DEIM Forum 2022 B43-5

ウェブ検索エンジンを用いた地域関連トリビア情報の網羅的抽出

中野　裕介†　　山本　祐輔†

† 静岡大学情報学部　〒 432–8011 静岡県浜松市中区城北 3–5–1
E-mail: †nakano@design.inf.shizuoka.ac.jp, ††yamamoto@inf.shizuoka.ac.jp

あらまし　本研究では，日本の地域に関連するトリビアをウェブから自動抽出する方法について検討する．トリビアとはある事柄に対する興味を喚起するような面白さや意外性のある事実を指し，ウェブ検索など複数の場面でエンゲージメントに対する効果が指摘されている．本研究では，トリビアについて言及する際に典型的な言い回しをクエリとするウェブ検索を行い，検索結果スニペットから地域に関連するトリビアの主題を抽出する．また主題語として適切な語を評価するために，言い回しに対して期待されるジャンルの語かどうかを Wikipedia のカテゴリ情報を利用して判別する．提案手法によって抽出されたトリビア情報は，その事実性と面白さという点で対抗手法より優れていることが実験によって明らかとなった．

キーワード　トリビア抽出，トリビア推薦，Web 検索，テキストマイニング

1 はじめに

情報技術の発展によって，人々はウェブ検索などを通して多様な情報に簡単にアクセスできるようになった．ウェブ検索はもはや日常における一般的な行動であり，多くのユーザは目的をもってそれを行うのではなく，予期しない知識に出会うことを期待している側面もある [1] [2]．クエリに対して通常期待される検索結果に加え，関連性があるが意外な事実を含む検索結果を提示する検索システムでは，ユーザが関連情報の検索行動を続けたという事例が報告されている [3]．この点でトリビアは，ウェブページやアプリケーションへのユーザの滞在時間を高めることに貢献できる可能性が考えられる．

トリビアとは些末な知識だが，意外性や面白さを内包しているものである [4]．トリビアの例を表 1 に示す．これらは日本の特定地域に関するトリビアであり，1 列目に示した関係性に基づいて，地名とトリビアの主題となる要素が主語述語の関係として紐づけられている．例えば「滋賀県にはけいおんの聖地がある」というトリビアの場合，地名である「滋賀県」と主題である「けいおん」が「聖地」であるという関係性でつながっているとみなす．このようにして一般的に地名と主題のつながりが予測できないものが関係性によって結び付けられることで，それは意外性や面白さを内包する知識となる．

トリビアにはユーザの関心を喚起する効果が存在し，探索的なウェブ検索行動を促す関連性のある情報として利用することができる．またウェブ検索に限らず，実際に複数のケーススタディにおいて，トリビアを用いたゲームやウェブサイトがユーザの滞在時間を向上させた事例が指摘されている [1]．このことからビジネス的な応用が期待できるほか，ユーザが自身の検索行動が特定の範囲に限定されていたことを自覚させる効果も期待することができる [5]．

一方でトリビアを含めたある事柄の「面白さ」は一意に定義することが困難である．例えば Masood は，確立されたモデルにおいて，何らかの変化を示唆する記録やパターンは面白いという指摘を行っている [7]．すなわち特定の文脈で用いる話題や単語をグループ分けできればその中での特徴から面白さを類推できるが，そのグループ化を行うことは難しい．そのため既存のトリビアを抽出する研究では，データセットが豊富な特定の文脈を対象にした評価方法を研究していることが多い．

本稿では，ウェブ検索エンジンを用い，日本の地域に関連するトリビアを網羅的に自動収集する方法を提案する．日本の地域に関連するトリビアとは表 1 のような，地名を主語として，表 1 の一列目のような関係性において何らかの主題を述語とするトリビアである．提案手法では，地名と前述のような関係性を表す語をクエリとしたウェブ検索で得たタイトルとスニペットから，トリビアの主題となるような要素を形態素解析と固有表現抽出によって収集する．検索の際に，地名との関係性が真に適切な主題を抽出するために，クエリとする関係性については，カテゴリベクトルという出現する名詞句と直接関連を評価できる手法を用意する．これはある関係性における抽出したい名詞句のカテゴリを，Wikipedia[2]の構造を用いてあらかじめ設定しておくアプローチである．これらのプロセスを経ることで，日本の地域という領域で幅広い関係性で構成されたトリビアを網羅的に抽出・収集することが可能となる．また評価対象の名詞句が関係性に属するかどうかを，その名詞句に関するデータの量にかかわらず評価をすることが可能となる．

2 関連研究

近年トリビアは研究の題材としてたびたび取り上げられている．また以前から，文章や単語の関係を面白さや意外性の観点で評価しようとする試みはなされてきた．トリビアは一般的に

1：Using trivia and quiz products to engage your customer: http://www.slideshare.net/woverstreet/using-trivia-and-quiz-products-to-engage-your-customer

2：Wikipedia: https://www.wikipedia.org/

表 1　地域に関するトリビア例

地域	関係性	主題	トリビア
神奈川県	発祥の地	ナポリタン	神奈川県にはナポリタン発祥の地がある
滋賀県	聖地	けいおん	滋賀県にはけいおんの聖地がある
大阪府	最大級	水族館	大阪府には世界最大級の水族館がある
鹿児島県	モデル	もののけ姫	鹿児島県にはもののけ姫のモデルがある
滋賀県	ロケ	朝ドラ・スカーレット	滋賀県には朝ドラ・スカーレットのロケ地がある

自然言語的に人が理解できる形式をとるため，多くは成果物として短文を生成あるいは取得している．ここではそのアプローチとして，構造型と非構造型という 2 つにアプローチを大別して紹介する．

2.1　構造型アプローチ

ここで構造型と形容する手法においては，既存のデータセットなどを用いてそのエンティティの関係構造からトリビア的である要因を導き出そうとする試みがなされている．

Sahar は，興味深い関係を見つけるのではなく，興味のない関係を排除するためのプロセスを提案している．ここでは対象を排除するためのルールを，ユーザに提示する質問に基づいて設定することでプロセスを実現している [6]．Doi は，日本の郷土料理に関するトリビアを収集し，クイズ形式のアプリケーションとしての利用を提案した．トリビアの生成にあたっては，観光雑誌などから手動で郷土料理に関する情報を収集している [7]．Fatma は，興味深さの算出を行うために DBpedia という構造化されたデータ群を利用し，そこに存在するドメイン間の興味深さを基準とした CNN モデルを構築した [8]．Korn は，Wikipedia の表データを利用して，ある分野における最上級であることに関連するトリビアを生成した [9]．Tsurel は，Wikipedia のカテゴリ構造に着目し，カテゴリに属する記事の中での意外性を評価することで意外な人物がある属性を有していることを示すトリビアを作成する方法を提案した [10]．Lin は，希少性という概念に基づき，文書データセットの接続関係からあるカテゴリにおいて特殊であるものを検出する方法を提案した [11]．Kamigaito は，意外性に着目し，Wikipedia の人物に関する記事動詞の関係性を抽出してその関係同士が意外なものであるかどうかを評価する手法を提案した [12]．Merzbacher は，データベースからトリビアに関する質問を抽出するという問題に取り組んでおり，属性や値を関係式として抽出する方法を提案した [13].

2.2　非構造型アプローチ

ここで非構造型と形容する手法においては，機械学習などを用いて自然言語におけるトリビアを表現する特徴をルール化する．そしてそれに基づいて既存の文書の中から雑学的な特性を持つものを抜き出そうとする試みが為されている．

Prakash は，Wikipedia 上のテキストを WTM アルゴリズムという手法を用いてトリビア的であるかスコア付けする方法を提案した．IMDB というサイトに存在する映画にまつわるトリビアをラベル付きデータセットとして扱い，教師付き学習を用いて言語やエンティティに基づく特徴を抽出できるようにしている [4]．Niina は，SVR などを利用して，文章のトリビアスコアを推定する手法を提案した．この手法では文章の主題となる単語とその他の語の関係に着目し，それらのペアから特徴量を算出するというアプローチを行っている [14]．Kwon は，Wikipedia の記事要約を利用して，与えられたエンティティに対するトリビア文を抽出する手法について提案した [15]．Gamon は，ユーザのブラウジング行動を観測し，それらを特徴量とすることによって面白さをモデル化することを試みた．これに基づいてユーザが文書に興味をもつための潜在的な要因をとらえ，Wikipedia 記事中のリンクをクリックする確率の推定を行った [16].

2.3　本研究の位置づけ

本研究では，関係性を限定せず，汎用的なトリビアの収集に適用可能な手法を目指す．関連研究の多くでは関係性を限定するか，対象とする抽出領域を限定することでそのカテゴリにおける文脈のトリビア性を評価するものである．例えば映画に関するトリビアの収集を試みる場合は映画に関するトリビアを集めたデータセットからよりトリビアとして優れたテキストを得るためのルールの学習を試みる．このアプローチから導き出されたトリビアの評価指標はあくまで「映画」に限定されたもので，汎用的に有名人や料理などのトリビアに適用できる枠組みであるとはいえない．

本研究では，一部のトリビアに特化するのではなく，普遍的にトリビアに関係のある単語を評価するための手法を提案する．限定的な関係性を対象としないため，大規模なデータセットが存在しないマイナーな話題に対するトリビアを収集することも期待できる．またデータ収集の対象領域を特定のウェブサイトなどに限定しない．そのためある対象についてまとめられた既存のウェブサイトと比較して，提案手法による対象の抽出ではより広範な情報を入手できることが期待される．

3　提 案 手 法

本節ではウェブ検索によってトリビアに関連するテキスト情報を収集し，そこからトリビアを生成する方法について提案する．トリビアの生成の流れについては図 1 に示す通りである．

本稿の提案手法では，トリビアを地域 l，関係性 r，主題 p の 3 要素から成るトリプル（3 つ組）として定義し，ウェブから抽象・収集を行う．このトリプルは，表 1 に定義するような自然言語的なテキストとしてのトリビアを想起させることを志向したものである．トリビア t を構成する要素については，以下の式によって表すことができる．

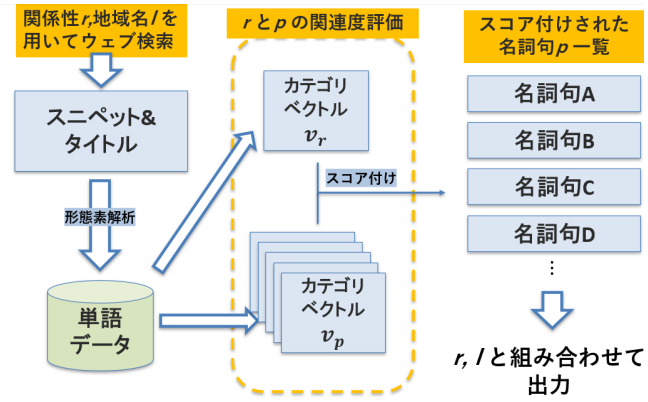

図 1　トリビア生成手法の概要

$$t = r(l, p)$$
$$s.t.\ \ l \ \text{is a} \ \ p \quad or \quad l \ \text{has a} \ \ p$$

トリビアを構成する要素として，地域名 l，関係性 r，主題 p が定義される．r は表 1 の 1 列目で例示されるように，様々な語で表現される．例えば 1 行目の「神奈川県にはナポリタン発祥の地がある」というトリビアでは，l に神奈川県，p にナポリタンが当てはまり，これらが r の発祥の地によって紐づけられることで，2 つの語の間のトリビア性を評価することが可能となる．この 3 つの要素によってトリプルは構成される．またこの構造から，広義に l と p の間の関係は，is-a あるいは has-a 関係に属していると解釈できる．

本研究では，上述のように定義された l，r，p で構成される名詞句トリプルをウェブから取得するため，カテゴリベクトルというものを用いる手法を提案する．カテゴリベクトルを用いたトリビアトリプルの生成は，下記のプロセスで行われる．

（1） 収集したいトリビアに関する地域 l，関係性 r を入力する．

（2） r と l を AND で結合したクエリでウェブ検索を行い，検索結果として得られるタイトルとスニペットを回収する．

（3） 回収したテキストを形態素解析，固有表現抽出し名詞句のみを抽出する．得られた名詞句が Wikipedia 上に記事タイトルとして存在するかどうかを調べ，当該記事に設定されたカテゴリを収集する．

（4） 収集したカテゴリを形態素解析し，出現頻度の高い 20 の単語を次元とし対象記事での出現回数の合計値を値としたものを r のカテゴリベクトル $\boldsymbol{v}_r$ と定義する．

（5） 同様に p の対象となる前述のウェブ検索で得られた各名詞句について，その名前を記事タイトルとする単一の Wikipedia 記事のカテゴリを収集し，r と同様に個々の p に対するカテゴリベクトル $\boldsymbol{v}_p$ として定義する．

（6） r と p のカテゴリベクトル $\boldsymbol{v}_r$，$\boldsymbol{v}_p$ を比較し，そのスコアが高い名詞句を最終的な p として扱い，r，l，p のトリプルをトリビアとして出力する．

以下に詳細を記す．

表 2　実際に用いた r 一覧

実際に用いた r	l	p
聖地	滋賀県	けいおん
生産量日本一	鹿児島県	オクラ
消費量日本一	岡山県	菓子パン
数が日本一	山梨県	セブンイレブン
日本初	熊本県	男女共学
偉人	岩手県	石川啄木
有名人	埼玉県	瀬戸大也
発祥の地	神奈川県	ナポリタン
B 級グルメ	三重県	四日市とんてき

表 3　カテゴリベクトルの一部（聖地）

次元	年	作品	人物	英語	企業	アニメ	日本語	映画	文化
値	62	50	41	38	36	35	32	29	27

3.1　データの収集

本節では，関係語 r と関連するカテゴリベクトルの生成や主題語 p の候補語となるテキストデータの収集方法について記述する．

まずは収集したいトリビアに関連する関係語 r を指定する．r は「発祥の地」や「日本一」などのように，地域に対するトリビアを表現する際に用いられる語である．r の例を表 2 に示す．

次に，用意した r と l を用いてウェブ検索を行う．ウェブ検索では，表 2 に示す r それぞれに対し，l として 47 の都道府県名を組み合わせたペアを用い,「“北海道” “聖地”」のようなクエリを用いるフレーズ検索を行う．各都道府県ごとの検索結果のうち上位 100 件のタイトルとスニペットをテキストデータとして収集し，形態素解析を行う．このとき得られた名詞及び固有表現について，地名に関連しないものを本研究で用いる名詞句データとして収集する．

3.2　カテゴリベクトルの作成

本節では関係語 r と主題語 p を直接比較し，r に対する適切な p を抽出するために用いるカテゴリベクトルについて記述する．

カテゴリベクトルは，Wikipedia カテゴリに含まれる単語を次元とする．このとき Wikipedia カテゴリを単語に分解して用いるのは，類似した意味合いであっても異なる Wikipedia カテゴリのバリエーションが多く存在することから，その集約を行うためである．各次元の値は各カテゴリ名が p もしくは r でウェブ検索した際のスニペットに出現する頻度であり，表 3 のように表される．これは r を用いた検索結果あるいは p がどのようなカテゴリに属しているのかを表現する方法である．例えば r=“聖地”の場合，作品やアニメといったカテゴリ語がウェブページ中で高頻度で出現すると考えられる．ここから r が聖地であるウェブ検索によって取得された単語データに含まれる単語にアニメ作品のタイトルが含まれていた場合，その単語がアニメ作品であることを聖地と同様のカテゴリベクトルで表現できれば，ベクトル間の類似性から聖地における p として適切であるかを評価できるといえる．このようにカテゴリベクトル

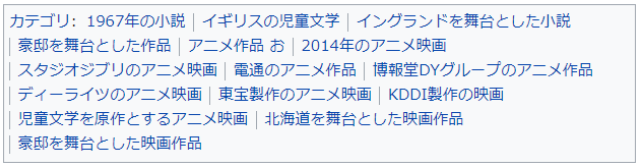

図 2　Wikipedia 記事のカテゴリ（思い出のマーニー）

を用いることで，p と r という異なるコンセプト同士の文脈における意味の類似性を直接評価することが可能になる．

r のカテゴリベクトル $\boldsymbol{v_r}$ は，以下のプロセスで作成される．

（1） 当該 r をクエリとして収集した単語データを対象に，含まれる単語それぞれがタイトルとなった Wikipedia 記事が存在するかどうかを検索する．

（2） 記事が存在した場合，各記事に設定された図 2 のようなカテゴリ名を収集する．

（3） 得られたカテゴリ名それぞれを形態素解析し，出現した単語を出現頻度順に記録する．またこの時，多くの単語が該当記事となる Wikipedia 記事のカテゴリに頻出する，「姓，名，日本」を対象から除外する．

（4） 単語の出現頻度が上位の 20 単語を次元とし，出現頻度を値としたベクトルを，r のカテゴリベクトル $\boldsymbol{v_r}$ として定義する．

また r のカテゴリベクトルとの比較に用いる単語データの各 p 候補のカテゴリベクトルについては，Wikipedia 記事が存在した場合に上記プロセスと同様に，p をタイトルとする一つの記事のカテゴリのみを用い，そのカテゴリに含まれる単語を次元とし，出現頻度を値とする p のカテゴリベクトル $\boldsymbol{v_p}$ とする．

3.3　r と p の関連度の評価

本節では，上述するカテゴリベクトルを用いた r と p の関連度評価の方法について述べる．

r のカテゴリベクトル $\boldsymbol{v_r}$ と p 候補のカテゴリベクトル $\boldsymbol{v_p}$ の関連性をスコアとして算出する方法は以下の式で表される．これは各次元間の関連性とその重みを反映した計算を行うことを意図しており，r に関連性の高いカテゴリを持つ p がより高いスコアを得られることを意図している．この評価式を用いることによって，p と r という異なるコンセプトや次元をもったもの同士をカテゴリベクトルとして直接評価することを目的としている．

$$\boldsymbol{v}_r = (f_r^{(1)}, f_r^{(2)}, \cdots f_r^{(m)})$$

$$\boldsymbol{v}_p = (f_p^{(1)}, f_p^{(2)}, \cdots f_p^{(n)}) \quad \text{とするとき}$$

$$Score(r,p) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} f_r^{(i)} f_p^{(j)} \cdot cos(\boldsymbol{w}_r^{(i)}, \boldsymbol{w}_p^{(j)})$$

ここで m は r のカテゴリベクトルの次元数を表し，n は p 候補のカテゴリベクトルの次元数を表す．また $\boldsymbol{w}_r$ と $\boldsymbol{w}_p$ は r と p 候補それぞれの各次元の意味する単語に対応する Word2Vec [17] を用いて得られた単語ベクトルであり，cos は単語ベクトル間のコサイン類似度である．また f_r，f_p は各次元に対応する値

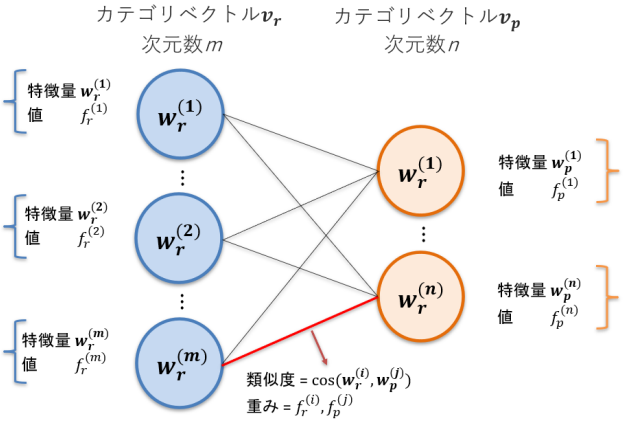

図 3　カテゴリベクトルの計算イメージ

を示しており，カテゴリベクトルを作成する過程で対応する各名詞句が出現した回数である．r と p 候補それぞれのカテゴリベクトルは次元数が異なり，また各次元の意図する値も異なるといえる．例えば，表 3 に一部を示した聖地のカテゴリベクトルは全部で 20 次元であるのに対し，あるアニメのタイトルを同様の方法で p 候補のカテゴリベクトル化すると，出現した単語数が少なく次元数は 20 以下となる場合も想定できる．そのため上記の式ではベクトル同士の各次元を図 3 のように総当たりで比較し，次元間の単語の類似性と出現頻度の 2 つにより，各要素を適切な重さで計算する．これにより異なる次元や異なる特徴量を持ったベクトル同士を計算することが可能となる．この手法を用いて，名詞句データについて対応する r との計算を行い，その点数が上位であったものを r に対する最終的な p として扱う．

以上の手続きによって r に対する p とその検索に用いた l を得て，その組み合わせを最終的な出力とする．

4　実　　験

本節では，提案手法によって生成されたトリビアの品質及び提案手法の機能の有効性について評価を評価するための実験について示す．

4.1　比 較 手 法

提案手法によるトリビアがトリビアとして面白いものであるのか，また事実であるのかについてその有効性を評価するための提案手法および 3 つの比較手法を以下の通り用意する．

（1） 提案手法

（2） コサイン類似度を用いてのカテゴリベクトルの比較

（3） EMD によるカテゴリベクトルの比較

（4） 各単語の検索ヒット数に基づく順位付け

4.1.1　コサイン類似度を用いてのカテゴリベクトルの比較

カテゴリベクトル同士を比較するにあたり，コサイン類似度を用い，ベクトル同士の比較を行う．なお比較にあたり r と p 候補それぞれのカテゴリベクトルの次元数が異なる場合が想定

される．そのためそれぞれの次元における単語に基づき，2 つのカテゴリベクトルの単語の重複なしの合計をそれぞれの次元数として統一する．またそのとき元のベクトル上で値が設定されていない次元については値は 0 として扱うものとする．

$$Score(r,p) = cos(\boldsymbol{v_r}, \boldsymbol{v_p})$$

4.1.2 EMD によるカテゴリベクトルの比較

Earth mover's distance(EMD) [18] は，輸送問題における最適化手法の 1 つであり，本研究において以下のように定式化することが可能である．

$$EMD(r,p) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}$$

r と p 候補の 2 つのベクトルの距離を算出するとき，両者の各次元の値を総当たりで計算する．d は r と p 候補のすべての次元ごとの組み合わせの距離であり，Word2Vec [17] における次元の特徴語のコサイン類似度を $cos(\boldsymbol{w}_r^{(i)}, \boldsymbol{w}_p^{(j)})$ とするとき以下のように定式化できる．

$$d_{ij} = 1 - cos(\boldsymbol{w}_r^{(i)}, \boldsymbol{w}_p^{(j)})$$

また f は r と p 候補のすべての組み合わせにおける最適化された重みの割り当てであり，各次元の値を用いて算出される．これによって算出される値を用いて，p 候補の順位付けを行う．

4.1.3 各単語の検索ヒット数に基づく順位付け

p 候補の各単語を元の l とともにウェブ検索エンジンでフレーズ検索する．その際に得られるヒット件数を評価指標として p 候補の評価を行う．

4.2 アンケート項目

提案手法と比較手法を用いて収集したトリビアについて，それがトリビアとして成立しているのかどうかについての以下の指標を用いる．

（1） 事実性：トリビアとして提示するテキストは事実であるか

（2） 面白さ：トリビアとして面白いと評価できるか

それぞれの指標はリッカート尺度を用い，三段階あるいは五段階で評価を行う．また事実性の評価を行うにあたっては，提示するトリビアごとにウェブ検索等を用いて事実であるかの調査を行ってもらう．

評価対象が単語群となることから，事実であるかの評価にあたってはその組み合わせによって，評価者が l における p について，r で示されるような関係性が存在するかどうかを主観的に評価する．面白さについても同様に，主観的に想定しうる意味を見出し，かつ面白いと感じられるかを評価する．

各指標の尺度は事実性については「事実ではない (1)」,「おそらく事実である (2)」,「事実である (3)」を，面白さについては「面白くない (1)」,「やや面白くない (2)」,「どちらともいえない (3)」,「やや面白くない (4)」,「面白い (5)」を用いた．

4.3 実験手法

表 2 に示す 9 つの r について，手法毎に関係性 r の上位 50 件ずつの名詞句について上述の事実性と面白さを評価してもらった．評価者はクラウドソーシングサービスによって募り，1 件の名詞句につき 3 人ずつ評価を行った．評価は 1 人当たり 50 件行ってもらい，200 円の謝礼を支払った．評価は手法及び r ごとに行われ，108 人の参加者によって実施された．

4.4 ランキング評価指標

アンケートによって 9 つの r について，提案手法および 4 つの対抗手法によって作成したランキングに事実性と面白さに対する評価を行った．このときランキングを評価する手法として以下の 2 つの指標を用いた．

4.4.1 適　合　率

適合率は，上位 k 件における r ごとに，事実性と面白さという基準での正しい p の割合を表すものである．ここで上位 k 件中の各 p について事実性においては評価の平均値が 2 より大きいものを，面白さにおいては 3 より大きいものを正しいとみなした．上位 k 件の適合率 ($p@k$) は，以下の式によって表される．

$$p@k = \frac{k \text{ 件中の正しい } p \text{ の個数}}{k}$$

4.4.2 nDCG

$nDCG$ は各手法によって得られたランキングの予測能力を評価する指標である [19]．$nDCG$ はランキング上での出現順を考慮した重みづけをする指標 DCG を，理想のランキングで作成された DCG（$IDCG$）で割ることで正規化した値である．ここで $nDCG$ は，ある手法において対象 r の事実性あるいは面白さにおける前述のスコア集合 A_k を用いて以下で表される．

$$A_k = (a_1, a_2, \cdots a_k) \quad \text{とするとき}$$

$$DCG_k = \sum_{i=1}^{k} \frac{a_i}{log_2(i+1)}$$

$$nDCG_k = \frac{DCG_k}{IDCG_k}$$

このとき $IDCG$ には，ある r において提案手法を含む 4 つの手法において上位 50 件にランキングされた計 200 件の p を高スコア順に並び変えた上位 50 件を A_k として用いた．

5 結　　果

本節では，9 つの r について提案手法及び対抗手法によって評価した実験の結果を示す．上位 10 件，30 件，50 件での適合率を事実性については表 5 に，面白さについては表 6 示す．k 件ごとの適合率の値は $p@k$ として示される．手法別の各 r のランキング上位 50 件についての事実性の $nDCG$ を表 7 に，面白さの $nDCG$ を表 8 に示す．

5.1 事　実　性

事実性については表 5 に適合率，表 7 に nDCG を評価した結果を示す．全ての指標において提案手法は最多の関係性 r で

最も高い値を示した．nDCG においては「発祥の地」以外の r で対抗手法よりも優れた結果であり，手法別平均値も最大であった．

5.2 面 白 さ

面白さについては表 6 に適合率，表 8 に nDCG を評価した結果を示す．全ての指標において提案手法は最多の関係性 r で最も高い値を示した．nDCG においては 5 つの r で最大の値をとり，手法別平均値も最大であった．一方で「偉人」と「有名人」の適合率において対抗手法の方が高い値を示している．

5.3 出力結果例

提案手法による実際の出力例を表 4 に示す.「聖地」や「偉人」といった関係性 r においては適切な関連のある用語が上位に多く出現している．正解が多い r では上位 5 件でもアニメ作品や学者といった特定の対象が主題語 p として出現している傾向がうかがえる．その一方で,「生産量日本一」や「B 級グルメ」といった r では，食品にまつわる語が出現していると考えられるが，実際には地域名 l と関係のないものであることがわかる．

6 考　　察

本節では，提案手法についての考察と今後の応用への展望を述べる．

6.1 評価実験の結果について

評価実験においては，多くの関係性 r において，提案手法が最も優れた性能を示した．nDCG においては事実性で 8 つ，面白さで 5 つの r で対抗手法を上回ったほか，適合率でも最も良い結果を示す r の個数で最多となった．ここから提案手法はトリビア抽出の手法として一定の能力があったと考えられる．

一方で r によって提案手法の結果の値は大きく異なった．提案手法は本質的に r と p の関係性を直接計算可能にするものであり，直接的に事実性や面白さを評価しているわけではない．そのため，r のカテゴリベクトル作成や p 候補の取得に用いたスニペットによって最終的な結果が簡単に左右されてしまう可能性が考えられる．例えば r として用いられた語に関係のある記事が少ないと，関連度の低いページがスニペットとして得られる可能性がある．それらから p や r のカテゴリベクトルが作成されることによって，もともと適切でない p を適切でないカテゴリベクトルで計算してしまう可能性は増加する．

この点から事前にスニペットから適切な語を抜き出すことを検討する必要があると考えられる．スニペットから適切な単語を抽出する研究として，大島が行った両方向構文パターンによる関連語抽出があげられる [20]．この手法ではある単語の関連語を取得するために，取得したい単語の前後に出現するパターンを用意してそれぞれ検索を行う．こうすることで結果の信頼性を高め，関連度の高い単語の取得が可能となる．これを応用して本手法における r の両方向構文パターンを作成し，検索を行うことにより，事前に収集される単語品質のさらなる向上が期待できる．

6.2 カテゴリベクトルについて

本研究では，r と p の次元数の異なるベクトル同士の関連性を計算するカテゴリベクトルという手法を提案した．事前に r に対応する 20 次元のカテゴリベクトルを設定することで，それぞれの r での出現語の傾向を定量的に評価できるようになった．一方で実験では，同様のカテゴリベクトルを用いたにもかかわらず提案手法と EMD の間で結果に大きな差が見られた．この 2 つの手法は類似しており，結果も類似した値になるはずである．この点で，本研究で作成したカテゴリベクトルには本来 p として期待されない結果を高評価とするリスクが存在していると考えられられる．例えば r が聖地である場合，カテゴリベクトルはアニメや映画に関係する p を評価する状態であることが期待される．しかし実際には表 3 に示すように，直感的には無関係であるような単語も含まれていることがわかる．この点で提案手法と EMD の間では，カテゴリベクトルの各次元に対する重みづけの違いが，異なる結果につながったと考えられる．

ここからカテゴリベクトルの作成にあたり，その対象とするカテゴリについてより検討することが重要であると考えられる．Wikipedia カテゴリに関しては，その構造に関する研究が盛んに行われている [21]．カテゴリは記事に付与されたテキストデータというだけでなく，階層構造を持った概念であり，その特徴からいくつかの種類に分類することもできると指摘される．ある親カテゴリを持つ子カテゴリは例えば「XX 年の映画作品」のようなフォーマットでいくつものバリエーションをもつこともあり，単語ごとの出現頻度にも大きな差がある．それらの特性はカテゴリベクトルを作成するにあたり，一部の単語に大きな影響力を与える可能性がある．

特性を適切に考慮した重み付けなどを行うことによって，上述の無関係な単語の出現や影響力の差を軽減し，より公平に多様な r に対応することが可能になると考えられる．

6.3 他分野への応用

本研究の提案手法は，地域に関するトリビア情報の取得を指向したものである．一方で本質的には入力語の意図する関係に適切な語の収集手法と一般化することができ，地域にとどまらない汎用的な応用が期待できる．例えば地域名 l の代わりに食品の名称を入力することで食品に関するトリビア情報を取得できる可能性がある．またトリビア情報ではなく，単純に関係語 r に対する語の上位概念や下位概念を取得することもできる可能性がある．いずれの応用についてもカテゴリベクトルの利用法には改善が必要だと考えられるが，発展的なアプローチとして検討していく．

6.4 想定アプリケーション

今回の手法によって得られたようなトリビア情報は，ウェブサービスにおけるユーザの滞在時間の向上などの効果が期待できる．またトリビア情報そのものがコンテンツとして消費可能である．そのためトリビア情報を位置情報と紐づけた観光アプリケーションなどを開発することなどが考えられる．観光情報

表 4　各関係性の出力結果上位 5 件 (一部)(太字は事実性の正解データ)

関係性 r		1	2	3	4	5
聖地	**地域名 l**	**山口県**	**青森県**	**大分県**	**長崎県**	**高知県**
	主題語 p	**サバイバルファミリー**	**シャーマンキング**	**ReLIFE**	**ちゃんぽん**	**竜とそばかすの姫**
生産量日本一	**地域名 l**	奈良県	和歌山県	福井県	**鹿児島県**	**北海道**
	主題語 p	おやつ	ヨーグルト	ピラフ	**オクラ**	**農作物**
消費量日本一	**地域名 l**	富山県	滋賀県	千葉県	奈良県	大阪府
	主題語 p	ロシア人	えび豆	ビーフン	柿の葉寿司	五平餅
日本初	**地域名 l**	岩手県	**島根県**	京都府	**滋賀県**	岡山県
	主題語 p	沢田	**松原亘子**	教育課程	**シトロエン**	児島
偉人	**地域名 l**	**群馬県**	**島根県**	**大分県**	**東京都**	山口県
	主題語 p	**萩原朔太郎**	**潮恵之輔**	**広瀬淡窓**	**葛飾北斎**	迫田
有名人	**地域名 l**	**長野県**	**愛媛県**	石川県	宮城県	**宮崎県**
	主題語 p	**酒井宏樹**	**長友佑都**	丸山桂里奈	香川真司	**伊野波雅彦**

表 5　手法ごとの適合率 (事実性)(太字は条件ごとの最大値)

関係性 r	$p@k$	提案手法	ヒット数	コサイン	EMD
聖地	*k=10*	**1.000**	0.100	0.000	0.500
	k=30	**0.767**	0.133	0.033	0.333
	k=50	**0.640**	0.140	0.120	0.360
生産量日本一	*k=10*	0.200	0.100	**0.300**	0.200
	k=30	**0.400**	0.100	0.233	0.167
	k=50	**0.400**	0.160	0.220	0.180
消費量日本一	*k=10*	**0.500**	0.200	0.000	0.000
	k=30	**0.433**	0.133	0.067	0.033
	k=50	**0.400**	0.100	0.060	0.100
数が日本一	*k=10*	**0.100**	0.000	**0.100**	0.000
	k=30	**0.200**	0.000	0.033	0.067
	k=50	**0.260**	0.020	0.040	0.040
日本初	*k=10*	**0.500**	0.200	**0.500**	0.400
	k=30	**0.633**	0.333	0.333	0.367
	k=50	**0.600**	0.400	0.380	0.400
偉人	*k=10*	0.600	0.100	0.600	**0.700**
	k=30	**0.800**	0.033	0.600	0.700
	k=50	0.700	0.020	0.520	**0.760**
有名人	*k=10*	**0.900**	0.300	0.500	0.100
	k=30	**0.767**	0.333	0.433	0.167
	k=50	**0.760**	0.340	0.400	0.200
発祥の地	*k=10*	0.600	0.200	**0.800**	0.400
	k=30	0.633	0.233	**0.667**	0.433
	k=50	**0.700**	0.360	0.660	0.460
B 級グルメ	*k=10*	**0.300**	0.100	0.000	0.000
	k=30	**0.233**	0.100	0.033	0.000
	k=50	**0.160**	0.120	0.060	0.000

表 6　手法ごとの適合率 (面白さ)(太字は条件ごとの最大値)

関係性 r	$p@k$	提案手法	ヒット数	コサイン	EMD
聖地	*k=10*	**0.800**	0.100	0.000	0.400
	k=30	**0.600**	0.133	0.067	0.267
	k=50	**0.500**	0.160	0.120	0.280
生産量日本一	*k=10*	0.200	0.100	**0.500**	0.300
	k=30	**0.333**	0.167	0.300	0.233
	k=50	**0.380**	0.220	0.260	0.280
消費量日本一	*k=10*	**0.500**	0.100	0.000	0.000
	k=30	**0.467**	0.100	0.067	0.033
	k=50	**0.440**	0.140	0.080	0.040
数が日本一	*k=10*	0.000	0.000	**0.100**	**0.100**
	k=30	0.100	0.000	0.067	**0.133**
	k=50	**0.080**	0.000	0.040	**0.080**
日本初	*k=10*	**0.600**	0.300	0.000	0.400
	k=30	**0.667**	0.400	0.000	0.300
	k=50	**0.600**	0.480	0.000	0.360
偉人	*k=10*	0.100	0.000	0.500	**0.700**
	k=30	0.233	0.000	**0.600**	0.567
	k=50	0.280	0.020	0.520	**0.600**
有名人	*k=10*	0.100	0.200	**0.500**	0.000
	k=30	0.067	0.100	**0.400**	0.133
	k=50	0.120	0.120	**0.280**	0.100
発祥の地	*k=10*	0.500	0.200	**0.600**	0.300
	k=30	**0.500**	0.167	**0.500**	0.367
	k=50	**0.540**	0.160	0.400	0.360
B 級グルメ	*k=10*	**0.200**	0.100	0.000	0.100
	k=30	**0.100**	0.067	0.033	0.033
	k=50	**0.080**	**0.080**	0.040	0.040

としてはあまり知られていない知識を提供することができれば，現実空間においてウェブサービスのような滞在時間の向上などの効果をもたらすことも考えられる．

7　おわりに

本稿では，地域に関連するトリビアを生成するために，l，r，p という 3 つの概念を設定し，それらの単語群を収集する手法としてのカテゴリベクトルを用いたアプローチを提案した．関係語 r と地域名 l をクエリとしたウェブ検索を行い回収したスニペットから p 候補を抽出し，Wikipedia のカテゴリを利用を利用して r と p を直接評価することで関連性をスコア付けした．

また実験によって r による差異はあるものの，対抗手法より

表 7　手法ごとの nDCG(事実性)(太字は条件ごとの最大値)

関係性 r	提案手法	ヒット数	コサイン	EMD
聖地	**0.828**	0.567	0.542	0.678
生産量日本一	**0.636**	0.516	0.610	0.548
消費量日本一	**0.761**	0.519	0.500	0.516
数が日本一	**0.660**	0.540	0.525	0.578
日本初	**0.758**	0.708	0.688	0.700
偉人	**0.863**	0.487	0.742	0.766
有名人	**0.898**	0.622	0.654	0.572
発祥の地	0.751	0.588	**0.773**	0.673
B 級グルメ	**0.693**	0.615	0.591	0.508
手法別平均値	**0.761**	0.574	0.625	0.615

優れた事実性と面白さを持った情報を取得できることが明らかとなった．今後は現在の手法がより汎用的な関係性に適用でき

表 8　手法ごとの nDCG(面白さ)(太字は条件ごとの最大値)

関係性 r	提案手法	ヒット数	コサイン	EMD
聖地	**0.767**	0.537	0.523	0.657
生産量日本一	**0.647**	0.528	0.599	0.547
消費量日本一	**0.750**	0.508	0.529	0.327
数が日本一	0.571	0.412	0.493	**0.581**
日本初	**0.772**	0.712	0.460	0.637
偉人	0.630	0.428	0.733	**0.747**
有名人	**0.724**	0.528	0.783	0.543
発祥の地	0.725	0.549	**0.687**	0.670
B 級グルメ	0.570	**0.602**	0.523	0.474
手法別平均値	**0.684**	0.534	0.592	0.576

ることを目指す．またトリビアの対象となった事物に関する興味関心を引き起こすようなアプリケーションへの応用についても検討していく．

謝辞 本研究は JSPS 科研費 JP18H03244，21H03554，21H03775 の助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Paul André, Jaime Teevan, and Susan T. Dumais. From x-rays to silly putty via uranus: Serendipity and its role in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, p. 2033–2036, New York, NY, USA, 2009. Association for Computing Machinery.

[2] Amanda Spink, Howard Greisdorf, and Judy Bateman. From highly relevant to not relevant: Examining different regions of relevance. *Inf. Process. Manage.*, Vol. 34, No. 5, p. 599–621, September 1998.

[3] Iris Miliaraki, Roi Blanco, and Mounia Lalmas. From "selena gomez" to "marlon brando": Understanding explorative entity search. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, p. 765–775, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee.

[4] Abhay Prakash, Manoj K Chinnakotla, Dhaval Patel, and Puneet Garg. Did you know?- mining interesting trivia for entities from wikipedia. *Proceedings of Twenty-Fourth International Joint Conference on Artificial Intelligence,*, 2015.

[5] Yelena Mejova, Javier Borge-Holthoefer, and Ingmar Weber. Bridges into the unknown: Personalizing connections to little-known countries. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, p. 2633–2642, New York, NY, USA, 2015. Association for Computing Machinery.

[6] Sigal Sahar. Interestingness via what is not interesting. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, p. 332–336, New York, NY, USA, 1999. Association for Computing Machinery.

[7] Syunya Doi, Yiqing Wang, Chen Zhao, and Takehito Utsuro. Design of a trivia game for traveling and domestic enjoyment in japan. In *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication*, IMCOM '17, New York, NY, USA, 2017. Association for Computing Machinery.

[8] Nausheen Fatma, Manoj K. Chinnakotla, and Manish Shrivastava. ¡i¿the unusual suspects¡/i¿: Deep learning based mining of interesting entity trivia from knowledge graphs. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, p. 1107–1113. AAAI Press, 2017.

[9] Flip Korn, Xuezhi Wang, You Wu, and Cong Yu. Automatically generating interesting facts from wikipedia tables. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, p. 349–361, New York, NY, USA, 2019. Association for Computing Machinery.

[10] David Tsurel, Dan Pelleg, Ido Guy, and Dafna Shahaf. Fun facts: Automatic trivia fact extraction from wikipedia. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, p. 345–354, New York, NY, USA, 2017. Association for Computing Machinery.

[11] Shou-de Lin and Hans Chalupsky. Using unsupervised link discovery methods to find interesting facts and connections in a bibliography dataset. *SIGKDD Explor. Newsl.*, Vol. 5, No. 2, p. 173–178, December 2003.

[12] Hidetaka Kamigaito, Jingun Kwon, Young-In Song, and Manabu Okumura. A new surprise measure for extracting interesting relationships between persons. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 231–237, Online, April 2021. Association for Computational Linguistics.

[13] Matthew . Automatic generation of trivia questions. In *Proceedings of the 13th International Symposium on Foundations of Intelligent Systems*, ISMIS '02, p. 123–130, Berlin, Heidelberg, 2002. Springer-Verlag.

[14] Kazuya Niina and Kazutaka Shimada. Trivia score and ranking estimation using support vector regression and RankNet. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong, 1–3 December 2018. Association for Computational Linguistics.

[15] Jingun Kwon, Hidetaka Kamigaito, Young-In Song, and Manabu Okumura. Hierarchical trivia fact extraction from Wikipedia articles. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4825–4834, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[16] Michael Gamon, Arjun Mukherjee, and Patrick Pantel. Predicting interesting things in text. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1477–1488, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[17] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, p. II–1188–II–1196. JMLR.org, 2014.

[18] Tomasi C. Rubner, Y and L.J Guibas. The earth mover's distance as a metric for image retrieval. In *Proceedings of International Journal of Computer Vision 40*, p. 99–121, 2000.

[19] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. Vol. 20, p. 422–446, New York, NY, USA, oct 2002. Association for Computing Machinery.

[20] Hiroaki Ohshima and Katsumi Tanaka. High-speed detection of ontological knowledge and bi-directional lexico-syntactic patterns from the web. *Journal of Software*, Vol. 5, pp. 195–205, February 2010.

[21] Masaharu Yoshioka. Analysis of japanese wikipedia category for constructing wikipedia ontology and semantic similarity measure. pp. 470–481, 12 2014.