze based attention extraction on materials for learning and instruction in physics. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct*, pp. 1041–1045, 2016.

[7] Yuki Daiku. Estimation of understanding and interest based on reading behavior in extensive reading with japanese comics translated in english. In *Master Thesis of Osaka Prefecture University*. 2019.

[8] Ryota Takahashi. Estimation of understanding based on eye information in extensive reading with japanese comics translated in english. In *Graduation Thesis of Osaka Prefecture University*. 2021.

[9] Georg Buscher, Andreas Dengel, and Ludger van Elst. Eye movements as implicit relevance feedback. In *CHI'08 extended abstracts on Human factors in computing systems*, pp. 2991–2996. 2008.

[10] Pavel Pudil, Jana Novovičová, and Josef Kittler. Floating search methods in feature selection. *Pattern recognition letters*, Vol. 15, No. 11, pp. 1119–1125, 1994.

[11] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, Vol. 16, pp. 321–357, 2002.