固有表現分析による知識グラフの自動作成

武 賢倖† Thatsanee Charoenporn†† Virach Sornlertlamvanich†††

アジア AI 研究所，武蔵野大学 〒135-8181 東京都江東区有明 3-3-3

E-mail: †s2122035@stu.musashino-u.ac.jp, {††thatsane, †††virach}@musashino-u.ac.jp

あらまし インターネットに発信されている様々な情報の中から、重要な情報についてのキーワードを取り出し、知識グラフの作成を主に考える。具体的な方法としては、注目されている文章から固有表現を抽出し関連性ごとに分類してネットワーク化することである。それによって一貫した情報をユーザーに提供する。その結果、検索する情報が網羅的に辿り着き、より簡潔的な情報が得ることが出来る。本研究では、観光スポットを注目して観光情報を知識グラフにし、可視化することによって効率よくアクセスできるようにする。客観的に書いた説明文を（Wikipedia や公式サイトなどから引用）分析して、関連する事柄、人物、場所、組織の 4 種類のジャンルの情報を抽出し、関係分析の結果によって知識グラフの作成方法を提案する。

キーワード 固有表現，知識グラフ，キーワード，分類

1. はじめに

技術の発達の進歩によって、インターネット上にさまざまな情報が散らばり、携帯電話を 1 人 1 台持つのが当たり前と言っても過言ではない便利な世の中になった。その結果、人々の日々の情報源は新聞、週刊誌といった紙媒体からネットニュース、Web サイトといった電子媒体に移り変わっていった。そして、人々は以前よりも簡単に自分が欲しいと思う情報を入手できるようになった。現在必要とされている技術としては、人々が欲しい情報と関連した新たな情報を提供できる技術である。

関連した情報を見つける際に重要なこととしては、情報を形成する説明文である。説明文を分析して固有表現を抽出し、その固有表現を関連する情報として新たな分析を行うというのが、関連した情報の収集の基本である。

本研究では、観光地の説明文を収集して、それを分析していき、その観光地に関連する新たな情報を収集することによる知識グラフの作成を考察した。

2. 説明文の収集、分析

観光地のことを書いた説明文は公式サイト、Wikipedia、個人のブログなど様々なものがあるが、その中からなるべく簡潔で断定的である常体（だ、である調）を用いて書かれた文章を選ぶ。また、固有表現を多く含んだ文章であるかということにも気をつける。

1. 常体の文章の収集

本研究では、築地本願寺の説明文を中心にして知識グラフを作成することにした。以下は築地本願寺についての説明文である。（Wikipedia から引用）

> 築地本願寺は江戸時代の 1617 年に、西本願寺の別院として浅草御門南の横山町（現在の日本橋横山町、東日本橋）に建立。「江戸海岸御坊」「浜町御坊」と呼ばれていた。しかし明暦の大火（振袖火事）により本堂を焼失。その後、江戸幕府による区画整理のため旧地への再建が許されず、その代替地として八丁堀沖の海上が下付された。そこで佃島（現：中央区佃）の門徒が中心となり、本堂再建のために海を埋め立てて土地を築き（この埋め立て工事が地名築地の由来）、1679 年に再建。「築地御坊」と呼ばれるようになった。なお、このときの本堂は西南（現在の築地市場）を向いて建てられ、場外市場のあたりが門前町となっていた。
> 現在の本堂は 1934 年の竣工。古代インド様式をモチーフとしたこの建物は、当時の浄土真宗本願寺派法主・大谷光瑞と親交のあった東京帝国大学工学部名誉教授・伊東忠太による設計である。当時の宗教施設としては珍しい鉄筋コンクリート造で、松井組（現：松井建設）の施工により建築された。大理石彫刻がふんだんに用いられ、そのスタイルは現在においても斬新かつ荘厳で、築地の街の代表的な顔である。本堂

は重要文化財に指定されている。 浄土真宗本願寺派の新体制移行（2012 年 4 月 1 日付）に伴い、正式名が従前の「本願寺築地別院」から「築地本願寺」になった。これにより、築地本願寺は全国唯一の直轄寺院となる。 著名な人物の葬儀が宗派を問わず多く執り行われている。 2015 年に銀行員・経営コンサルタント出身の安永雄玄が宗務長に就任し、他宗派信徒や訪日外国人などにも開かれた寺を目指す「寺と」プロジェクトが進められている。 電話で相談を受け付けるコールセンターなどに加えて、2017 年 11 月 8 日にはカフェなどが入るインフォメーションセンターと合同墓を開設した。

図 1 築地本願寺の説明文

説明文の収集に当たっては、以下の事柄に留意した。

① 正確な情報に基づいて作成された文章である。
② 同じ単語が必要以上に繰り返し使用されていない文章である。

2. 分析、固有表現抽出

説明文の中から固有表現を抽出し、その中から関連する事柄、人物、場所、組織の 4 種類の固有表現を選び出す。今回の説明文では以下の結果となった。

事柄	江戸時代 関東大震災
人物	大谷光瑞
場所	西本願寺 浅草御門南 日本橋横山町 江戸海岸 浜町 八丁堀 築地御坊 築地市場
組織	江戸幕府 浄土真宗 東京帝国大学工学部

図 2 抽出した固有表現

分析の手順としては以下の通りである。

① 文章の中から、分析のために使うことができると思われる部分を取り出す。
② Python を用いて固有表現を抽出する。フレームワークには当初は spaCy を用いた[2]。
③ 抽出した固有表現を図 2 のような表に表示する。

3. 知識グラフの作成

第 2 節で説明文から抽出した 4 種類のジャンルの説明文から、新しく説明文を入手し、同じ要領で分析を行った。

中央に今回の元となったキーワードである「築地本願寺」を配置し、周りに関連する 4 種類のジャンルのキーワードを配置した。

図 3 作成した知識グラフ

4. 知識グラフの修正

完成した知識グラフには、以下の問題点があった。

1. 関連するものが見つけられないキーワードがある。
2. 関連するものの中にほとんど関係ないと思われるキーワードが混ざっている。

今回の知識グラフで、上記の問題点があったキーワードは以下の通りである。

キーワード	類型
江戸時代 関東大震災	1. 関連するものが見つけられないキーワードがある。
大谷光瑞 浅草御門南 江戸海岸 築地御坊	
西本願寺 八丁堀 築地市場 江戸幕府 東京帝国大学工学部	2. 関連するものの中にほとんど関係ないと思われるキーワードが混ざっている。

図 4 問題点のあるキーワード一覧

この問題を解決するために、以下の解決策を行った。

2.1. 固有表現抽出の再試行

本研究では固有表現抽出の際、前述の通り spaCy を用いたが、spaCy ではそこまで関係のない固有表現を抽出したり、重要な語句が途中で途切れて認識されたりするなど正確な固有表現抽出ができないことが判明したため、自然言語処理ライブラリ（NLP ライブラリ）の GiNZA を利用して、新たに固有表現抽出を行った[2][3]。

新たに固有表現抽出を行った結果を以下に示す。なお、地名、組織名などで当時と現在とで名称が変わっているもの（例：松井組→松井建設、佃島→中央区佃など）に関しては、今後のデータ収集の合理化、簡素化を目的として現在の名称のみを収集した。

事柄	明暦の大火 関東大震災
人物	大谷光瑞 伊東忠太 安永雄玄
場所	日本橋横山町 東日本橋
	中央区佃 築地市場 築地
組織	西本願寺 江戸幕府 浄土真宗本願寺 東京帝国大学 松井建設

図 5 GiNZA を使用して抽出した固有表現

2.2. 単語の重要性の測定による一部の固有表現の削除

固有表現抽出の再試行で、新たに抽出された 15 の単語のうち、重要性が薄い情報を削除することにした。

まず、図 1 の説明文の中から、重要度の高い単語をピックアップし、その中から図 5 の表にある固有表現を残すことにした。

まず、図 1 の説明文を利用してワードクラウドを作成し、出現頻度の高い固有表現を可視化することにした[4]。

図 6 作成したワードクラウド

ワードクラウドでは抽出した固有表現以外の単語が、出現頻度が多いように表示されたため、固有表現を整理することはできなかった。

そのため、本研究では TF-IDF 値を利用した重要度の測定を考えた。手順は以下の通りである。

1. 図 1 の説明文の他に、もう一つ別の説明文（例えば公式サイトに表示されている紹介文）を入手し、2 つの説明文に形態素解析を行う。
2. 2 つの文章を形態素解析して完成した 2 つの文書を利用して、図 5 に記載した固有表現の出現頻度（TF）と逆文書頻度（IDF）を求める[5]。
3. 出現頻度と逆文書頻度を掛け合わせて TF-IDF 値

を算出し、その値が高い固有表現だけを残して作業を行う[6]。

今回は時間の都合で TF -IDF 値を求めることができなかったので、固有表現抽出の再試行のみで知識グラフを修正した。

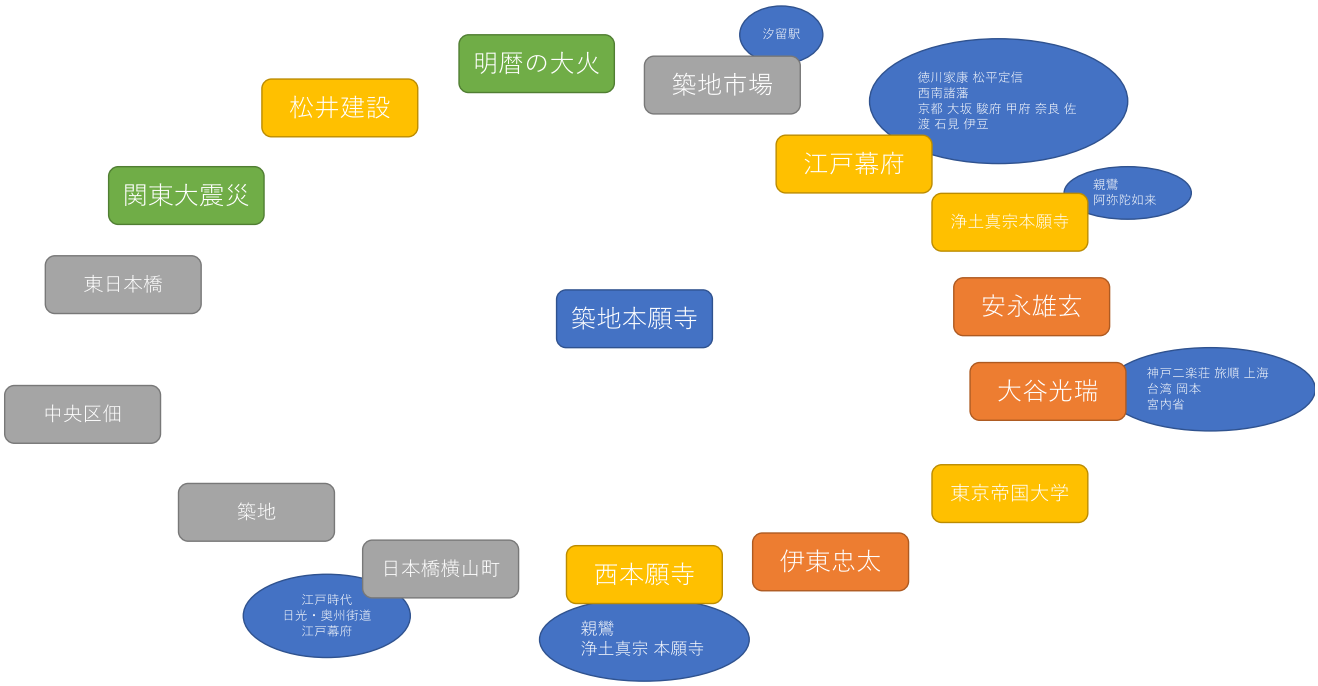

図 8 修正した知識グラフ

5. ウェブスクレイピングによる説明文の再収集

説明文に関して、冒頭で 2 つの条件に合う説明文を手動で抽出したが、今回はウェブスクレイピングによる説明文の収集を考察した。築地本願寺の公式サイトから、説明文を収集して新たな知識グラフの作成を行い、さらに簡略化された知識グラフが完成した。しかし使用した文章が異なるため、ウェブスクレイピングの使用と知識グラフの簡略化との関係性は不明である。本文抽出には Python3 の extractcontent3 のモジュールを使用した[7]。

1617 年に浅草近くに創建されましたが、1657 年の「明暦の大火」とよばれる大火事で焼失してしまいます。その後、再建のため江戸幕府から与えられた土地が現在の場所ですが、当時は海上でした。そこで海を埋め立てて土地を築き本堂を建立したことが「築地」という名称の由来になっています。また 1923 年には関東大震災に伴う火災により再度本堂を焼失しましたが、1934 年に再建し現在の本堂の姿となりました。

現在の本堂は、東京帝国大学(現在の東京大学)名誉教授で建築史家の伊東忠太博士の設計によるものですが、建築研究のためアジア各国を旅した博士と、時を同じく、仏教伝来ルートを明らかにするために探検隊を結成し、シルクロードを旅した大谷光瑞(当時の浄土真宗本願寺派門主)との出会いが縁となっています。
築地本願寺の建物は、インド等アジアの古代仏教建築を模した外観や本堂入り口のステンドグラス、数多くの動物の彫刻などが特徴で、オリエンタルな雰囲気は、まさにシルクロードを伝わる仏教伝来のルーツを感じさせます。その一方で、内観においては僧侶のお勤めスペースよりも本堂内の参拝スペースの方が広く、中央正面に本尊阿弥陀如来が安置しているなど、伝統的な真宗寺院の造りとなっております。2014(平成 26)年には本堂及び大谷石の石塀と三門門柱が国の重要文化財に指定されました。
シルクロードを旅してきた伊東忠太博士でなければ作り上げることのできない、独特な仏教寺院をぜひお楽しみください。

図 9 ウェブスクレイピングによって収集した説明文

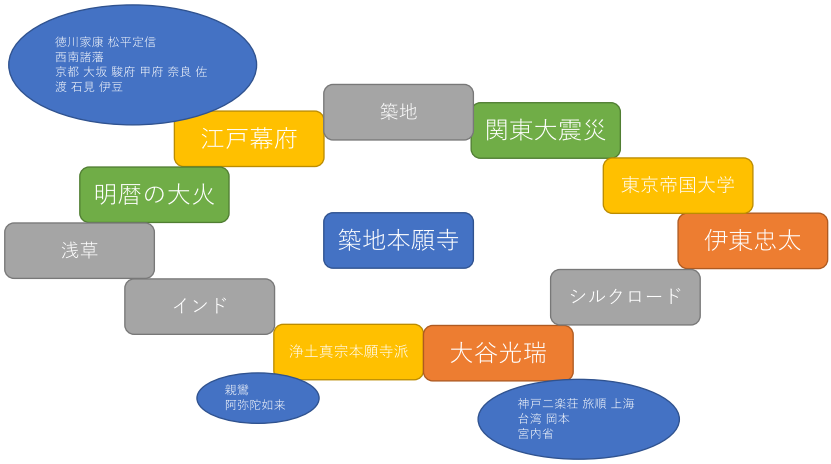

図 10 ウェブスクレイピングによって収集した説明文を使用して作った知識グラフ

6. 期待される活用方法

知識グラフの自動生成について、期待できる活用方法としては、本研究のテーマに沿ってもいる観光地の情報の収集がある。

例えば旅行の際、どの様な観光地に行けばいいのか迷うといった際に、自分の興味のある事柄を中心にして知識グラフを作成して、自分の好きなことに関連する行き先を探すことができる。

他にも、英語や国語の学習をする際に、資格試験の過去問などの長文の中で頻出する単語や句をリストアップすることが出来れば、実際の試験を受ける時の対策に大いに役立つと思われる。

7. おわりに

本研究では、中心となるキーワードを決定し、それについての説明文から固有表現抽出をおこなった。それによって抽出された固有表現から知識グラフを作成したが、これは固有表現の抽出を自動で、知識グラフの作成を手動で行なった半自動的な作成方法であった。そのため完成した知識グラフは範囲が狭いものになったので、実用化するには完全に自動的に作成し、知識グラフの範囲を広げることが必要であると考えられる。

参 考 文 献

[1] もふもふ技術部『spaCy + GiNZA を使って固有表現抽出とカスタムモデルの学習をしてみる』https://tech.mof-mof.co.jp/blog/spacy-ner/（令和 4 年 1 月 11 日）

[2] Qiita『自然言語処理ライブラリ GiNZA で固有表現抽出してみた』https://qiita.com/yuto16/items/1fc1f2b79195a503c681（令和 4 年 1 月 11 日）

[3] DenDenBlog『Colab で GiNZA v5 を試してみた！【固有表現抽出】』https://dendenblog.xyz/ginza-v5/#Colab%E3%82%92%E4%BD%BF%E3%81%A3%E3%81%A6GiNZA_v5%E3%82%92%E5%8B%95%E3%81%8B%E3%81%99%E6%96%B9%E6%B3%95（令和 4 年 1 月 11 日）

[4] GMO AD Partners TECH BLOG by GMO『日本語テキストをワードクラウドで可視化する』https://techblog.gmo-ap.jp/2021/06/15/text-visualization-wordcloud/（令和 4 年 1 月 11 日）

[5] Shingo の数学ノート『Python で文章の近さを計算しよう 1(形態素解析)』http://mathshingo.chillout.jp/blog12.html（令和 4 年 1 月 11 日）

[6] ギークなエンジニアを目指す男『【初学者向け】TFIDF について簡単にまとめてみた』https://www.takapy.work/entry/2019/01/14/141423z（令和 4 年 1 月 11 日）

[7] やってみる『HTML から本文テキストだけを抽出したい（python-extractcontent）』https://ytyaru.hatenablog.com/entry/2021/10/03/000000（令和 4 年 2 月 11 日）