新聞記事の特徴語を用いたクロスワードパズルの生成

日置 竜輔[1,a)]　角 康之[1,b)]

概要：本研究では，過去の新聞記事から抽出した年代ごとの特徴語をデータベースとして，クロスワードパズルを自動生成し，日本語を学習する外国人や子供を支援する言語学習システムとしての確立を目標とする．クロスワードパズルの盤面の自動生成は黒マスが縦や横に連続してはならない制約や，盤面が黒マスによって分断されてはならないといった黒マスルールが存在する．これらの制約に対しては，素集合データ構造の Union Find Tree や幅優先探索といったアルゴリズムを駆使し，自動生成には貪欲法のアルゴリズムを使用することで制作を試みる．クロスワードパズル内で問題となる単語には，時代・難易度・カテゴリなどといった属性を兼ね備えた特徴のあるクロスワードパズルの作成を検討する．さらに，興味を持った単語に関連する新聞記事へ誘うことで，新聞に新たな付加価値を与えることを提案する．

1. はじめに

本研究では，過去の新聞記事から抽出した年代ごとの特徴語をデータベースとして，クロスワードパズルを自動生成し，使用した単語に関する新聞記事へ誘導することで，各年代の社会的背景・歴史を学ぶ機会を提案する．雑誌や新聞に掲載されているクロスワードパズルは，一般的に使用されている語彙から抽出した単語が多い傾向にある．これを新聞記事から抽出した特徴語にすることで，時代・難易度・カテゴリなどといった属性を兼ね備えた特徴のあるクロスワードパズルが出来上がることが期待される．さらに，日本語を学習する外国人や子供を支援する言語学習システムとしての確立も提案する．

2. 関連研究

日置らの研究では，新聞記事との新い出会いの場を作り出すために言葉遊びのスマートフォンゲームを制作した [1]．本稿では，その際制作しなかったクロスワードパズルに焦点を当て，自動生成することに取り組んだ．

Patrick らの研究 [2] により，薬理学を勉強する学生にクロスワードパズルを用いて学習させることで成績が向上する結果になった．薬理学は，専門用語の単語や意味を暗記するのが難しいため，様々な学習戦略と方法が検討されている．第 2 プロフェッショナル MBBS コースの第 5 学期の学生，合計 139 人の学生を対象に内分泌薬理学に関する 32 問からなるクロスワードパズルを作成し取り組ませた．その結果，学生全体で平均 52.69% の得点を獲得し，全ての学生がホルモン避妊薬に関する質問に正しく答えた．また，その大多数は，薬の知識を高め，病気や薬の名前を覚え，学習に役立つと回答した．このことから学習補助ツールとしてのクロスワードパズルは，大多数の学生が学習態度を改善し，それにより成績が向上した結果となった．

さらに Pearson の研究 [3] では，COVID-19 の直前と渦中に，薬学部の学生，1 年生 132 名と 2 年生 120 名の計 252 名に復習補助として，オンライン講義後の化学クロスワードパズルを実施した．学生のオンライン学習行動パターンに基づいて評価を行い，2 年生の約 80% がクロスワードパズルが役に立つと回答し，化学をテーマにしたクロスワードパズルは，COVID-19 の最中に実施すると好評であるとの結論が出た．その結果，多くの教育機関でリモート授業が続く中，準備が簡単なクロスワードパズルの学習教材は，講義の一部に検討されるべきであると結論付けた．

このような研究結果から，本研究の目的である日本語を学習する外国人や子供を支援する言語学習システムとして試すには十分な価値があると考えられるため，本稿ではクロスワードパズルに焦点を当て制作することにした．

3. 使用データについて

3.1 新聞記事データ

今回使用するデータは北海道新聞の過去 33 年間に及ぶ実際に新聞に掲載された実データである．提供データは全

[1] 公立はこだて未来大学
[a)] r-hioki@sumilab.org
[b)] sumi@acm.org

て csv ファイルの形式となっており，膨大な量であった．Rigutini らの研究 [4] では，Web 上からダウンロードしたページから語彙とその定義を抽出し，自然言語処理技術を応用することで，テーマに特化したクロスワードパズルの生成に成功した．本研究では Web ページからのスクレイピングによってデータを収集するのではなく，新聞記事からデータを集めるといった点に違いが存在することから独自性の高い研究であると考えられる．

3.2 データの加工・整形

北海道新聞社から提供していただいた新聞記事データのままでは，実際の新聞には掲載されていないノイズや欠損値が存在していたため，データの前処理を行う必要があった．データの前処理には，正規表現を使いノイズ除去を行った．その後の整形に関しては，主に Python の pandas ライブラリを利用し，整形を行った．

4. 特徴語の抽出

本章では，膨大な量の新聞記事データから特徴語を抽出する手法について述べる．

4.1 形態素解析

クロスワードパズルの単語で使用する品詞は全てが名詞であるため，新聞記事を品詞分解し，名詞のみを抽出する必要がある．今回は形態素解析を使用して品詞分解を行った．形態素解析とは，自然言語処理の 1 つで，自然言語で書かれた文章を最小単位である形態素に分け，それぞれの品詞や活用形などを判別する解析手法の 1 つである．今回は，MeCab と呼ばれる形態素解析ソフトウェアを使用して形態素解析を行った．MeCab とは，オープンソースで誰でも利用できる形態素解析器であり，Python で使用できる上に，処理速度が高速であるという点から MeCab を使用した．

4.2 mecab-unidic-NEologd

オープンソースで利用しやすい MeCab と同梱されている IPA 辞書の精度に限界がある．特に，固有名詞の抽出には失敗してしまうケースが多数存在し，例えば「公立はこだて未来大学」という単語に対して形態素解析を行うと，[”公立”, ”はこだて”, ”未来”, ”大学”] のように期待と異なる出力結果になってしまう．その点を補うのが mecab-unidic-NEologd [*1] である． mecab-unidic-NEologd は，UniDic [*2] に多数の Web 上の言語資源から得た新語や固有表現，絵文字などのエントリを足して MeCab のシステム辞書としてインストールするためのシードデータとスクリプト群のセットである．特徴として，UniDic には含まれていない固有表現などの語とフリガナの組を約 338 万組採録している点や，毎週 2 回辞書のアップデートが行われている点である．これを使用することで，最近の特徴語や固有名詞に対しても形態素解析を行って抽出することが可能となった．

[*1] https://github.com/neologd/mecab-unidic-neologd
[*2] https://clrd.ninjal.ac.jp/unidic

4.3 ストップワードによる頻出単語の除去

クロスワードパズルの問題として使用する単語は基本的には名詞であるが，「これ」，「あれ」，「わたし」などのいつの時代であっても一定数出現する単語は特徴語ではない．このような単語は事前にストップワードとして除去を行った．

4.4 TF-IDF

特徴語を抽出するための手法として， TF-IDF (Term Frequency-Inverse Document Frequency) を使用する．TF-IDF とは，ある文書内 d_j での単語 t_i の出現頻度を表す TF (TermFrequency) に，全ての文書内での単語が含まれる文書の割合の逆数である IDF (Inverse Document Frequency) を掛け合わせたものである．具体的には以下の数式で表される．IDF を求める際に，分母に 1 を足しているのは，一度も文書内に特定の単語が出現しなかった場合，分母が 0 になるのを防ぐためである．

$$tf(t_i, d_j) = \frac{f(t_i, d_j)}{\sum_{t_k \in d_j} f(t_k, d_j)}$$

$$idf(t_i) = \log\left(\frac{N}{df(t_i) + 1}\right)$$

$$tfidf(t_i, d_j) = tf(t_i, d_j) \cdot idf(t_i)$$

これらを用いて新聞記事から特徴語を抽出する．抽出した 2020 年の特徴語上位 20 件の結果を表 1 に示す．

TF-IDF 値のみに基づいた特徴語の抽出では，「新型コロナウイルス」や「感染拡大」などといった，その年代を風刺する単語が抽出出来ている一方で，「札幌」や「北海道」など，通年に渡って出現する単語も抽出する課題が浮上した．

4.5 閾値の設定

そこで，1990 年から 2020 年までの TF-IDF 値のみに基づいた特徴語の抽出結果から，各単語が計 31 年間で何回出現したか計測するようにした．そして，その出現回数

表 1 TF-IDF に基づいた 2020 年の特徴語

単語	TF-IDF 値
新型コロナウイルス	0.004611158399909591
感染	0.0034447680486929537
札幌	0.003263287289859507
道内	0.0027232448427818522
作る	0.0026049861865424864
感染拡大	0.002578076198481379
べし	0.0025448729589117765
地域	0.002456991458283796
対策	0.002454951478869493
時間	0.002384288224627461
北海道	0.0023704972115276366
必要	0.0023254549414860413
ながら	0.0022415355070587034
午前	0.0022294026598467347
事業	0.002214221185367941
参加	0.002173512994908174
販売	0.0021522609233730135
施設	0.0021085989412959203
デン	0.0020608174204038788
函館	0.0020307353410474837

が閾値を超えた場合には，抽出した特徴語から除去する処理を取り入れた．このような処理を行った結果を図 1 に示す．

新型コロナウイルス　感染拡大
新型コロナ　オンライン
コロナ禍　休校
緊急事態宣言　感染症
外出　消毒
新規感染者　感染予防
感染対策　ストレス
拡大防止　クラスター
営業時間　寄付金
飲食業　外出自粛
新型コロナウイルス感染症　感染者数
重症　バイデン
テレワーク　医療従事者
アルコール　武漢
Zoom　官房副長官
You Tube　入院患者
自粛要請　宿泊客

図 1 出現回数を考慮した 2020 年の特徴語

今回は閾値を 1 に設定し，2 年以上で特徴語として抽出された単語は除去を行った．その結果，「札幌」や「北海道」など，通年に渡り抽出された単語は除去される結果となった．さらに，抽出された特徴語は，2020 年にしか抽出されなかった単語が全てを占める結果となった．最終的に表に表示していない単語も含めると，2020 年の特徴語は計 138 個抽出し，他の年代と比較しても最も多く抽出が出来た結果となった．これは新型コロナウイルスが蔓延したことで，生活が大きく変わり，特徴語も大きく変化したためだと考える．

5. ゲーム制作

ゲームの作成には，OpenSiv3D [*3] を使用した．OpenSiv3D は，ゲームやアプリを C++ コードで開発できるフレームワークであり，豊富なクラスや関数を組み合わせて，2D/3D ゲーム，メディアアート，ビジュアライザ，シミュレータなどのアプリを，効率的に開発することが可能である．

5.1 ゲーム画面

ゲーム画面の遷移を図 2 に示す．

ゲームを開始するとタイトル画面に遷移し，ルール説明か年代選択への画面へと遷移する．また，クロスワードパズルのプレイ時間が長くなることを考慮し，BGM を取り付けた（図 2 ①）．

タイトル画面の後，最初に年代選択を行う画面に遷移する．年代選択をユーザにしてもらうことで，興味を持った年代の特徴語が使用されたクロスワードパズルを遊んでもらうためである．選択した年代の特徴語を少なくとも 2 単語以上は含めたクロスワードパズルを生成した．年代は提供データと同様の 1990 年 から 2020 年までの 31 年分の年代でクロスワードパズルを遊ぶことができる（図 2 ②）．

年代を選択した後，クロスワードパズルのゲーム画面に遷移する．クロスワードパズルのゲーム画面では画面左にタイマーとクロスワードパズルの盤面を配置し，右上にキーボードによる入力ボックス，その下にヒントを配置した（図 2 ③）．

6. クロスワードパズルの自動生成

本章ではクロスワードパズルを自動生成するアルゴリズムについて紹介する．クロスワードパズルの盤面を作成する上で，黒マスルールという黒マスが上下，左右で連続してはならないといった制約や黒マスによって盤面が分断されてはならないといった制約が存在する．このような制約は，素集合データ構造を扱うアルゴリズムである Union Find Tree や幅優先探索（Breadth First Search）といった探索アルゴリズムで制約を満たしているかは確認することが可能である．これらの制約は絶対に厳守しなければならないというわけではないが，制約を満たした上で生成された盤面の方が質の高いクロスワードパズルであると考えられるため，出来る限り制約を満たすようにアルゴリズムを

[*3] https://github.com/Siv3D/OpenSiv3D

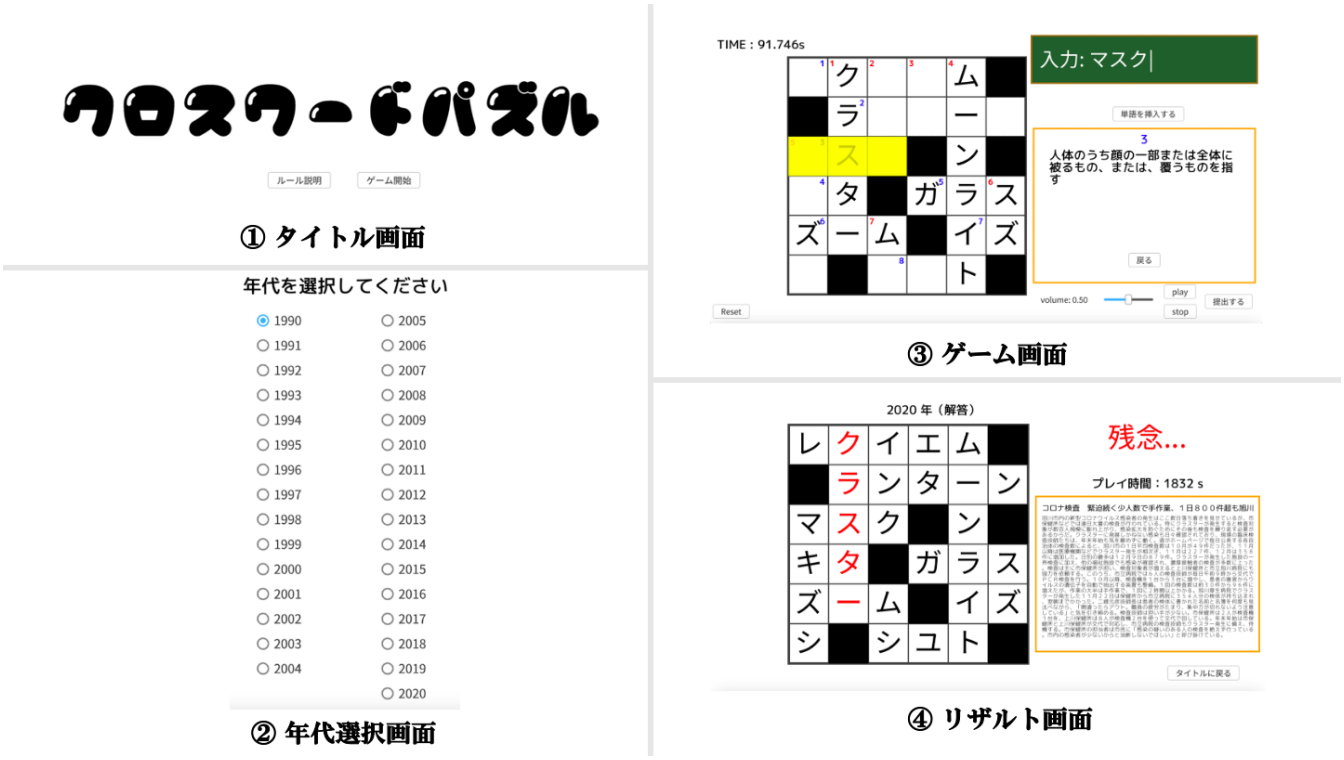

図 2　ゲーム画面の遷移図

構築していく．

6.1 貪欲法による盤面の自動生成

貪欲法とは，その場その場で最善の選択を取っていくアルゴリズムである．貪欲法によるアルゴリズムを図 3 に示す．

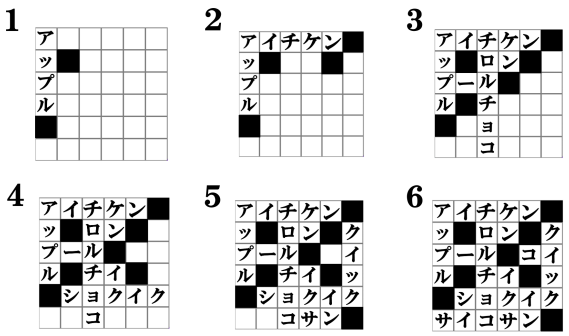

図 3 貪欲法で生成した盤面

まずはじめに盤面の左上に縦に単語を挿入する．ここで，「ン」や「ー」などその文字を先頭として作れる単語が存在しない場合，横に続けて単語を挿入することは不可能であるため，その横のマスは黒マスとする．次に，横に繋げられる単語を挿入する．同様に「ン」や「ー」などといった文字があった前後は黒マスとする．これを全てのマスが埋まるまで続けてやっていくことで貪欲法による盤面の自動生成が完成する．しかし，このアルゴリズムには欠点もあり，盤面に言葉が埋まってくるほど制約が厳しくなるため，そもそも解が存在しない場合もあり得る．また，解が存在しても黒マスが上下左右に連続してしまう可能性も考えられる．このように貪欲法による盤面の自動生成は，構築が簡単である反面，解が存在しないといったケースも存在するので最適なアルゴリズムであるとは言えない．そこで次に紹介するのが山登り法というヒューリスティックの探索アルゴリズムである．

6.2 鍵の自動生成

クロスワードパズルを解くためのヒントとなる鍵の生成も本研究の目的の 1 つである．鍵の生成には Wikipedia の情報を利用することを考える．Python にある wikipedia ライブラリを使用することで，Wikipedia の情報を取得することが可能である．Wikipedia の記事の中でも鍵の生成には summary 文章を使用する．Wikipedia の summary 文章は，検索ワードで始まることが多い傾向にある．そこで，冒頭の検索ワードの部分を除去することにより，鍵の生成を自動で行うことが可能となる．鍵となる文章の前の数字は縦・横それぞれの埋めるマスの番号を表している．これをクロスワードの鍵として使用すると図 4 のようになる．

図 4 クロスワードパズルの鍵の様子

7. おわりに

本研究では，過去の新聞記事から抽出した年代ごとの特徴語をデータベースとして，クロスワードパズルを自動生成し，使用した単語に関する新聞記事へ誘導することで，各年代の社会的背景・歴史を学ぶ機会を提案することを目的として取り組んだ．現時点では，特徴語の抽出は完了しているが，難易度・カテゴリといった属性に基づくクラスタリングは実装段階である．また，言語学習システムとしての効果があるかについては調査段階であるため，これらが今後の課題となる．

謝辞

本研究では北海道新聞の新聞記事データの利用を許可頂きました．記事データの入手やシステム試作において北海道新聞長の三浦辰治氏に多大なるご協力を頂戴したので感謝申し上げます．

参考文献

[1] 日置竜輔，岩上慎之介，田中龍仁，保土沢朋和，伊藤太一，小山内魁人，中川翔真，金澤快飛，服部俊紀，寺沢憲吾，美馬のゆり，角康之，坂井田瑠衣．新聞記事との出会いを促す言葉遊びゲームの試作．第 26 回 一般社団法人情報処理学会シンポジウム．インタラクション 2022．2022，p.579-582

[2] Shilpa Patrick, Kirti Vishwakarma, Vishal P Giri, Debranjan Datta, Priyanka Kumawat, Preeti Singh, Prithpal S Mawat. The usefulness of crossword puzzle as a self-learning tool in pharmacology, Journal of Advances in Medical Education & Professionalism, 2018. 6. 4. 181 – 185

[3] Russell J. Pearson. Online Chemistry Crossword Puzzles prior to and during COVID-19: Light-Hearted Revision Aids That Work, Journal of Chemical Education. 2020. 97. 9. 3194-3200

[4] Rigutini Leonardo, Diligenti Michelangelo, Maggini Marco, Gori Marco. Automatic Generation of Crossword Puzzles, International Journal on Artificial Intelligence Tools (IJAIT). 2012