平 成 30 年 度　卒 業 研 究 論 文

題目　Twitter トレンドを始点としたユーザの関心を煽る雑学探索に関する研究

指導教員　　服部 峻

提出者　　室蘭工業大学 情報電子工学系学科

氏　名　　松田　純哉

学籍番号　　15024160

提出年月日　　平成 31 年 2 月 13 日

目次

図目次

表目次

第 1 章

まえがき

「月面着陸に成功したアポロ 11 号のコンピュータの性能はファミコン以下」,「江戸時代にはオナラをした人の身代わりになる役職があった」などといった知って得するかどうかを考えない雑多な面白い知識のことを雑学と言う．雑学は，一般的にはくだらないものとされ，日常生活を送る上でそこまで重要となる知識ではない為，調べるほど価値のある情報とはされず，実際に調べる人はかなり少数である．もし調べるとしても，Web 上の非常に限定的な部分に存在する雑学のみを調べることで満足し，より幅広いカテゴリの雑学や専門的な雑学を調べるまでには至らないケースがほとんどであろう．また,「テレビ」や「雑誌」などの限られたメディアから知る程度で，日常生活上で雑学を知る機会というのは限られている．このように，日常生活を送る上で活用できる知識ではなく，あまり知る機会というのは多くないが，人と人との会話においては，ある話題に沿ってそのような雑学を披露できることから，会話を盛り上げる題材として十分に活用可能である．また，雑学の中には知識という性質上，実際に役立つものも少なからず含まれており，一概に全てが無駄な知識とは言えない為，雑学を知るということは有意義なことであると言える．2000 年代中期には，視聴者から投稿された雑学を紹介する「トリビアの泉 〜素晴らしきムダ知識〜」というテレビ番組が放送されており，ゴールデンタイムでのレギュラー放送が平均視聴率 17.8% を獲得し，それが 2003 年以降に放送を開始したバラエティ番組では 1 位であるという調査 [1] からも分かるように，雑学の娯楽としてのポテンシャルには非常に大きな可能性が秘められていると考えられる．ここで，その知っておいて損はない有用な雑学というものを，まだまだ未知の雑学が眠っているであろう Web という情報が膨大に蓄積されているデータの中から探し出し，提供できないかと考えた．しかし，現時点では，数多くの雑学が眠っているであろう Web 上から，雑学を探し出し面白さを測る明確な方法は存在していない．また，雑学をテキスト処理する上では「雑多な知識ゆえの雑学というカテゴリの曖昧さ」,「雑学の面白さの不明瞭さ」などが大きな問題となる．

そこで本研究では，雑学のテキスト特徴の分析から立案した仮説によって，これらの問題の解決を試み，Web 上から雑学を探索，及び評価することで，より多くの人が関心を持つ面白い雑学を提供することを目指す．

第 2 章

関連研究

雑学に対する考え方や，ある雑学について紹介するような論文はいくつか存在するものの，雑学のテキスト特徴の分析などといった雑学というカテゴリにフォーカスした研究は行われておらず，雑学を Web 上から自動で収集するような研究も行われていない．その為，雑学のテキスト特徴や面白さに対する分析などは新たに行わなければならないが，本研究の目的を達成するうえで大きな問題の 1 つである「雑学かどうかの判定」に関わるテキスト分類についての研究は，昨今の偽情報によって起こる被害拡大防止の流れもあり，数多く行われている．佐々木ら [2] は，文書クラスタリング手法を用いたスパムメール判定手法を提案している．機械学習によってスパムメールであるか非スパムメールであるかをラベリングし，スパムメール判定を行う手法は，同じように雑学の文であるかそうでないかを判定する上で有用であると考えられる．また，同じく本研究の目的を達成するうえで大きな問題の 1 つである「雑学の面白さの評価」に関わるテキストの面白さを評価する研究も少なからず行われていた．天谷ら [3] は，テキストの面白さを評価すべくユーモアの有無に着目し，ユーモアの認識を Naive Bayes と SVM のハイブリッドで行う手法を提案している．本研究とテキストの面白さを評価するという点では一致しているが，この手法における判定対象が話者の体験談や近況，あるいは知識に基づくストーリー型のテキストであり，特に文章量や文章表現という点において，雑学のテキスト特徴との乖離が大きい．また，雑学の面白さにユーモアが関係しているというケースは比較的少数なこともあり，この手法は本研究に対してあまり有用ではないと考えられる．よって，本研究では雑学文とノイズ文といったようにクラスタリングによる分類によって雑学の抽出を行い，新たに雑学の面白さに繋がるファクターを発見すべく，雑学の面白さを分析する必要がある．

第 3 章

提案システム

本章では，最終的な目標である Web 上から自動的に雑学を集め，面白さ順にランキング表示することを達成する為の雑学探索システムについて提案する．まず，提案システムの対象である雑学の特徴に関して整理した後，システム全体の概要，及び各処理の詳細について述べる．

3.1　雑学の定義

本研究では，雑学を研究対象としている為，まず雑学とはどのようなテキストなのかをはっきりさせる必要がある．雑学は，多岐のジャンルにわたる系統立っていない様々な事柄についての知識であり，その大きな特徴として，知識としての面白さを重視しているものとされている．ここから「知識」「面白い」という 2 つの要素が雑学には必要であるとし，これら 2 つの要素を満たしたテキストを，本研究では雑学として取り扱うこととする．

3.2　雑学のパターン化

雑学は，図 3.1 のように大きく分けて 2 つのパターンに分けられる．まず 1 つ目のパターンとして，「タヌキ寝入りを英語で言うと Fox sleep（キツネ寝入り）となる.」や「和菓子のコンペイトウは，角が 24 個あるのが良品とされている.」などといった 1 つの文で成立する雑学（以下，単文成立型雑学と呼ぶ）が挙げられる．この単文成立型雑学は 1 つの文のみで意味が伝わり，情報量が少ないもののインパクトが強いものが多い．

一方で，知識の文には大抵主語が存在しているが，指示語などによって「どこで」「何が」といった情報が文中に存在しない場合がある．具体例で言うと「また，防水にするには，コストがかかるとともに完全密封になる為，熱がこもりやすくなり壊れやすくなることがある.」といった文では「何が」に当たる情報が欠損しており，この文だけではどういった知識なのか意味が伝わらない．この文は前文の「日本産のケータイには防水が多いが，海外では防水ケータイ，スマートフォンはほとんど存在しない.」という文があって初めて意味が伝わる．これが 2 つ目のパターンの複数の文が揃うことで初めて意味が伝わる雑学（以下，複文成立型雑学と呼ぶ）である．この複文成立型雑学は，ある人物に対するエピソード中の意外な一面や，ベー

スとなる文の補足雑学として存在している傾向があり，多くは前の文章から引用する為に，文の始まりに「それ」や「これ」などといった指示語が使われている．ただし，複文成立型雑学は，その名の通り複文であるが故に情報量が多く，読むことだけでユーザに負荷がかかる．また，文と文の関係が複雑化しているものであれば，意味を理解できるまで時間を要することになる．そういった意味で複文成立型雑学は単文成立型雑学に比べストレスがかかりやすく，雑学としての面白さ，分かりやすさという点で単文成立型雑学の方が優れていると考える．

以上より本研究では，雑学探索の対象として単文成立型雑学を中心に Web 上から雑学を抽出し，面白さ順にランキングして提供することを目指す．

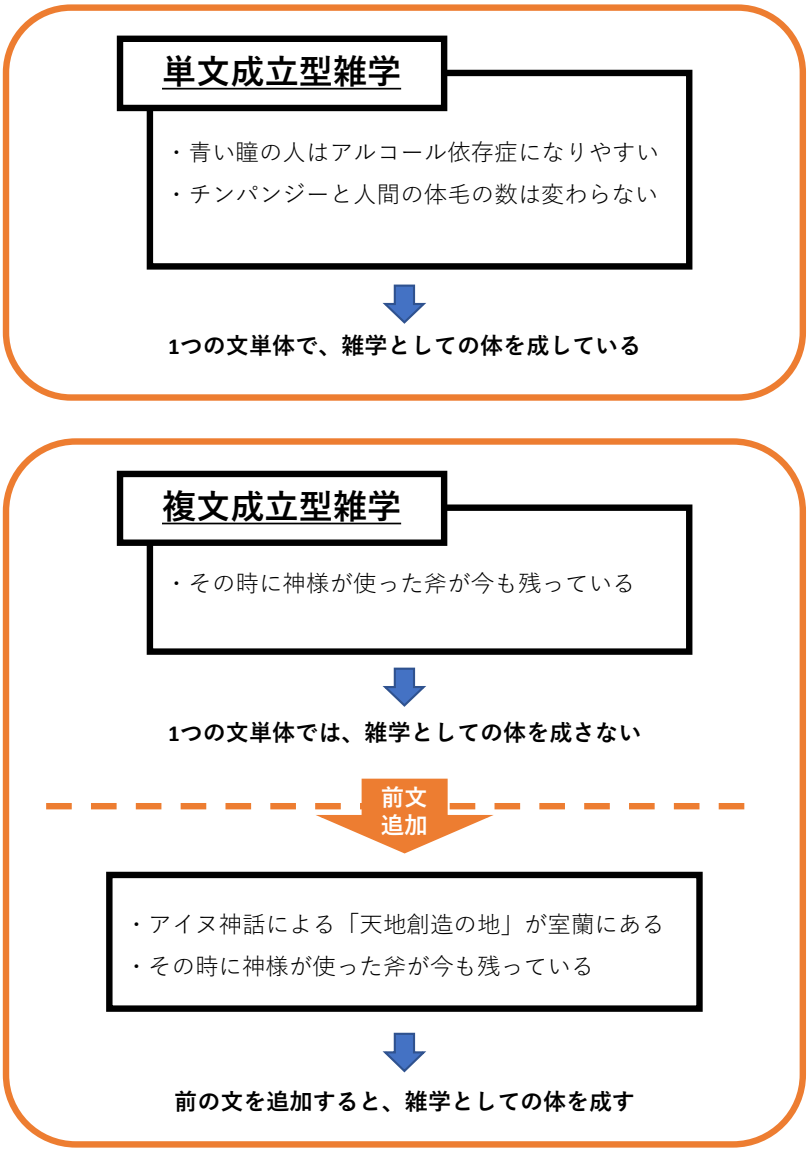

図 3.1　単文成立型雑学と複文成立型雑学

3.3　雑学探索システム全体の概要

本研究における雑学探索システムの流れを図 3.2 に示す．始めに，Twitter からその時点におけるハッシュタグを除いたトレンドキーワード上位 10 件を抽出し，それぞれそのまま検索エンジンに入力する．そのトレンドキーワード 1 つから関連性を持つ幅広い分野に属する雑学の探索を可能とする為に，トレンドキーワードで Web 検索した結果表示された Web ページの要約テキスト各々から人物，場所などといったトレンドキーワードに関連する名詞を雑学検索単語として抽出する．これらを用いて再度 Web 検索を行い，URL からリンクされている各 Web ページのコンテンツより雑学となるテキストを抽出する．このように抽出された雑学の面白さを評価し，その後ランキングしてユーザに雑学を提示する．以上の流れでユーザの関心を煽る雑学提供を目指す．

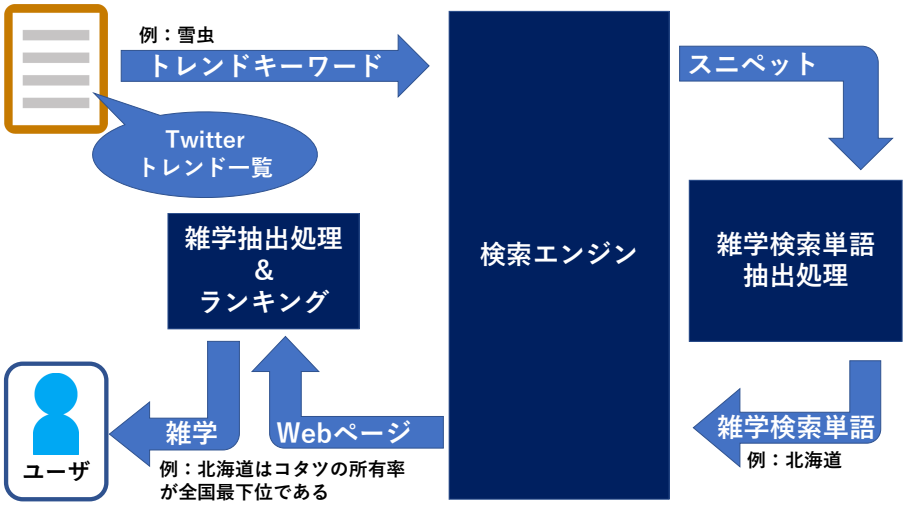

図 3.2　システムの流れ

3.4　雑学探索の始点の決定

Web 上で雑学を見つけようとする時，Web 検索における検索ワードが必要となる．そこで，この検索ワードにトレンドキーワードを使用する．トレンドキーワードは対象の Web リソースにおいて多くのユーザの検索対象となっているワードであり，現時点でのユーザの興味関心を反映しているワードとも言える．また，Web リソースの種類に依って情報の伝播する速度は変化し，特にツイートと呼ばれるメッセージや画像等を投稿できる SNS の 1 つである Twitter においては，リツイートという投稿に対する拡散機能により，トレンドキーワードの伝播する速度が他の Web リソースに比べて突出している．実際に吉田ら [4] の研究でも，Twitter のトレンドキーワードの言及量とトレンドの即時性の高さについて述べられている．その為，Twitter は Web 上で比較的早くトレンドキーワードを掴むことが可能であり，早期のトレンドに関する雑学は将来的に話題性を持つ可能性が高いと考えられる為，Twitter のト

レンドキーワードを雑学探索の始点とすることが最適であると考える．しかし，トレンドキーワード中にはハッシュタグと呼ばれるツイートの共有目的に利用されるものも含まれる場合があり，これは純粋なトレンドキーワードとは言えない為，本研究では除くこととする．

3.5　雑学検索単語抽出

Twitter から雑学探索の始点となるトレンドキーワードを取得後，そのトレンドキーワードを検索ワードとして Web 検索を行う．検索結果として表示された Web ページから URL 各々のスニペットを取得する．スニペットとはそれぞれの URL の Web ページの要約文であり，その Web ページ内で検索ワードに関連した部分がテキストとして抽出されたものである．取得したスニペットから雑学検索単語を取得する為に Google Cloud Natural Language API[5] を用いる．この API によって，入力されたテキストに含まれている既知の名詞を「人物，場所，組織，イベント，メディア，商品」の 6 種類にラベル付けした名詞として抽出することが可能となり，本研究ではこの抽出された名詞を雑学検索単語として使用する．また，ラベル付けによって，より少ない雑学検索単語で幅広いカテゴリの雑学を探索することも可能となる．

ただし，雑学と言ってもユーザが知らない雑学検索単語の雑学は関心を煽ることができないと考えられる．その為，比較的認知度が高い雑学検索単語に絞る為に，各ラベルの雑学検索単語毎に Google での検索数を比較し，検索数がトップの雑学検索単語をそのラベルにおける高認知度の雑学検索単語として抽出を行う．

3.6　雑学抽出の概要

雑学が載っている Web ページを見つける為に，3.5 節で抽出された雑学検索単語と「雑学」という単語を組み合わせて AND 検索を行う．この時，検索結果として表示された上位約 10 件で各 URL からリンクされた先の Web ページ内のテキスト部分全体を抽出する．そして抽出したテキストを「。」「！」「？」「!」「?」を区切り文字として文単位に分割し，雑学検索単語が含まれている文を雑学候補のテキストとして抽出する．この雑学候補となるテキスト群に対し，雑学かどうかの判定，及び雑学としての面白さを評価し，最終的に，これまで抽出された雑学検索単語それぞれの雑学全てを面白さ順のランキングにしてユーザに提示する．雑学かどうかの判定手法の詳細は 4.1 節，雑学の面白さの評価手法の詳細は 4.2 節で後述する．

第 4 章

提案手法

本章では，雑学探索システムによって得られた Web ページ内のテキスト部分全体から雑学だけを抽出する手法，その取得した雑学の面白さをギャップの大きさという観点から評価する手法について述べる．

4.1　テキスト分類による雑学のピックアップ

雑学探索システムによって得られた Web ページ内のテキスト部分全体にはアフィリエイト目的の広告による文や，雑学とはなんら関わりもない文が多分に含まれている．このままではユーザにノイズだらけの雑学を提供することになってしまう為，これらのノイズとなる文を除去する必要がある．そこで，以下の 2 つの手法について提案する．

- 機械学習による分類

雑学の文とノイズの文とを分けることができれば良い，文のある単語から，その前後に出現する単語を推測するモデルを生成するアルゴリズムである Skip-gram を実装し，文書分類に長けた機械学習の手法の 1 つである fastText[6] という手法を利用する．この fastText で学習用テキストに対し，「雑学」，「ノイズ」とカテゴリ別ラベルを付与し，Skip-gram にて教師付き学習を行うことで学習モデルを構築する．このモデルを活用することで新たなテキストに対する雑学文とノイズ文の分類を行う．

- ルールベースによる分類

雑学，ノイズ双方にはそれぞれ特徴的な単語が存在しており，特にノイズ文には特徴的な単語（以下，ノイズキーワードと呼ぶ）が多く含まれている．具体例としては，「いたします」，「スポンサーリンク」，「PR」などが挙げられる．このノイズキーワードを手作業で抽出し，ノイズキーワードが含まれているか否かで雑学文とノイズ文の分類を行う．

4.2　雑学の面白さの評価

面白い雑学を見た場合，「へぇ～そうなのか！」といったような「感心する」という感想が思い浮かぶ傾向にある．この感想が浮かぶ要因として，文章に感じる意外性が面白さを生んでいると考えられる．意外性を感じさせる雑学のパターンとしては「A が実は B だった！」，「A は過去に B をしていた！」というようなものが多く，単語 A と単語 B のギャップが雑学としての面白さの肝であると考えられる．このギャップが生まれる要因として，「雑学中の単語 A と単語 B の類似度は低い」という仮説を立てた．この仮説から，単語の意味をベクトル化しそのコサイン類似度の算出をもって雑学の面白さを評価する手法が考えられる．そこで，本研究ではコサイン類似度を算出するのに，fastText と同じく Skip-gram を実装した，単語の意味をベクトル化するときによく使われる Word2Vec[7] を利用する．この手法により，単語 A と単語 B のギャップの大きさを評価することができ，雑学の面白さの定量的な評価が可能になると考えられる．

しかし，雑学中の単語 A と単語 B が Word2Vec の学習モデルに含まれていない未知語であった場合，単語の意味のベクトル表現が得られず，意図的にその雑学の評価を避けるようにしなければならない．そこで，4.1 節で提案した手法で利用されている fastText は，未知語の意味に対するベクトル化が可能である為，fastText もコサイン類似度の計算に利用する．何故，未知語の意味のベクトル化が実現可能となっているかについては，subword[12] という仕組みを取り入れているからである．この仕組みは，未知語を既知の語に分割した上で，未知語の意味をその既知の語の意味をベクトル化したものの合算によって，未知語の意味のベクトル化を可能としている．例えば図 4.1 で，Word2Vec では未知語である「ゴジラ座」は，fastText では既知の語「ゴジラ」，「座」に分割され，その各既知の語のベクトルの和として「ゴジラ座」のベクトル表現が可能となる．

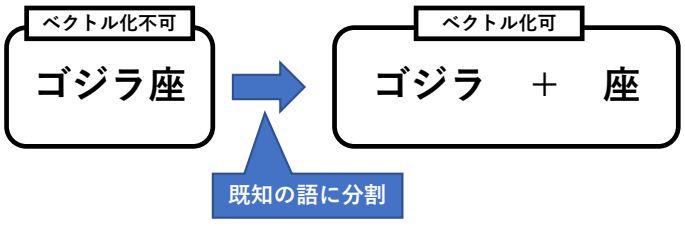

図 4.1　subword の仕組み

ここで事前実験として，雑学から単語 A，単語 B の抽出を行い，それら単語間のコサイン類似度を算出することで雑学の面白さを評価することができるか検証する．文中の単語 A と単語 B の抽出は手作業で行い，単語 A と単語 B のコサイン類似度の算出を Word2Vec と fastText を利用し行う．Word2Vec の学習モデルには白ヤギコーポレーションが提供している日本語版 Wikipedia を学習したもの [8] を利用し，fastText の学習モデルは Facebook Open Source から公開されている日本語版 Wikipedia を学習したもの [9] を利用する．実験に使用

するデータは，以前放送されていた人気バラエティ番組「トリビアの泉」にて紹介された雑学全 1034 件 [10] を使用する．また，この番組において雑学は「へぇ」という 0 から 100 までの値で雑学の面白さが評価される為，各獲得「へぇ」数ごとにデータ数の上限値を 5 個に定め，各々の雑学から抽出した単語 A と単語 B のコサイン類似度を Word2Vec と fastText それぞれを利用して算出し，各雑学の面白さ（へぇ数）とコサイン類似度との関係性を表す散布図を作成すると，図 4.2 と図 4.3 のようになる．

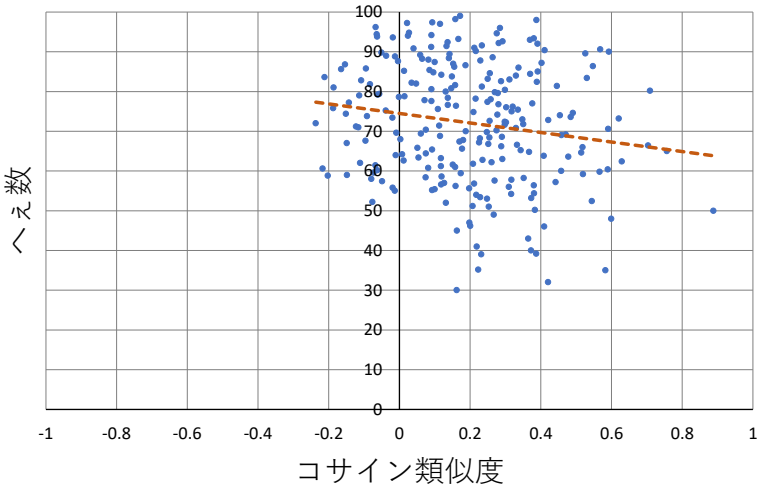

図 4.2　Word2Vec による単語 AB 間のコサイン類似度と各雑学におけるへぇ数の相関関係

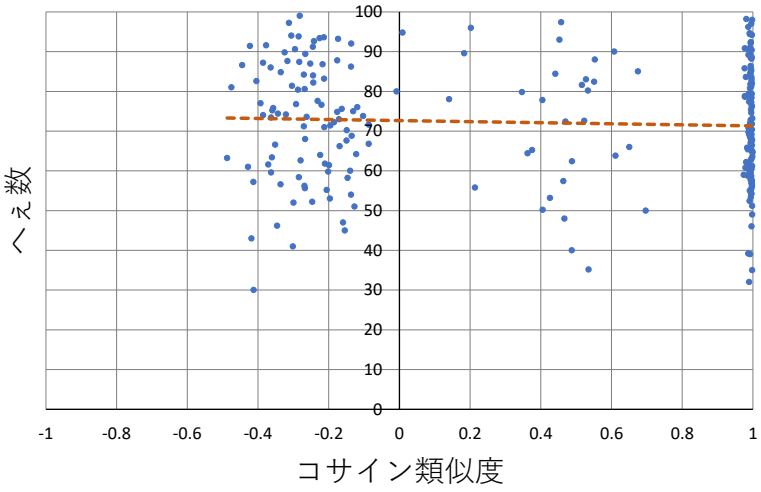

図 4.3　fastText による単語 AB 間のコサイン類似度と各雑学におけるへぇ数の相関関係

図 4.2 と図 4.3 において，コサイン類似度が低いほどシステム的に面白い雑学となり，へぇ数が多いほど番組で高い評価を受けた雑学となる．また，線形近似（回帰直線）を点線で示している．実験結果から，Word2Vec を利用した場合の相関係数は $r = -0.165690091$，fastText を利用した場合の相関係数は $r = -0.052330319$ となり，Word2Vec を利用した場合で，若干ではあるが「へぇ」数とコサイン類似度に負の相関の傾向が見られた．しかし，「へぇ」はあくまで番組基準の面白さ評価値であり，番組で雑学紹介タレントのユーモアのあるトークや，その 1 文に関わる補足トリビアの紹介によっても「へぇ」数は伸びる傾向にあった為，「へぇ」数が単純にその雑学の評価となっていない場合が多く含まれていた．よって，「へぇ」数は雑学単体の面白さを表しているわけではないと言える．

また，この事前実験では単語 A と単語 B を手作業で選出していたが，システム化に向けて Support Vector Machines に基づく日本語係り受け解析器である CaboCha[11] を利用する．この CaboCha を利用することで，ある単語に係る単語が自動的に抽出でき，大規模な雑学データに対する評価を容易に行うことが可能となる．単語 A は雑学において名詞となっている傾向がある為，本研究のシステムでは，同じく名詞である雑学検索単語を単語 A として扱い，雑学文に CaboCha を適用して，この雑学検索単語に係る語を単語 B として抽出する．

第 5 章

評価実験

本章では，雑学検索単語で Web 検索された結果の各 Web ページ中の雑学候補の文からノイズを取り除き，雑学のみを抽出する手法，及び雑学の面白さを定量的に評価する手法に関する評価を行う．

5.1　Web ページ上の雑学の抽出手法の評価

本節では，4.1 節で提案した手法によって Web ページ中から抽出された雑学とノイズがどの程度正しく分類できたかを明らかにしていく．

- fastText を用いた機械学習による分類

ここでは機械学習の fastText による分類精度の検証を行う．使用するデータは，雑学収集システムから取得できた Web ページのテキストより，単体で雑学の定義を満たしている文のみを雑学の正解データとし，それ以外の文はノイズ文の正解データとして手作業で分類する．また，2018 年 11 月 28 日 16 時頃に実施した雑学収集システムにより収集し，テキスト全 4383 件を対象に分類した雑学文 561 件，ノイズ文 3822 件から各 500 件を学習データとする．それとは別に雑学文 61 件，ノイズ文 477 件を判定データとして取り扱う．分類精度に大きく直結するパラメータとして主に学習データ数，単語ベクトルの次元数，学習を繰り返す回数であるエポック数の 3 要素がある為，それぞれ数値を変更した場合ごとの分類精度を比較する．

図 5.1 に雑学文，ノイズ文各 50 件を合わせた学習データ数 100 件，各 100 件を合わせた 200 件，300 件,...，1000 件ごとにそれぞれ学習を行ったモデルを利用した際の分類精度について示す．図 5.1 で F 値に着目すると，見ても分かる通り学習データ数の増加に依る分類精度の向上は見られなかった．唯一，学習データ数 700 から 800 にかけて一時的に分類精度が向上しているが，これは学習データの内容に依るところが大きく，この時の学習データ数に追加された 50 件のノイズ文が，広告や宣伝などの内容に傾倒しており，出現する単語にノイズキーワードが多く含まれていたからであると考えられる．

次に図 5.2 に，前述の図 5.1 で F 値が最も高かった学習データ数 100 件で次元数 50，100，150,...，300 ごとにそれぞれ学習を行ったモデルを利用した際の分類精度について示す．図

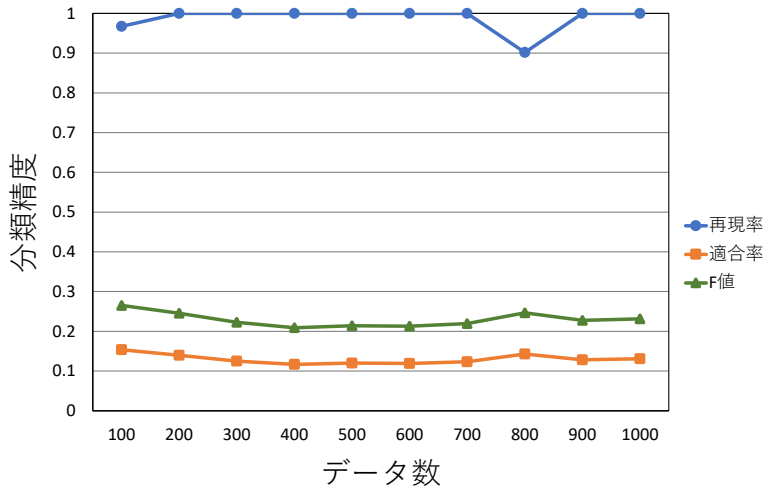

図 5.1　学習データ数に依る機械分類精度の変化

5.2 から見ても分かる通り次元数が低くなると再現率は急激に悪化している．一方，適合率に関しては，安定的な推移となっている．F 値で見ると，次元数 100，150 辺りが最も高い値を取っており，これは，本研究で取り扱っているデータセットの大きさが小さい為，低い次元が適当であるからであると考えられる．

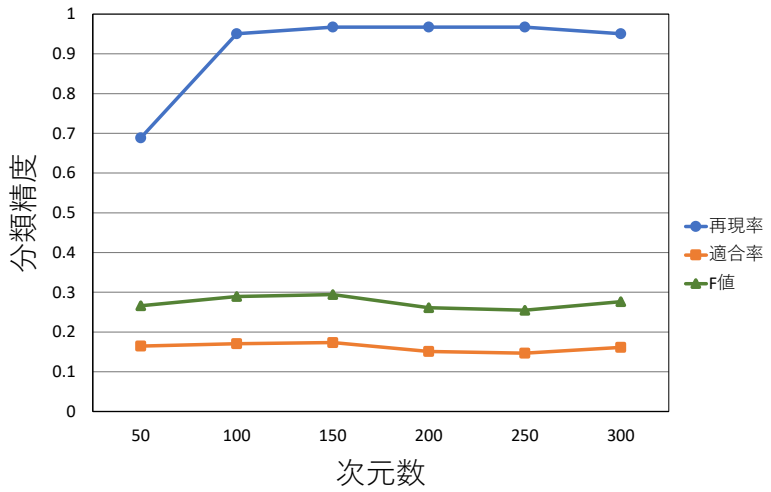

図 5.2　単語ベクトルの次元数に依る機械分類精度の変化

最後に図 5.3 に，前述の図 5.1 で F 値が最も高かった学習データ数 100 件でエポック数 5，10，15，. . . ，30 ごとにそれぞれ学習を行ったモデルを利用した際の分類精度について示す．

図 5.3 から見ても分かる通りエポック数の増加に依って，再現率は好転しているが，適合率，F 値はともに悪化している．

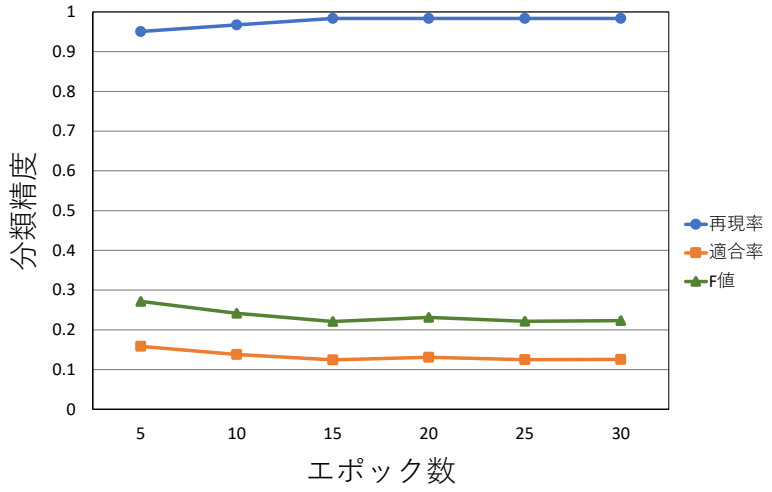

図 5.3　エポック数に依る機械分類精度の変化

上記よりまとめると，次元数は対象となる学習データ数に応じて適宜調整が必要となるが一般的に増やせば増やすほど分類精度が向上するとされている学習データ数とエポック数の増加に依る分類精度の向上は見られなかった．このような結果となってしまった大きな原因として，複文成立型雑学がノイズ文の学習データとして含まれてしまっている為，学習データのノイズとしての専門性が低くなってしまい，分類精度の低下に繋がっているかと考えられる．また，雑学文とノイズ文の双方に言えることであるが，そもそもこれらの取り扱っている文書は多岐のジャンルにわたる系統立っていない様々な事柄について述べられているものであり，出現する単語の傾向が表れにくいカテゴリである為，Word2Vec と同じく Skip-gram を実装している fastText ではこれらのカテゴリ特徴を捉えにくいと考えられる．

- ルールベースによる分類

ここではノイズキーワードを用いたルールベースによる分類精度の検証を行う．本研究で，ノイズ文に分類される文から手動で選出したノイズキーワード全 161 個を用いて分類を行う．選出したノイズキーワードの一例を図 5.4 に示す．

あの，あるある，あるよ，いかが，いきます，いたします，
いただき，いるよう，うれしい，おすすめ，コツ，マジ，
ヤバ，サイト，シェア，ハンパ，ベスト，ページ，リンク，
オススメ，一方，一覧，出典，方法，以来，今回，比較，
当時，追記，記事，ID，PR，TOP，NAVER，by，jp，com，etc，
jpg，net，↑，→，↓，←，⇒，・・，..，※，>，≫

図 5.4　ノイズキーワードの一例

選出した単語群が 1 つでも含まれている文をノイズ文と分類し，それ以外の文を雑学文として分類することで，正しい分類を行うことができるのかを検証する．精度評価には，fastText による分類時に利用した判定データをこの手法でも利用する．

表 5.1 に前述の fastText における最も精度が高い，学習データ数 100 件，次元数 150 の場合の精度と，ルールベースによる分類精度を示す．表 5.1 の F 値に注目すると分かるように単純な精度を比較するとルールベースによる分類の方が優れているという結果となった．また，分類が上手く行われなかった文について分析すると，ノイズキーワードに含まれる「あの」などの指示語による分類ミスが多く発生していた．複文成立型雑学をノイズとして分類する上で指示語をノイズキーワードに選出しており，雑学とノイズの手動による分類は指示語が文中に含まれていても，その文単体で意味が伝わるものであれば単文成立型雑学として分類を行っている．具体例としては「その初テレビ放送の第一声は“JOAK-TV，こちらは NHK 東京テレビジョンであります”でした.」などがある．よって，ノイズキーワードに含まれている指示語の存在が，そのようなケースの雑学をノイズとして分類するということが再現率の低下の主な要因になっていると判明した．

表 5.1　機械学習とルールベースの分類精度

手法	再現率	適合率	F 値
fastText	0.967213	0.173529	0.294264
ルールベース	0.524590	0.359551	0.426667

以上，2 つの手法についてまとめると，本研究では単文成立型雑学に限定して分類を行った為，複文成立型雑学がノイズ文として扱われていることで，このことが両手法いずれにも悪影響を与えていると考察できる．また，Web ページから雑学候補テキストを取得する際，文の分割処理が「。」「！」「？」「!」「?」の 5 つの区切り文字によって行われているが，必ずしも文末がこれらの区切り文字になっている訳ではなく，上手く分割できていない文も見受けられ

た．よって，その分割不備が同じように両手法に悪影響を与えていると考えられる．これらの問題の有効的な解決策として，形態素解析などにより前文との関係性を分析することで，複文成立型雑学を取り出す，または正しく文を分割することなどが挙げられる．更に精度が上昇する手法として，ルールベースによる分類は，ルールを増やし厳しくすればするほど適合率が上がる為，適合率 9 割程度までルールを増やし，テキスト群からほぼ確実にノイズ文を分類する．その後，ルールベースによる分類によって失われた再現率を機械学習による分類で補うという，機械学習による分類手法とルールベースによる分類手法を組み合わせるハイブリッドな手法が考えられる．

5.2　雑学の面白さ算出手法の評価実験

本節では，4.2 節で提案された手法による雑学の面白さの評価が，どの程度人による評価に近いのかを検証する実験を行った．使用するデータは，5.1 節で利用していた雑学文 561 件とする．この判定データ 1 件ごとに，雑学検索単語に CaboCha を適用し，雑学検索単語に係る単語 B を抽出する．そして，その 2 単語間のコサイン類似度を算出後，コサイン類似度が低い順にランキングすることで雑学の面白さを評価する．コサイン類似度の算出には Word2Vec と fastText を適用する．Word2Vec と fastText の学習モデルは，それぞれ 4.2 節で適用していたものを利用する．また，両手法によってランキングされた雑学を順位に基づいて等間隔にそれぞれ 30 件雑学をピックアップし，被験者 40 人に面白いと思った雑学を任意の数，選択してもらうアンケートを実施した．ランキングの一例として，2018 年 11 月 28 日 16 時頃に取得し，ハッシュタグの除いたトレンドキーワード上位 10 件（「メタル化」「アー写感」「Mazda3」など）を始点に探索した雑学ランキング結果を図 5.5 に示す．

そして，Word2Vec と fastText の両手法により単語間のコサイン類似度を算出し，各雑学の被験者に面白いと選択された人数とコサイン類似度との関係性を表す散布図を作成すると，図 5.6 と図 5.7 のようになる．

図 5.6 と図 5.7 においてコサイン類似度が低いほどシステム的に面白い雑学となり，選択人数が多いほど人から見て面白い雑学となる．また，線形近似（回帰直線）を点線で示している．実験結果から，Word2Vec では相関係数 $r = -0.00371$ となり，fastText では相関係数 $r = -0.35252$ となった．理想では強い負の相関が得ることができれば良かったが，結果として両手法ともにはっきりとした相関は得られなかった．両手法ともに，はっきりとした負の相関が得られなかった原因としては，雑学検索単語に対して係り受け関係にある単語 B はコサイン類似度を算出する単語 B として適当ではないケースが多いことが大きな原因であると考えられる．具体例として，「秋刀魚は立って泳ぐ」という雑学では「秋刀魚」という雑学検索単語に対し「立つ」という単語 B の取得が理想であるが、CaboCha では「秋刀魚」に対し「泳ぐ」という単語 B が取得されてしまう．このように CaboCha ではほとんどの場合，係り受け解析で主語の名詞に対し，「ある」「する」などといった述語の動詞が単語 B として取得される．ギャップのある単語のペアは，主語と動詞の組み合わせ関係ばかりではない為，単語 B が動詞ではなく修飾語などであった場合，ギャップのある単語のペアがほぼ取得不可能と

【Word2Vecでのランキング】
1位.　「六甲おろし」と言えば間違いなく通じるので、
曲名だと思っている方は大半ですが、正しくは「阪神タイガースの歌」です。
2位.　1980年代、ノルウェーの水産業者が日本を訪れたことがきっかけで、
日本にサーモンが輸入されることになった。
.
.
.
29位.　42.チンパンジーと人間の体毛の数は変わらない。
30位.　えきねっとを利用した予約方法であれば、
従来の手順に比較して手間が減らせて手軽にチケットの購入が済ませられるのですが、
優れているのは購入に際する手順に限らず購入時に支払う料金も例外ではありません。

【fastTextでのランキング】
1位.　36.日本には「謎のフルーツ味」、「天才エネルギー」など、
奇妙なフレーバーのファンタが70種類以上ある。
2位.　25.日本にはウサギだらけの島があるウサギ島よばれるうさぎたちの楽園は、
広島県の大久野島、瀬戸内海にある周囲4.3kmの小さな無人島にある。
.
.
.
29位.　しかし、チャーチルがヒトラーに対してＶサインをした写真が世界に出回った際に、
現在の「平和」という意味でピースサインが広まりました。
30位.　■補足同様に、手に汗をかくという行為は、木の枝をつかみやすくして
すばやく逃げるためという、人が木の上で生活をしていた時のなごりだそうです。

図 5.5　Word2Vec と fastText によるランキングの一例

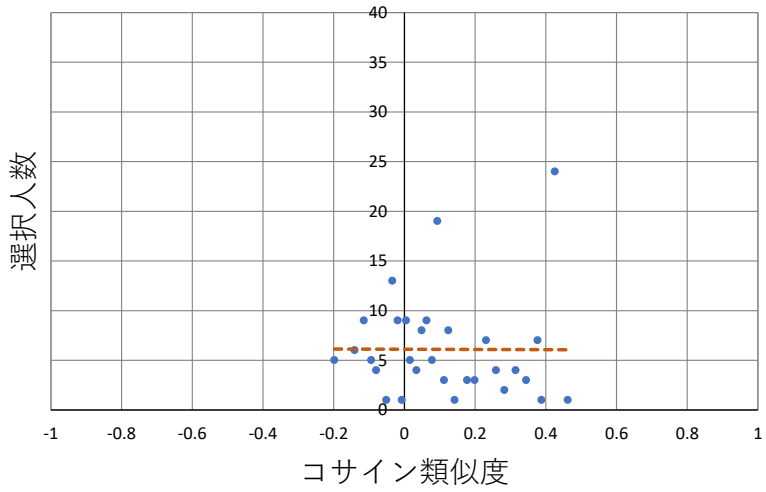

図 5.6　Word2Vec によるコサイン類似度に基づく雑学ランキングと主観的面白さ評価との相関関係

なってしまう．そこで，雑学文には係り受け関係が重なり合う階層構造が存在している場合がある為，その場合において階層構造中から抽出すべき単語 B を探し出す手法を検討する必要がある．以上のまとめとして，全てのケースにおいて最も適当な単語を抽出することは不可能であるが，1 つの係り受け関係のみでギャップのある単語のペアを決定せず，係り受け関係の階層構造を考慮した適当な単語ペア抽出の実現が面白さ評価の改善に繋がるのではと考えられ

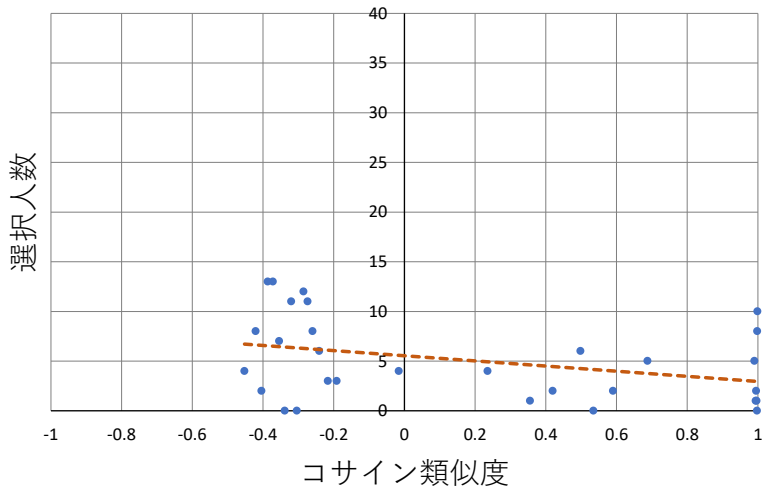

図 5.7　fastText によるコサイン類似度に基づく雑学ランキングと主観的面白さ評価との相関関係

る．ただし，これらの手法では雑学検索単語との関連性をほとんど失ってしまう為，単語間のギャップとは別の評価基準を見つけるべく，面白さが高く評価されている雑学の更なる分析も必要である．

また，Word2Vec に比べて fastText は若干負の相関が得られており，比較的 fastText による雑学の評価の方が優れていると言える．このような結果となった要因として，表 5.2 からも分かる通り fastText の評価可能雑学数が Word2Vec に比べ 2 倍近い差があることから，ランキング対象数が大幅に減少し評価精度が下がったのではとも考えられる．ちなみに，判定データの 561 件に対し fastText の評価可能雑学数が 510 件に減少している原因は，CaboCha によって雑学検索単語に係る単語 B が抽出できなかった文に因るものであると考えられる．実験設計においても，「面白いかどうか」という選択基準であると被験者の興味が結果に影響を及ぼす可能性がある為，さらに被験者を増やす，選択基準を「へぇ～と思うか」というようにアンケートの仕様を変更することなどが必要である．

表 5.2　雑学の面白さ算出手法における Word2Vec と fastText のコサイン類似度算出不可回数と評価可能雑学数の比較

手法	コサイン類似度算出不可回数	評価可能雑学数
Word2Vec	548	229
fastText	0	510

第 6 章

まとめと今後の課題

本研究では，単文成立型雑学を中心に Web 上から幅広いカテゴリの雑学を探索し，収集した雑学の面白さを評価してランキングする雑学探索システムを提案した．その中で，雑学文とノイズ文を分類する幾つかの手法，及び雑学の面白さを単語間のギャップという観点から定量的に評価する手法を考案し，評価実験を行った．評価実験の結果から，雑学分類手法全ての悪影響に関わる今後最優先に解決すべき問題として，複文成立型雑学の認識，及び取得が明らかとなった．この問題に対しては，多くの複文成立型雑学に含まれる指示語によって失われている欠損情報を，文の前後関係を解析し，前の文から補填する方針で今後対策を行うことを考えている．最終的に本研究によって，Web 上から数多くの雑学抽出が容易になり，単語間のギャップというテキストの面白さの新たな観点を確立したと言える．現状では，ノイズ文のテキスト特徴を捉えて，テキスト群からノイズ文を除去していく手法を採用しているが，今後，雑学文のテキスト特徴を分析し，新たに雑学をダイレクトに抽出する手法を確立することで，雑学に関わる Web ページのみからではなく，Web 上全てのテキストから潜在的雑学をマイニングすることが可能になると考えられる．また，評価実験を行う過程で他にも解決すべき課題が複数存在することが判明した．以下にその解決すべき問題点について示す．

6.1　抽出された雑学の重複傾向の解消

現状では，各トレンドキーワードの各ラベル付けされた名詞の中から最も Google での検索数が多いものを雑学検索単語として抽出しているが，取得される名詞全てにそのプロセスを用いている故に，雑学検索単語として抽出されるものが日常的に使われている一般名詞に偏ってしまう傾向が見られた．具体例としては「人」,「日本」,「川」などが挙げられる．このままではある特定の単語，特に一般名詞に関わる雑学ばかり集められてしまい，幅広いカテゴリの雑学を取得することができない．そこで，あまりにも日常的に使用されている一般名詞は除外することや取得できる一般名詞の割合を本研究で収集した雑学中での出現確率を参考に設定するなどの対策が必要となる．

6.2　パーソナライズ情報の入力に依る関心誘発力の向上

本研究で提案したシステムは Web 上から網羅性を重視した雑学を探索するシステムと言える．しかし，網羅性を重視するということは，広く浅い雑学を探索することと同義であり，システムを複数回利用した際に出力される雑学は浅い部分である一般的な事象に沿った雑学に偏ってしまう．また，「関心を煽る」という上で少なくともユーザの好みを分析することは大切であると言える．そこで，例えば「パーソナライズ情報としてユーザに関心のある分野に即したキーワードを入力として受け付け，そのキーワードのカテゴリを重み付けした探索を行う」ことで，よりユーザの好みに沿った専門性の高い雑学が探索できるのではと考えている．

6.3　客観的な雑学の定義の必要性

評価実験で学習データとして用意した Web 上からの手作業による抽出で得られた雑学であるが，その手作業で抽出する際，雑学かどうかを判断する為に「知識」「面白い」というテキストとして 2 つの要素を満たしているかどうかによって抽出を行った．ただし，雑学であるかそうでないかの判断基準は人によって大きく異なることが，5.2 節によるアンケート調査によって明らかとなった．本研究では，より多くの人に面白い雑学を提供することを目標としている為，一般的に雑学とはどのような要素を持って成り立つのかを具体的に明らかにしなければならなかった．雑学という曖昧なカテゴリの情報を取り扱う上で，主観的な定義は悪影響を及ぼす為，改めて雑学の成立条件に関わるアンケート調査を行う必要がある．

6.4　雑学の信憑性の評価

Web 上には，誰でも手軽に情報を発信できるが故に，偽情報が存在している．この偽情報の拡散によって，災害時では実際に被害が出ているケースも過去には存在している為，Web 上から情報を収集する上で，偽情報は可能な限り排除しなければならない．このシステムでは，重要な情報が発信されるわけではない為，ユーザに被害が出ることは考えにくいが，間違った知識を披露するようなことは出来るだけ避けるべきである為，偽情報かどうかを判定する処理の実現も考慮すべき問題である．

謝辞

本研究に際して，様々なご指導を頂きました服部峻助教に厚く御礼申し上げます．また，日常の議論を通じて多くの知識や示唆を頂いた服部研究室の皆様にも深く感謝の意を表します．

参考文献

[1] Video Research Ltd.（2003）2003 年 年間高世帯視聴率番組 30（関東地区）, <https://www.videor.co.jp/tvrating/past_tvrating/top30/200330.html>.

[2] 佐々木 稔, 新納 浩幸, “文書分類を用いたスパムメール判定手法,” 情報処理学会研究報告 自然言語処理 93, pp.75–82 (2004).

[3] 天谷 祐介, 荒木 健治, ジェプカ ラファウ, “テキストの面白さの評価によるユーモアの認識,” ファジィシステムシンポジウム講演論文集 28, pp.233–238 (2012).

[4] 吉田 光男, 荒瀬 由紀, “トレンドキーワードに関するウェブリソースの横断的分析,” 情報処理学会論文誌, データベース, Vol.9, No.1, pp.20–30 (2016).

[5] Google（2018）Google Cloud Natural Language API, <https://cloud.google.com/natural-language/?hl=ja>.

[6] Facebook Inc.（2018）fastText, <https://fasttext.cc/>.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, “Distributed Representations of Words and Phrases and their Compositionality,” Advances in Neural Information Processing Systems 26, pp.3111–3119 (2013).

[8] Tanida Kazuaki（2017）word2vec の学習済み日本語モデルを公開します, <http://aial.shiroyagi.co.jp/2017/02/japanese-word2vec-model-builder/>.

[9] Facebook Inc.（2018）Wiki word vectors・fastText, <https://fasttext.cc/docs/en/pretrained-vectors.html>.

[10] Noncky.（2011）トリビアの泉 パーフェクトデータベース, <http://www.noncky.net/trivia/>.

[11] 工藤 拓, 松本 裕治, “チャンキングの段階適用による日本語係り受け解析,” 情報処理学会論文誌, Vol.43, No.6, pp.1834–1842 (2002).

[12] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov, “Enriching Word Vectors with Subword Information,” Transactions of the Association for Computational Linguistics, Vol.5, 2307-387X, pp.135–146 (2017).