様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１９　日現在

機関番号：　１２６０８
研究種目：　挑戦的萌芽研究
研究期間：　2014～2016
課題番号：　２６５４０１１３

研究課題名（和文）計算言語学的手法を利用した人間の単語認識における定量的法則の発見

研究課題名（英文）Computational approach for finding a quantitative law on human word recognition

研究代表者
高村　大也（Takamura, Hiroya）

東京工業大学・科学技術創成研究院・准教授

研究者番号：８０３６１７７３

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：編集距離、文字n-gramベクトルの余弦距離、文字列カーネルなど、単語間の表層的類似度を算出するプログラムを作成した。
また、新聞記事データで学習した言語モデルを用いて、ある文脈において次の単語が持つ情報量を算出するプログラムを作成した。さらに実験で使う読み時間測定プログラムを作成した。学内の倫理審査を経て、実際に人間の読み時間の測定を行った。その際に、文内の一部の単語の綴りを意図的に変更した。20人程度の実験データから、読み時間と、単語の綴りの特徴、文脈から単語が予測できる度合いの間の統計的傾向が得られらた。

研究成果の概要（英文）：We wrote computer programs : (i) a program measuring the similarity between words such as edit distance, cosine similarity of character n-gram vectors, and string kernel, (ii) a program measuring the information amount that a word has given its preceding context using the language model trained on news paper articles, (iii) a program that is used in the experiment to measure the reading time of human subjects. With an ethical approval, we conducted experiments to measure reading time, where the spellings of some words are intentionally changed. From the experimental data, we derived the statistical trend among reading time, spelling, and the information amount of a word given its preceding context.

研究分野：　自然言語処理

キーワード：　単語認識　読み時間

１．研究開始当初の背景

「こんちには　みさなん　おんげき　ですか？」という文は、いくつかの単語の綴りが不正確であるにも関わらず、「こんにちは　みなさん　おげんき　ですか？」と人間は理解できてしまう。Rayner らは、この人間の情報処理能力を学術的に確認しようとした[Rayner et al., 2006]。両端以外の文字を無作為に置換した文字列を含む文は、文字列全体を無作為に置換した文字列を含む文よりも、読み時間が短くなることを被験者実験によって確認し、両端の文字を固定した場合の処理負荷が軽いことを示唆する結果を得た。Rayner ら[1]の研究は一定の成果を得たものの、依然として多くの疑問が残る。４文字から成る単語の両端を固定すれば、内側の隣り合った文字を置換するのみとなるが(「みさなん」)、これは元の文字列(「みなさん」)との表層的類似度が高いので、読み時間の増加を抑制する要因が、両端の固定にあるのか、それとも高い類似度にあるのかは不明である。本研究課題の目的は、どのような文脈において、文字列のどの部分にどのような操作をすると、人間の単語認識にどの程度影響するのか、という疑問に対する答えを見つけることである。そのために、計算言語学的な手法が必要である。

[Rayner et al., 2006] Rayner et al., “Raeding wrods with jubmled lettres: There is a cost”, Psychological Science 17, 192-193, 2006.

２．研究の目的

人間はテキストを読む際、単語が正しい綴りで書かれていなくても、その単語を認識できる場合が多くある。しかし、本来の綴りと非常に大きく異なっている場合は、正しく認識できない。また、文脈により単語が予測できる場合は、認識が容易になると考えられる。本研究課題では、正しい単語と綴りの間違った単語の表層的類似度と、文脈から単語が予測できる度合いの二つの要素を計算言語学的な手法を用いて定量化し、それらが人間の単語認識における負荷に与える影響について定量的な法則を発見することを目的とする。人間の単語認識における負荷は、文の読み時間を用いて測定する。その知見を元に、正しくない綴りの単語が含まれているにも関わらず人間が正しく読めてしまうような文の生成アルゴリズムを考案する。

３．研究の方法

まずは、使用するコンピュータ・プログラムの作成を行う。次に、被験者実験を行う。被験者には、コンピュータの画面に出てくる文を読んでもらい、その読み時間を測定する。読み時間が人間の処理負荷と高い相関を持つことは知られており[Miyamoto, 1998]、アイトラッキングと比較して実験のコストが低いことからここでは読み時間を測る。文の中のある単語になんらかの置換操作が加えられているとき、その置換操作の有無や種類によって、あるいは文脈によって読み時間がどのように変化するかを調べる。測定データを分析し、どの要因がどの程度読み時間に影響を与えるかを推定する。次に、読み時間の増加が小さいような設定において、置換などの操作に人間が気付くかどうかを詳細に実験によって調べる。最後に、与えられた文に対して、人間に処理負荷の多大な増加を与えない範囲で単語に変更操作を加えるアルゴリズムを構築する。

[Miyamoto, 2008] Edson T. Miyamoto, “Processing sentences in Japanese”, The Oxford Handbook of Japanese Linguistics, pp. 217-149, 2008.

４．研究成果

まずは、必要となるコンピュータ・プログラムの作成を行った：

・編集距離、各単語の文字 n-gram ベクトルの余弦距離、文字列カーネル関数による距離など、様々な文字列間の類似度を算出するプログラム

・ある文脈において次の単語が持つ情報量を算出するプログラム(新聞記事で学習した言語モデルを利用した)。現在、最も精度が高いとされているニューラル言語モデルを使用した。

・被験者実験で使う読み時間測定プログラム(これについては以前の研究で利用した

ものを今回の実験用に変更した)。これは、モニタに日本語の文を提示し，文節ごとに被験者が読んだ読み時間を測定するものである。測定は通常のコンピュータ上で、スペースキーを押すと日本語文の文節が前から１つずつ表示される仕組みを用い、そのときスペースキーを押した時刻が記録されるという仕組みである。文の提示画面の例を次に示す。

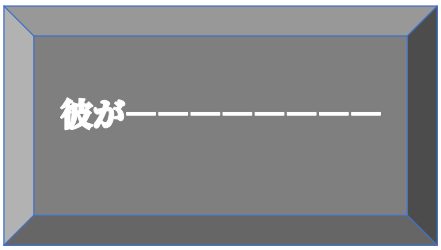

図. 文の提示画面の例

ただし、「3.研究の方法」で述べたように、一部の単語については、意図的に綴りに誤りを含め、その誤りが読み時間にどのように影響を与えるかを考察する。例えば、「私は 昼食に ヤソキバを 食べました」のような文が表示される。この「ヤソキバ」には綴り誤りが含められている。また、綴り誤りは、単語の先頭部分、中央部分、語尾部分に含めている。これにより、綴り誤りの場所が読み時間に与える影響についても調査を行う。

学内の倫理審査を経て、実際に人間の読み時間の測定を行った。20 人程度の実験データが得られ、どのような箇所で人間が文の読解に時間がかかるのかがわかるようになった。このデータから読み時間と、単語の綴りの特徴、文脈から単語が予測できる度合いの間の統計的傾向が得られた。

平均読み時間は次の通りである：

綴りが正しい単語: 617
単語の先頭に綴り誤りを含めたもの: 1,104
単語の中央に綴り誤りを含めたもの: 1,105
語尾部分に綴り誤りを含めたもの: 936

ここで、単位はミリ秒である。綴りが正しい単語とそうでない単語の平均読み時間の間には差があり、綴りの不正確さが読み時間を増加させていること、すなわち単語認識の負荷を高めていることがわかる。また、綴り誤りを含むものの中では、単語の先頭と中央に綴り誤りを含むものは、ほぼ同程度の平均読み時間であった。これに対し、語尾部分に綴り誤りを含むものは、読み時間が少なく、認識負荷が比較的小さいことが示唆されている。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　0　件）

〔学会発表〕（計　0　件）

〔図書〕（計　0　件）

〔産業財産権〕

○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
高村 大也　（TAKAMURA, Hiroya）
東京工業大学・科学技術創成研究院・准教授
研究者番号：80361773

(2)研究分担者
なし

(3)連携研究者
なし

(4)研究協力者
なし