Vol.2021-DBS-173 No.12
Vol.2021-IFAT-144 No.12
2021/9/17

ウェブ検索エンジンを用いた地域関連雑学情報の自動抽出

中野 裕介[1,a)]　山本 祐輔[1,b)]

概要：
本研究では日本の地域に関連する雑学文をウェブから自動抽出する方法について検討する．雑学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す．本研究では，雑学について言及する際に典型的な言い回しをクエリとしウェブ検索を行い，検索結果スニペットから地域に関連する雑学要素要素を抽出する．さらに，得られた雑学要素と地域名をテンプレートに当てはめることで雑学文の生成を目指す．

1. はじめに

今日において，特定の事物に関する情報を集めるためにウェブ上に存在する文書等を利用することは一般的である．また様々な領域でそうしたデータを収集することを目的とした研究は盛んに行われているが，対象とするものによっては困難を伴う場合も多く，雑学はその一例として考えられる．雑学はトリビアとも表現され，これを対象とした文書の自動生成や情報検索の研究においてはいずれも「人の注意や関心を引き起こす特定の事物に関する瑣末な知識」として扱われてる [1][2][3][4]．この定義において雑学は様々な領域において存在し，映画などのフィクション作品の裏話から建造物の高さや歴史を対象にするものなど多岐にわたる．それゆえに文書における雑学の表現方法は様々であり，語義を満たすか否かの機械的判定には困難を伴う．過去に行われた研究では大量のデータを機械学習することで雑学の定義の定量的な判断を試みたり [1][2]，データの収集領域を一部のウェブサイトに限定しての収集を試みたりした [3][4]．しかしこうした手法では，前者では自然言語的な文書の品質を担保することが課題となり，後者では検索領域が狭く，汎用的にウェブ上に存在する様々な文書に適用することが難しい．

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雑学情報を網羅的に自動収集する方法を提案する．地域に関連するというのは雑学情報を述語とした場合に地域が主語の関係となるようなものを指す．例えば「ひな祭り発祥の地」や「日本一深い湖」といった名詞句を雑学情報とした場合，これらは自明に特定の地域との関連を有しており「和歌山県はひな祭り発祥の地」や「秋田県には日本一深い湖がある」といった地域名を主語，雑学情報を述語とする関係を想定することができる．本稿ではこうした関係をもつ雑学情報を扱う名詞句を「雑学名詞句」と定義する．その上で地名と雑学について言及する際に典型的な言い回しである表 1 に示すような「雑学テーマ」をクエリとしたウェブ検索を行う．その結果得られたスニペットから雑学名詞句となりうる文章のまとまりを抽出し，地名との主語述語関係を持つ文章（雑学文）とする．さらに地名と雑学名詞句から雑学文を生成するためにテンプレートを用い，これに当てはめた場合の文法的な整合性評価を行う．またその雑学情報が実際にクエリとした地名と関連しているのかについても評価を行った．

また上記によって得られた雑学名詞句集合について，テンプレートに適用した場合に雑学文として成立するかどうかを検証した．提案手法と一部の提案手法を用いた 2 つの対抗手法について各雑学名詞句集において成立可能な品質の雑学名詞句の割合を，人による分類から収集した．その結果，提案手法は対抗手法と比較して雑学名詞句を分類できていることが明らかになった一方で，文法的な破綻や地域との実際の関連がないものも依然として見受けられた．これらの解決方法については今後の課題として提示した．

2. 関連研究

Niina らはウェブ上からトリビア文を収集し，その主語と文集に含まれる各単語の関係性やその珍しさを特徴量として SVR による回帰分析を行い，Wikipedia 上の文章のトリビア度合いの算出を試みた．またあわせて Ranknet と呼ばれるランキング学習を行うニューラルネットワー

[1] 静岡大学
Shizuoka University, Hamamatsu, Shizuoka 432–8011, Japan
[a)] nakano@design.inf.shizuoka.ac.jp
[b)] yusuke_yamamoto@acm.org

Vol.2021-DBS-173 No.12
Vol.2021-IFAT-144 No.12
2021/9/17

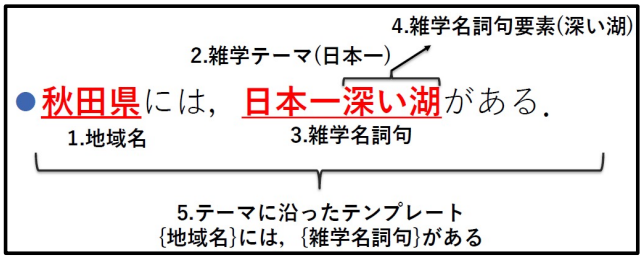

図 1 本稿における雑学文とその構成要素

クに基づいたトリビア度合いについても算出し，前者よりも高い精度を示した.[1] Prakash らは IMDB というサイトの映画に関連するトリビアを用い，それに付与された評価スコアを教師データとした機械学習を行った．それによって Wikipedia 上からあるエンティティに対する関連トリビア文をランキング化して出力するシステムを提案した.[2] Korn らは Wikipedia 上にある表データに着目し，各表の持つ属性について主題として比較を行うことでランキングに関連した雑学文を作成する方法を提案した.[4] Doi らは国内の観光情報に着目し，郷土料理に関するトリビアのクイズを半自動的に生成することを提案した.Web 上から集めた料理に関する情報をガイドブックなどを用いて補完し，また Wikipedia の類似する料理情報とともに用いることで四択クイズを作成した.[3]

上記のように，Wikipedia テキストの利用を中心として機械学習やルールを設定することで雑学情報を収集しようとする試みは盛んに行われているが，ウェブ上のすべてのテキストを対象とした雑学情報の抽出を試みる研究はあまり見られない．

3. 提案手法

本節ではウェブ検索によって雑学に関連するテキスト情報を収集し，そこから雑学文を生成する方法について提案する．雑学分の生成の流れについては図 2 に示す通りである．雑学文として読解可能かの文法的評価と真偽評価は，スニペットから抽出した名詞句が雑学文の構成要素として適当かどうかを選別するために行った．以下でその詳細について述べる．

3.1 雑学文とその構成要素の定義

ここで提案手法に用いる各要素の定義について記述する．

雑学文を構成する要素は主に地域名，雑学名詞句，テンプレートの 3 つあり，以下の式で表すことができる．

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

雑学文 $trivia$ はテンプレート t から構成され，地域名 l と雑学名詞句 p が必ず含まれる．例えば図 1 においては l に「秋田県」，p に「日本一深い湖」が該当し，「l には，p がある」というテンプレートにそれらを当てはめることで「秋田県には，日本一深い湖がある」という $trivia$ とする．また t 内において l と p は is-a 関係あるいは has-a 関係のいずれかをもつ．これはテンプレート上でのおける l と p の概念的な連接を意味する．すなわち has-a 関係の場合，「日本一深い湖」の例文のように l が雑学情報となりうるスポットなどである p を有していることを指し，is-a 関係である場合「ひな祭り発祥の地」のように，p 自体が l を表す．

また p についても「雑学テーマ」と「雑学要素」という 2 つの属性を持ち，前者は表 1 に示すような雑学について言及する際に典型的な言い回しを表す．また後者は雑学情報の主題となりうる要素を表し，前述の例における「深い湖」や「ひな祭り」が該当する．これらを組み合わせることによって，「日本一深い湖」や「ひな祭り発祥の地」のように，2 つの属性が修飾-被修飾の関係を持つような構造の p を得ることができる．

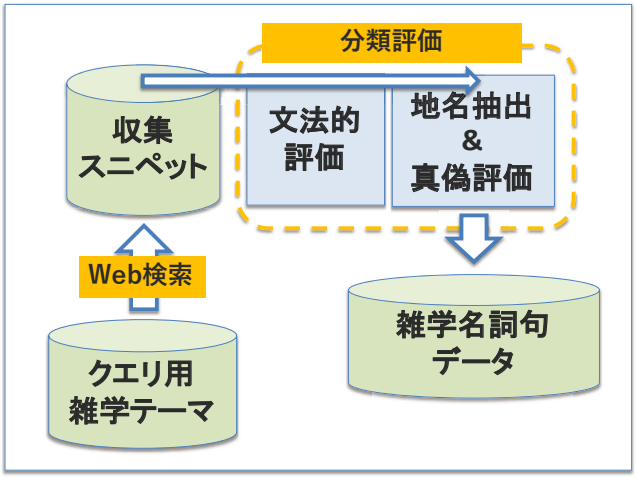

図 2 本稿における雑学文とその構成要素

表 1 クエリとして用いた雑学テーマ

雑学テーマ	雑学テーマを用いた雑学名詞句例
日本一	日本一かたい名刺
発祥の地	ナポリタン発祥の地
聖地	けいおんの聖地
最大級	世界最大級の水族館
舞台	もののけ姫の舞台
ロケ	朝ドラ・スカーレットのロケ地

3.2 名詞句抽出

本節では，雑学名詞句 p の候補となる名詞句の収集方法について述べる．

p を収集するにあたり、本研究ではウェブ検索エンジンを用いてデータの収集を行った．検索用のクエリには 47 の都道府県名と，35 の雑学テーマをそれぞれ一つずつ併記する形で用いた．例えば「秋田県 日本一」などのように，それぞれを空白を挟んで併記する．また後者に関しては表 1 で一例を示す．これらの 47 の都道府県名と 35 の雑学テー

表 2 p として扱う対象名詞句パターン

名詞句パターン	抽出した p の例
名詞-助詞-名詞	日本一の棚田
形容詞-名詞	美しい蕎麦
形容動詞-名詞	有名である仏像
名詞-助詞-動詞-名詞	キャベツを生産する農家

マのすべての組み合わせについて 1000 件ずつのスニペットを回収した.

p について，得られたスニペットからの収集をおこなう.p はスニペット内の文章の一部から作成され，その切り出し条件として形態素解析を行った際，表 2 の「名詞句パターン」の列に示すような品詞の構成であるものを回収した. これは竹本 [5] の分類した 4 単語以内の名詞句の抽出率上位のパターンを参考にした. またその中から p として候補を選定する際には，雑学テーマが名詞句中に含まれているもののみを候補とした. この操作によって，表 2 の「抽出した p の例」の列に示すような名詞句をこの時点での p として回収している. 以上のスニペットから p を回収する工程を改めて以下に記す.

(1) 47 の都道府県名と 35 の雑学テーマのすべての組み合わせをクエリとして（クエリ例：「秋田県 日本一」）検索結果のスニペットを各 1000 件ずつ回収する.
(2) p の候補として，表のに示すようなパターンに該当する文章の一部を抽出する.
(3) 上記で抽出したものから，抽出元スニペットの検索クエリとして用いた雑学テーマが含まれたものについて p として回収する.

これによって計 37207 件の p を回収した.

3.3 文法的評価

上記で得られた雑学名詞句 p から，$t(p,l)$ とした場合に文法的に不適当なものを取り除く操作を行う. この判断を行うにあたり，得られた p から無作為に 1000 件を取得し，$t(p,l)$ の概念的な連接を仮定した時に，その意味が推定できるか否かで正誤のラベリングを行った. これによって，326 件の正解データと 674 件の誤りデータが得られ，それぞれについて自然言語処理を行い品詞と固有表現ラベルの出現頻度について算出した. ここで得られた結果から，表 3，4 に示す品詞と固有表現ラベルが p 上の出現位置に含まれた場合に不適当として除去した. また表における出現頻度は出現位置として指定された場所におけるものを表す. この操作によって，上記の p から 19944 件を抽出した.

3.4 真偽評価

本節では雑学文 $trivia$ において p の主語となる l に関して詳細なエリアを特定するための地名抽出と，それを用いた p と l との関連の真偽判定を行った.

表 3 除去対象品詞の正誤ラベルにおける出現頻度と出現位置

品詞	出現頻度 正	出現頻度 誤	出現位置
名詞-普通名詞-サ変可能	0.0245	0.2329	文末
名詞-固有名詞-地名-一般	0.0061	0.0563	文末
名詞-数詞	0.0031	0.0237	文末
助詞-係助詞	0.0000	0.0051	文中
補助記号-句点	0.0000	0.0041	文中
名詞-普通名詞-サ変形状詞可能	0.0000	0.0019	文中

表 4 除去対象品詞の正誤ラベルにおける出現頻度と出現位置

固有表現ラベル	出現頻度 正	出現頻度 誤	出現位置
$Province$	0.0841	0.2151	文中
$OrdinalNumber$	0.0088	0.0646	文中
$NPerson$	0.0000	0.0087	文中
$NumexOther$	0.0000	0.0051	文中
$CorporationOther$	0.0000	0.0049	文中
$Percent$	0.0000	0.0019	文中

p はそれぞれがウェブ検索のクエリとして用いられた都道府県名を 1 つ l として有している. この l をより詳細な市町村などのエリアと紐づけるために，スニペットから l を抽出した. プロセスは以下の通りである.

(1) p の抽出に用いたスニペットから品詞が「名詞-固有名詞-地名-一般」である名詞の集合 $A = \{a_1, a_2 ... a_n\}$ を収集する.
(2) p を収集する際に使用した都道府県名 l の内包する地域名集合を L と定義するとき，$L \ni \langle a \rangle$ となる名詞 $\langle a \rangle$ を $true$ とラベリングする. またこれを満たさなかった $\langle a \rangle$ については $false$ であるとラベリングする
(3) $true$ と $false$ それぞれの $\langle a \rangle$ の集合を A_{true}, A_{false} と定義する. このとき $|A_{true}| > |A_{false}|$ および $A \neq \emptyset$ の条件を満たさなかった場合，当該スニペットから抽出された p を不適当として除去する.
(4) 上記条件を満たした場合は，l について，スニペット内で最も p に隣接していた $\langle a \rangle (\in A_{true})$ を l として都道府県名に付け加える.

この操作によって，上記の p から 8070 件を抽出し，これをテンプレートに挿入可能な最終的な p とした.

4. 実験と分析

提案手法によって抽出された雑学名詞句 p について，テンプレートに当てはめることを仮定した場合に雑学文 $trivia$ として成り立つかどうかを検証する. 検証には提案手法で抽出された p のうち 3 つの雑学テーマ（日本一，発祥の地，聖地）について各 100 件のデータを無作為に取得し，以下の定義に該当するかどうかで評価を行う.

(1) $trivia$ が概念的に $t(l,p)$ における 2 つのいずれかの関

表 5 手法毎の定義に該当する割合一覧

	割合		
雑学テーマ	**gc+tc**	**gc**	**plain**
日本一	**0.46**	0.34	0.14
発祥の地	**0.62**	0.40	0.27
聖地	**0.57**	0.35	0.28

表 6 実験に用いた p の例

評価	抽出した p の例	対応する l
成功	ハンドボールの聖地	富山県氷見市
	日本一のカレー県	鳥取県
	カラオケボックス発祥の地	岡山県岡山市
失敗	質ともに日本一	北海道
	全国 3000 箇所以上の聖地	福岡県
	発祥の地サイト	静岡県

係を満たし，その意味を理解することが容易である

(2) l,p について，p を抽出したスニペットのテキストの記載されたウェブページの内容を見ることで，l と p が適切に関連していることが容易に理解できる．

(3) p における雑学要素となりうる部分が具体的な事物を特定できている．

上記の 3 つの条件をいずれも満たすかそうでないかによって p が $trivia$ として成立するかを検証し，各雑学テーマについてその割合を算出する．また提案した 3 つの選別手法を用いた p の集合と比較するために，以下の提案手法の一部を用いた集合との比較を行った．

(1) **gc+tc**: すべての提案手法を用いて選別した p の集合

(2) **gc**: 文法的評価を用いて選別した p の集合

(3) **plain**: 名詞句抽出をしただけの p の集合

なおこれらの分類は筆者が行った．

各手法における 3 つのテーマの $trivia$ としての成立割合は表 5 の通りである．また提案手法によって得られた成功例・失敗例を表 6 に記す．

結果から，すべての提案手法を適用した p の集合が最も定義に該当する割合が多く，地域に関する雑学情報として成立していた．一方で雑学要素として扱われる対象が抽象的であったり，適切に固有名詞を抽出できていないものが多く確認された．とりわけ雑学テーマ「日本一」において，p が「なく日本一」や「日本一心」などのように文法的に破綻しているものが多く含まれた．また一部で p 自体は文法的に成立しているものの，元のウェブページを確認すると地域名 l との関連がないものも見られた．こうした不適格のパターンは今回実験に用いた雑学テーマごとに差異があることが考えられるため，今後それぞれにおける傾向をさらに検証する必要がある．

5. 今後の課題

本節では今回提案した手法を踏まえた上で現状の課題とさらなる発展手法についての提案を行う．また本稿においては得られた雑学名詞句 p について，実際にテンプレートに当てはめた上での雑学文 $trivia$ を生成しなかった．$t(l,p)$ の概念的な has-a,is-a 関係について分類を行えないことが要因であり，これについても今後の課題とする．また提案した p の各分類手法についてもそれぞれの有用性について十分な検証ができていないため，それについても今後検討する．

5.1 文法的評価

本稿では，文法ルールに基づき手動でラベリングした p の自然言語処理的分析に基づき，その正誤判定の基準を作成した．しかし **gc+tc** においても文法的に破綻した p が散見された．この課題について，大島の提案した両方向構文パターンを用いた関連語の取得法 [6] が有用ではないかと考える．この手法では，ある名詞についてそれの後方に文字列が続く構文パターンと前方の文字から接続される 2 つの構文パターンを用意する．そしてそれをクエリとしたウェブ検索によって両方のパターンに該当するものを抽出することによって，パターンに該当し，文法的に適切な名詞を得ることができる．この方法は本稿における雑学テーマの設定とも関連が深く，分類精度を向上させる上で有用であると考える．また当該手法ではこの両方向構文パターンについても逆説的に発見することができるため，雑学テーマの設定に応用することで品質の高いデータの収集量を増やすことが期待できる．

5.2 地名抽出と真偽判定

今回の真偽判定の方法において，その判定基準として都道府県名を格納した地域名 l と関係のある地域名集合 $Areas$ の個数のみを用いた．しかし地名抽出において p に隣接した $\langle a \rangle$ を用いたように，その位置関係を考慮に入れることも可能である．スニペットによっては，観光情報をまとめたブログなどから取得している場合には多くの地域名を含んでいることからも，$\langle a \rangle$ の p との位置関係を表す変数 $distance$ を定義するとき，

$$\frac{1}{n}\sum_{i=1}^{n} distance \cdot bool > x$$

を提案する．ここで n は $Areas$ の要素数を表し，$bool$ は $\langle a \rangle$ が $True$ か $False$ のいずれに属すかによって変化する重みを表す．この値の平均が閾値 x を超えた場合に，既存の l との関連を真とする．

5.3 雑学としての品質評価

今回の提案手法では，あくまで文法的な適切さと真偽の判定を行ったのみで，雑学としての面白さについての評価を行っていない．現状の仮説として，文法的破綻のない p

については，その収集元であるスニペット上に存在する感嘆や驚きを表す表現の個数によってその品質を評価できるのではないかと考える．今後は上記のような手法によって p の面白さをスコアリングすることを目指し，それによっての雑学の選別も検討する．

6. まとめ

本稿では，地域に関連する雑学文 *trivia* を生成するために，その主題となる雑学情報の収集方法について提案した．雑学テーマと地域名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雑学名詞句 p を抽出し，その品質を評価することによってテンプレートに当てはめた際に *trivia* として適切なものを抽出した．実験によって提案した評価手法の抽出能力について一定の可能性を示したが，文法的に破綻した p を十分に除去するには課題が残る．今後は文法的評価手法について異なるアプローチを試みるなどし，収集する p の品質を向上させ，*trivia* として自然言語的な出力を得ることを目指す．

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

参考文献

[1] Niina, K. and Shimada, K.: Estimating Trivia Score for Trivia Sentence Extraction, *Proceedings of IEICE Tech. Rep., vol. 118, no. 122, NLC2018-7, pp. 69-74.* (2018).
[2] Prakash, A., Chinnakotla, M. K., Patel, D. and Garg, P.: Did you know?- Mining Interesting Trivia for Entities from Wikipedia, *Proceedings of Twenty-Fourth International Joint Conference on Artificial Intelligence,* (2015).
[3] Doi, S., Wang, Y., Zhao, C. and Utsuro, T.: Design of a trivia game for traveling and domestic enjoyment in Japan, *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication* (2017).
[4] Korn, F., Wang, X., Wu, Y. and Yu, C.: Automatically Generating Interesting Facts from Wikipedia Tables, *Proceedings of the 2019 International Conference on Management of Data* (2019).
[5] 竹本　翔，徳久雅人，村田真樹：情緒推定のための名詞句の評価極性，第 76 回全国大会講演論文集 (2014).
[6] 大島裕明，田中克己：正解語ペア漸増による関連語取得のための両方向構文パターン発見，第 1 回データ工学と情報マネジメントに関するフォーラム (2009).