OCRを用いた視線情報と閲覧テキストの言語的特徴の関連付け

冨田 恭平 *　　相澤 彰子 *†　　Pascual Martinez-Gomez*
陳 琛 *　　原 忠義 †　　狩野 芳伸 ‡
* 東京大学, † 国立情報学研究所, ‡ 科学技術振興機構さきがけ

1　はじめに

人間の視線の振る舞いは興味深い研究の対象となっている。視線の振る舞いからユーザが見ている画像にどの程度興味を持っているかを調べたり [6]、文章を読む際の視線の振る舞いに関する調査も行われている [8]。また最近は、自動要約や自動分類、クエリ拡張の際にユーザの視線の振る舞いを考慮するなど [9][3][2]、視線認識と自然言語処理を組み合わせた研究も盛んになってきている。煩わしいフィードバックを必要とせず、個々のユーザに特化した自然言語処理が可能となることが、視線認識と自然言語処理を組み合わせた際の強みであろう。

視線情報と自然言語を組み合わせる際に、ユーザが見ている画面上の座標と、その座標に表示されている語をいかにして関連付けるかという問題が生じる。認知モデルの検証実験など、特別な状況下でのみ視線情報と閲覧テキストを関連付けられれば良い場合は、人手での対応付け [8] や専用ソフトウェアの開発 [9] などで対応可能であるが、インタフェースの一つとして視線認識を用いる場合には一般的な状況下での対応付けが求められる。そのような対応策として、OCR を用いた手法 [3] や Text 2.0 というフレームワーク [1] などが提案されている。

視線情報と文脈に依存する語の言語的特徴 (例えば品詞など) を組み合わせて利用することができたら、寺ら [8] が行ったような実験を自動で行うことができたり、自然言語処理への応用の幅が大きく広がるだろう。言語的特徴を利用するためには、各単語に事前に言語的特徴をアノテートしておき、ユーザが見ている単語を特定したあと言語的特徴を参照する必要がある。しかし、先述の手法では事前の言語的特徴のアノテーションは行われておらず、また日本語のように単語の境目が明示的でない言語の場合はユーザが読んでいる文字が分かったとしても単語を特定することは難しいので、視線情報と言語的特徴を組み合わせて利用することは容易ではない。

このような観点から、本論文では、OCR を用いることで任意のテキストに対して、画面上で各語が表示されている位置とその言語的特徴を特定し、テキストの言語的特徴とユーザの視線の振る舞いを組み合わせた処理を容易に行うための手法を提案する。また、実際に提案手法を使い再読に関する実験を行い、閲覧テキストの言語的特徴と視線情報を容易に組み合わせることが可能となったことを示す。

2　関連研究

近年、視線情報と閲覧テキストを用いた研究が盛んに行われている。Buscher ら [2] は、テキストを読んだあとにクエリ拡張を行う際、テキスト中の語の tf-idf のスコアと視線情報を組み合わせて用いることで、tf-idf のスコアのみを用いるよりも良いクエリ拡張を行なうことに成功している。またテキストの自動分類を行う際に、複数のトピックを含むテキストの分類方法をユーザの視線の振る舞いから学習する方法なども提案されている [3]。Xu ら [9] はテキストの自動要約を行う際に、従来の手法に加えて視線情報からユーザの各文への興味の度合いを予測して用いることで、従来手法よりも良い結果を残している。Biedert ら [1] が考案した Text 2.0 というフレームワークは、HTML, CSS, JavaScript を用いて、読まれた箇所の文字色を変えるなど視線を使ったインタラクティブなブラウジングを可能にしている。

これらの研究では画面上の語の位置の特定に、専用のドキュメントブラウザを用意する、HTML, CSS, JavaScript 等を用いて語の位置を特定するなど様々な手法を用いているが、その中で OCR を用いた手法が存在する [4]。この手法では、ユーザが読んでいるテキストの行を検知し、その行の周辺のスクリーンショットをとり、OCR を用いてテキストの内容を得る。認識結果には間違いが含まれることもあるが、認識した行と元のテキストをマッチングすることにより、ユー

ザが読んでいる行を正確に得ることができる。

語の位置を指定するような手法が使えるのは、テキストごとに語の位置を指定することができる特別な状況下のみであり、Text 2.0 や OCR を用いた手法に関しても、既存の言語解析ツールの出力を言語的特徴として視線情報と関連付けて利用することは難しい。よって、視線情報と言語的特徴を組み合わせて利用したい場合は、視線情報と言語的特徴をどのように関連付けるかが問題となる。この問題に対し、本論文では、OCR の手法を発展させ、言語解析ツールによりテキストを単語へ分割し、言語的特徴をアノテートしたテキストと OCR の認識結果をマッチングすることにより、各語の画面上での位置とその言語的特徴を特定する手法を提案する。この手法を用いると、単語の区切りが明確でない日本語に対しても語の位置が特定できるようになる。

3 提案手法

3.1 取り扱う問題

はじめに、「単語情報」という用語を、単語の字面、単語の言語的特徴、そして単語の Bounding Box の組であると定義する。ただし、本論文では単語の位置の指定にその Bounding Box(単語を囲む長方形) を用いることとする。視線情報と閲覧テキストの言語的特徴を組み合わせた利用を容易にするためには、閲覧テキストに含まれる単語の単語情報を正しく生成すればよいので、それをここでの目的とする。

本論文ではテキストは日本語で書かれたものを対象とし、言語的特徴としては品詞を用いる。また、ユーザが見ている画面のスクリーンショットと、閲覧テキストを入力として使用可能であると仮定する。

3.2 全体の流れ

閲覧テキストを言語処理ツールを用いて単語ごとに分割し、各語の言語的特徴をアノテートしておく (「フルテキスト」とする)。次にスクリーンショットに OCR を適用することで、画面上に表示されている文字列と、その各文字の Bounding Box を得る (「認識テキスト」とする)。ただし、認識テキストには間違いが含まれることが多いので、この間違いを修正する必要がある。認識テキストとフルテキストをマッチングすることによりこの修正を行い、更に両テキストの単語ごとの対応関係を得る。単語ごとの対応関係がわかったら必要な情報はすべて揃うので、単語情報を生成するのは容易である。マッチングの詳細は次節で述べる。

3.3 マッチング

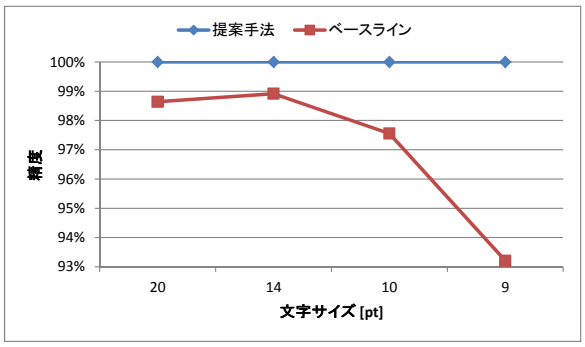
認識テキストとフルテキストの語の対応関係を取るために、まずは認識テキストの各行に対応する部分をフルテキストから探す。つまり、ある認識テキストの行 $line$ に対して、$levenshtein(line, str)$ が最小となるようなフルテキストの部分文字列 str を探し、str を $line$ に対応する行であるとする。ただし、$levenshtein(\cdot,\cdot)$ は編集距離を求める関数である。このように対応する部分文字列を探す問題は近似文字列マッチングとして知られており、編集距離を求める動的計画法を利用することにより求めることができる [7]。

この方法で OCR が認識した文字は閲覧テキストに含まれていない行やノイズであってもフルテキストのどこかに必ずマッチングしてしまい、意図しない単語情報が生成されてしまう。そこで、上述の編集距離を $line$ の長さで正規化した値がある閾値以上である場合にはマッチングは失敗したとして、その部分の単語情報を生成しないこととする。

行ごとの対応関係がわかったら、動的計画法のテーブルを逆にたどることによって文字ごとの対応関係を取る。この際、フルテキスト中の文字と認識テキスト中の文字が 1 対 1 に対応せず、単語ごとの対応が取れない場合があるが、その場合は文字数、文字が全角か半角かなどの情報を用いて文字の幅を推定し、単語情報を生成する。

3.4 実装と評価

我々は提案手法を実際に実装し、単語情報がどの程度正確に生成されているかを評価した。実装の際、OCR はメディアドライブ株式会社の活字文書 OCR ライブラリを、テキストを単語に区切り品詞を得る言語解析ツールとして MeCab[5] を使用した。また、実装言語は OCR で文字を認識する部分は C 言語、それ以外の部分は Python を用いた。3.3 節で述べたマッチングの失敗については、予備実験の結果より正規化した編集距離が 0.5 以上の場合にマッチングが失敗することとした。図 1 に、提案手法で計算した単語情報の Bounding Box を、スクリーンショット上に描画した例を示す。赤い四角形が Bounding Box を意味しており、各単語が Bounding Box 内に綺麗に収まっていることが確認できる。



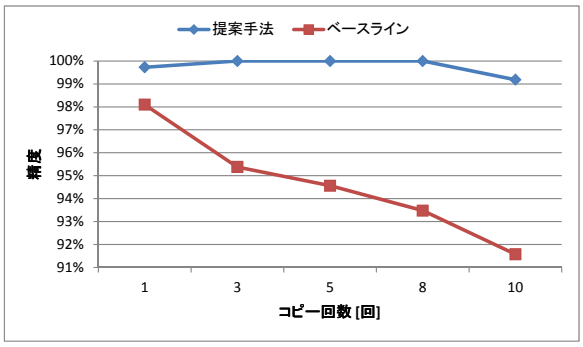
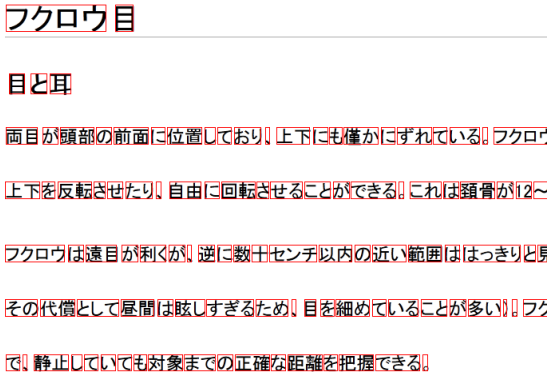


図 2: (左) 文字サイズと単語情報の生成精度の関係。 (右) コピー回数と単語情報の生成精度の関係 。ベースラインは、OCR で得られた認識結果をそのまま言語解析ツールの入力として単語情報を生成した場合の精度。


フクロウ目

目と耳

両目が頭部の前面に位置しており、上下にも僅かにずれている。フクロ

上下を反転させたり、自由に回転させることができる。これは頸骨が12～

フクロウは遠目が利くが、逆に数十センチ以内の近い範囲ははっきりと見

その代償として昼間は眩しすぎるため、目を細めていることが多い。フク

で、静止していても対象までの正確な距離を把握できる。

図 1: Bounding Box の例。テキストは日本語版 Wikipedia の「フクロウ目」のページ (http://ja.wikipedia.org/wiki/フクロウ目) を使用している。

単語情報の生成精度は OCR のの認識精度に大きく依存するため、文字サイズを変更する、コピーを繰り返してノイズを入れるという 2 種類の方法で OCR の認識精度を下げて単語情報の生成精度を評価した。文字のサイズは 9pt, 10pt, 14pt, 20pt の 4 種類を用い、コピーは 1 回, 3 回, 5 回, 8 回, 10 回の繰り返しで評価した。

単語情報の生成精度は以下の式で計算した。

$$\frac{\text{(正しい単語情報が生成された単語数)}}{\text{(テキストの単語数)}}$$

ただし、単語情報が正しく生成されているとは、単語の字面と言語的特徴がフルテキストと一致し、さらに Bounding Box 内に単語が正しく収まっている場合を言う。今回は Bounding Box 内に単語が収まっているかどうかのチェックは人手で行った。

マッチングを用いることでどの程度単語情報の生成精度が向上したかを見るために、マッチングを行わずに認識テキストから単語情報を生成した場合ををベースラインとして提案手法と比較する。図 2 は文字サイズ、コピー回数を変化させて評価した結果を示している。文字サイズを変えた評価では、ベースラインは 93.2～98.6%の精度であったが、提案手法ではいずれの文字サイズに対しても 100%の精度で単語情報を生成していることがわかる。また、コピー回数を変えた評価では、ベースラインは 91.5～98.1%の精度であるのに対し、提案手法ではそれを大きく上回る 99.1～100%という精度で単語情報を生成できていた。

4 実験

4.1 実験設定

提案手法が実際に機能することを示すため、再読する際の視線情報を集め、その視線情報と言語的特徴（本論文の場合は品詞）を組み合わせた解析を行う。再読とは一度読んだテキストをもう一度読み直すことであり、本実験では 1 度目、2 度目にテキストを読む際に視線の振る舞いがどのように異なるかを調べる。

本実験ではタスクを 2 つ用意した。タスク 1 は (1) テキストを読む、(2) テキスト再度読む、(3) 質問に答える、という流れで、タスク 2 は (1) テキストを読む、(2) 質問に答える、(3) テキストを再読する、(4) 質問に再度答える、という流れである。質問を用意した一つ目の意図は被験者にテキストを集中して読んでもらうためであり、もう一つの意図は質問の回答を探すための再読では視線の振る舞いがどのように変化するかを観察するためである。

使用したテキストは、日本語版 Wikipedia の「フクロウ目」のページと、「インフルエンザ」のページの一部を使用した。被験者を 2 つのグループに分け、タスク 1 とタスク 2 で使用するテキストを入れ替えて実験を行った。実験には 12 人に協力してもらい、そのうちデータにノイズが多く解析が困難だった 2 人分のデータを除き、各グループ 5 人分、計 10 人分のデータを解析に使用した。

視線情報の計測には Tobii TX300(解像度が 1920 x 1080 の 23 インチスクリーン、サンプリングレートは 300Hz) を使用した。視線認識装置の誤差を軽減するために、Wikipedia の記事のフォントサイズと行間を変更したものをウェブブラウザに表示して実験を行った。

品詞に関する分析を行う際、名詞に関してはに一般名詞、固有名詞等、より細分化された分類を使用した。また、いくつかの品詞をまとめた内容語 (名詞、動詞、形容詞、副詞) と機能語 (内容語に含まれない品詞) という観点でも分析を行った。

4.2 実験結果

図 3 に、内容語、機能語、一般名詞、動詞、そして助詞における読まれた語の割合を示す。まず、1 回目のリーディング (タスク 1/1 回目およびタスク 2/1 回目) に注目すると、一般名詞は 7 割程度読まれていることがわかる。これに比べ、助詞などの機能語は読まれている割合が低く、機能語全体では 3 割程度しか読まれていない。また、タスク 1 での 1 回目と 2 回目のリーディングを比較してみると、各品詞とも読まれている割合が低くなっている。読まれている割合は一般名詞が 15%程減少しており、主要な品詞の中では一番の減少幅であった。1 回目と 2 回目のリーディングの間に質問を挟んだ場合のタスク 2 の結果を見てみても、各品詞とも読まれている割合が低くなっていることがわかる。グラフを見てもわかるように、主要な品詞についてはタスク 1 の 2 回目のリーディングよりも、タスク 2 での 2 回目のリーディングのほうが読まれている語の割合が低くなっている。これは、1 度目のテキストを読み、質問に回答した後、もう一度テキストを読むことになるタスク 2 では、再読の際に質問の答を探すことに力を注ぐような読み方になり、その分質問の答とは関係のない箇所については多くの語を読み飛ばしているからではないかと考えられる。

このように提案手法を用いて視線情報と言語的特徴を組み合わせた実験・解析を行い、各品詞がどの程度読まれているかという、従来では得ることが難しいような結果を得ることができた。

5 おわりに

本論文では、OCR やマッチングなどの技術を用いて、視線情報と閲覧テキストの言語的特徴を関連付けて利用するための手法を提案した。提案手法にて特定される画面上の語とその言語的特徴の精度も、我々の

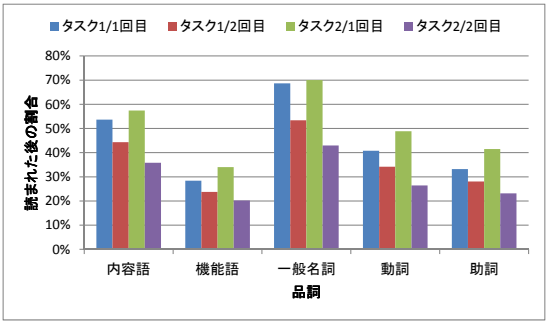


図 3: 各品詞における、読まれた語の割合（主要な品詞のみ表示)。再読時には、読まれた割合が低くなっていることがわかる。

検証した環境では 99%〜100%と非常に高い精度で正しく関連付けられていることを示した。また、提案手法を用いた再読の実験では品詞ごとの読まれた語の割合を示し、従来の手法では自動で行うのが難しかった解析を自動で行うことに成功した。

今回我々は日本語のテキストに大してのみ実装・評価を行ったが、英語など他の言語についても提案手法が正しく動作するかの検証も行い、正しい動作が確認されれば応用の幅は更に広がるだろう。また今回は OCR の認識精度を上げるためにテキストの背景がない環境で実験を行ったが、背景がある場合の精度の検証や、その際の精度の向上方法なども検討するとより実用的になるだろう。

参考文献

[1] R. Biedert, G. Buscher, S. Schwarz, M. Möller, A. Dengel, and T. Lottermann. The text 2.0 framework. In *Proc. of the International Workshop on Eye Gaze in Intelligent Human Machine Interaction*, 2010.

[2] G. Buscher, A. Dangel, and L. van Elst. Query expansion using gaze-based feedback on the subdocument level. In *SIGIR'08*, pp. 387–394, 2008.

[3] G. Buscher and A. Dengel. Attention-based document classifier learning. In *Document Analysis System'08*, pp. 87–94, 2008.

[4] G. Buscher, A. Dengel, L. van Elst, and F. Mittag. Generating and using gaze-based document annotations. In *Proceedings and Extended Abstracts of the Conference on Human Factors in Computing Systems*, pp. 3045–3050. ACM, 2008.

[5] T. Kudo. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.net/*, 2005.

[6] S. N. Haji Mirza and E.Izquierdo. Finding the user's interest level from their eyes. *SAPMIA '10*, pp. 25–28, 2010.

[7] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, Vol. 33, p. 2001, 1999.

[8] A. Tera, K. Shirai, T. Yuizono, and K. Sugiyama. Analysis of eye movements and linguistic boundaries in a text for the investigation of japanese reading processes. *IEICE TRANSACTIONS on Information and Systems*, Vol. E91-D, pp. 2560–2567, 2008.

[9] S. Xu, H. Jiang, and F. C. Lau. User-oriented document summarization through vision-based eye-tracking. In *IUI '09*, pp. 7–16, 2009.