

オープンデータの活用に関する研究

久永忠範^{†1} 能登大輔^{†2} 郭崇^{†1} 瀧田孝康^{†1}

概要: 近年、ビッグデータやオープンデータの活用が推進され、国や地方自治体をはじめ多くの団体がオープンデータの公開、活用に取り組んでいる。これらの開示されたデータ形式は、ワード形式、エクセル形式や CSV 形式のファイルがまだまだ多く、2012 年に策定された「電子行政オープンデータ戦略」に明示されている「機械判読可能で人手を多くかけずにデータの 2 次利用が可能である」というデータ活用までには至っていないのが現状である。開示されている多くのデータを RDF 形式へ簡単に変換できれば、複数のオープンデータの機械的な連携が可能となりオープンデータの活用を促進することが可能となる。本研究では、RDF 形式の述語に当たる語彙に注目し、語彙の共通化を行うために Wrod2Vec というニューラルネットワークを活用して、語彙の近似性について実験を行った。また IPA の推奨する共通語彙基盤の語彙等を活用して、メジャーな名前空間の活用と語彙の共通化を行う手法について提案を行う。

キーワード: オープンデータ, Word2Vec, RDF, コア語彙

The study on utilization of Open data

TADANORI HISANAGA^{†1} DAISUKE NOTO^{†2} CHONG GUO^{†1}
TAKAYASU FUCHIDA^{†1}

Abstract: In recent years, Utilization of big data and open data has been promoted, and many organizations including the national government, local government and other organizations are working on opening and utilizing open data. These disclosed data formats are still many files in word format, Excel format and CSV format. The data utilization that "It is machine-readable and second-hand use of data is possible without spending a lot of manpower" clearly stated in the 2012 e-administration open data strategy formulated in 2012 is not yet reached. If many disclosed data can be easily converted to RDF format, mechanical linkage of multiple open data becomes possible and it becomes possible to promote utilization of open data. In this research, we focused on the vocabulary that corresponds to the predicate of the RDF form, and experimented about the approximation of the vocabulary by utilizing the neural network called Wrod2Vec in order to share the vocabulary. We also propose a method to utilize major namespace and common vocabulary by utilizing IPA recommended vocabulary based on common vocabulary.

Keywords: Open data Word2Vec RDF kuromoji

1. はじめに

近年、ビッグデータやオープンデータの活用が推進され、国や地方自治体をはじめ多くの団体がオープンデータの公開、活用に取り組んでいる。

これらの開示されたデータ形式は、特定のアプリケーションに依存した形式や CSV 形式のファイルがまだまだ多く、2012 年に策定された「電子行政オープンデータ戦略」に明示されている「公共データは機械判読可能で人手を多くかけずにデータの二次利用が可能である」というデータ活用までには至っていないのが現状である。

開示されている多くのデータを RDF 形式[1]へ簡単に変換できれば、複数のオープンデータの機械的な連携が可能となりオープンデータの活用を促進することができるだけでなく、人間の考えでは連携することが難しかった異なるデータ間をつなぐ、新しい情報価値を創出することが可能となると考えられる。

本研究では RDF 形式の述語となる語彙を共通化することに目的として、各データの語彙同士の近似性を測るために word2vec[2]と呼ばれるニューラルネットワークを応用して、単語をベクトル空間に対応付ける実験を行った。

Word2vec とは、自然言語の文章を入力として学習を行うことで、単語をベクトル空間に写像し、単語の意味に基づいたベクトル演算を可能にする手法であり、近年さまざまな分野で応用されている。

すでに公開されているオープンデータのタイトル行に現れる文字列(項目名)をキーワードとして Google 検索によりページを検索し、上位に現れたページ中のソースコードから文章を抽出し、それらを Kuromoji[3]と呼ばれる構文解析機で解析して分かち書きし Word2Vec に入力することで、単語を名前空間に写像させた。収集した約 2900 万の単語を学習した結果、人間が判断しても近い単語が上位に現れることが確認できた。現在は、その結果を使って複数のオープンデータに現れる述語の共通化を試みる。

^{†1} 鹿児島大学院理工学研究科
Graduate School of Science and Engineering Kagoshima University
^{†2} 京都大学
Kyoto University

^{†3} 鹿児島大学工学部
Faculty of Engineering Kagoshima University

また Word2Vec の写像するベクトル空間の次元数を変えることで、得られる単語の語彙の軸が変化するので、これらの軸を利用することで単語のカテゴリライズが可能になると考えられる。

以上の手法を統合することで、オープンデータの持つ項目の語彙を共通化し、簡単にデータの RDF 形式への変換が行えるようになる手法の提案を行う。

2. 現在のオープンデータ

実際、現在のオープンデータの現状はどのようになっているかを地方自治体、政府、そしてオープンデータの活用促進を行う取り組みについて調査した。

2.1 地方自治体におけるオープンデータの現状

地方自治体は膨大な公共データを保有しているが、それがオープンデータとして公開され、各々のデータの連携した活用に至っていないのが現状である。そこで全国の 47 都道府県所在地のオープンデータの開示状況を調査した。その中でオープンデータサイトを構築しているのは 27 都道府県であった。(平成 28 年 9 月現在)

サイトで開示されているオープンデータのデータ形式をみると、html、PDF、xls 形式のデータが多く、機械判読に適していない状況にある。またオープンデータへの取り組みが各サイトに記載されているものの、データの活用方法や今後の方向性がしっかりと明示されていないものが多いのが現状である。

2.2 政府におけるオープンデータの現状

日本政府は、機械判読に適したデータ形式のデータを、営利目的も含めた二次利用が可能な利用ルールで広く公開する「オープンデータ」の取り組みを推進することとした。また 2015 年 10 月には、データの横断的検索等の機能を備えたオープンデータのデータカタログサイト Data. or. jp[4] (図 1) を開設した。このサイトでは、政府に関係する各省庁が保有するデータが 2016 年 12 月現在、18, 000 のデータセットとして開示されている。

特に開示データの多い省庁は、行財政、国土交通省、経済産業省の順になっている。ここで開示されているデータ形式も地方自治体と同じように、PDF、html、xls、csv の順に多く、機械判読に適していないものが多い。

より多くのデータを登録するためにこのような状況になっているが、しかし当サイトでは、開発者向け情報として RDF 形式でメタデータ取得の流れが揭示されている。



図 1 データカタログサイト DATA.GO.JP

Figure 1 Data Catalog Site DATA.GO.JP

2.3 共通語彙基盤について

「世界最先端 IT 国家創造宣言」[5] (2013 年 6 月 14 日) が内閣で承認され、中核である「公共データの民間開放」と「利便性の高い電子行政サービスの提供」を支える基盤となる「共通語彙基盤」の構築を行うプロジェクトが開始された。このプロジェクトを担うのが独立行政法人情報処理推進機構[6]であり、共通語彙基盤を支える「情報連携用語彙データベース」の構築及びそれと連携し、データ構築などに活用するためのツール類の整備を推進している。

情報連携用語彙データベースでは、行政で電子的に交換・公開されている情報に用いられる用語の意味の違い等が起きないようにすることを目的とし、用語の意味や使い方の規則、電子的な表記法等、多様な情報が提供されてきている。

この共通語彙基盤で扱われている語彙、特にコア語彙をベースとして、本研究の語彙の近似性の尺度を考えるものとする。

3. 提案手法の流れ

本研究では、Tim・Berners=Lee[7]が提唱したオープンデータの 5 つの段階のデータ形式で、河海判読可能である 4 段階の RDF 形式に焦点を当てて、その中の述語の共通化を図ることにした。その述語にあたる語彙の共通化を図るためにトマス・ミコロフ氏の開発 Word2Vec を活用することにした。

Word2Vec は、単語をベクトル化して表現する定量化手法である。日本人が日常的に使う語彙数は数万から数十万と

いわれるが、この単語を 100 から 300 次元くらいの空間においてベクトルとして表現する。このことにより今まで精度を向上するのに難しかった単語同士の類似度や、単語間での加算・減算などができるようになり、単語の「意味」をとらえることができるようになった。これを活用することにより RDF 形式における述語の語彙の近似性を測り、できるだけ共通の語彙を用いられるような語彙の抽出を行うことにした。

3.1 RDF 形式について

オープンデータのデータ形式は、Tim・Berners=Lee の提唱した 5 つの段階（図 2）に分けることができる。1 段階から 3 段階は、主に人が見て判別する形式であり、4 段階から 5 段階は、機械判読可能な形式で表されている。その 4 段階で活用される形式が RDF である。RDF は主語、述語、目的語の 3 要素から成り立っており、この述語が主語と述語の関係性を示している。この述語の語彙を共通化することにより、いろいろなデータとの連携が可能になり、新たなより多くの情報の価値を作り出すことができる。



図 2 5つのデータ形式

Figure 2 Five Data Formats

3.2 コア語彙について

I P Aを中心に共通語彙基盤整備事業が開始され、語彙の共通化を図るため、コア語彙、DMD、I E Pの3つを柱とするプロジェクトが進められている。とくにコア語彙は、R D Fの述語の共通化を図る中核的な用語の集合であり、ここで示されているコア語彙の用語とオープンデータの項目を共通化することにより、機械判読におけるデータ連携を促進することができる。

3.3 単語学習の流れ

Word2Vec で学習するための文章を抽出するために、コア語彙で定義されている単語と鹿児島市のオープンデータサイトで掲示されている項目の単語をもとに Google 検索をおこないそれに関連する文章を抽出した。また日本語 Wikipedia アーカイブの文章と組み合わせて、単語要素に

分ける「kuromoji」を活用して総単語数 157, 339, 847, 単語の種類 269, 339 を抽出した。

3.4 Word2Vec での学習

Word2Vec とは、単語の意味や文法を考えるために単語をベクトル表現化して次元を圧縮したものである。この Word2Vec で学習した結果で単語の近似性を測る実験をおこなった。

この学習トレーニング手法には、CBoW と Skip-gram という 2 つの手法があり、「kuromoji」で単語要素に分けた文章を各々100 次元、200 次元、300 次元の学習を行った。

CBoW は、文脈中の単語から対象単語が現れる条件付き確立を最大化し、前後の単語から対象単語を推測する手法である。Skip-gram は、出力層における周辺単語予測のエラー率の合計を最小化することで、指定された単語に対してどのような単語が当てはまるかを予測する手法である。

3.5 Word2Vec での実験

仮定を確認するために、いくつかのコンピュータシミュレーション実験を行った。この実験では、鹿児島市の公開データから施設から教育施設（特別支援学校）のデータを抽出して実験を行った。

実験の流れは、次のとおりである。

1. 各都市のウェブサイトからオープンデータを取得する。
2. 施設名を表す指定列データを抽出する。
3. 各列の単語をスペースで区切る。
4. 単語ごとに単語辞書から単語ベクトルを求め、来られるベクトルをすべて追加する
5. コア語彙及びすべての語彙から最も近い 10 単語を検索する。

この実験では、6 種類の単語辞書を使用した。2 種類の学習アルゴリズム（CBoW と Skip-gram）[8]ごとに単語ベクトル空間の 100 次元、200 次元、300 次元の実験をおこなった。

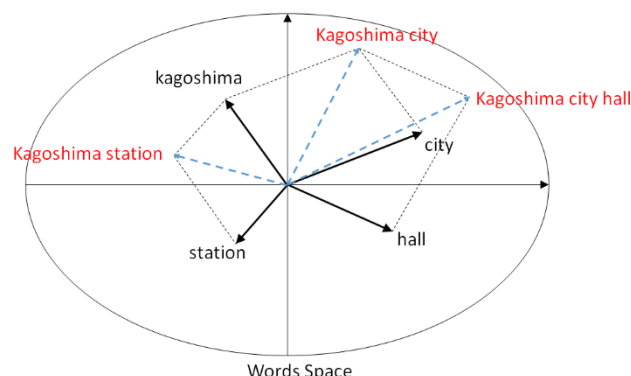


図 3 語彙の合成ベクトル

Figure 3 Composite vector of Vocabulary

任意の複合語は、単語ベクトル空間内に追加されたベクトル

ルで表現できると考える。例えば、「鹿児島」と「駅」の和で「鹿児島駅」や「鹿児島」と「市」の和で「鹿児島市」で表現できる。また「鹿児島市」と「ホール」で「鹿児島市民文化ホール」と語彙のベクトル和をとることによっていろいろな表現を作成することが可能である。前頁の図3に単語ベクトルの計算イメージを示す。

表1(a), (b)は、鹿児島市のオープンデータページから施設情報の教育施設の特別支援学校の csv データの名称列(図4の赤字)のデータ図4を抜粋してそれらを kuromoji を活用して分かち書きし、それらの単語を Word2Vec で学習させ、その結果上位から並べたものである。

大分類,小分類,名称,ふりがな,住所,電話番号,ファックス,URL,X座標,教育施設,特別支援学校,鹿児島盲学校,,西谷山1丁目3-3,099-263-666
教育施設,特別支援学校,桜丘養護学校,,桜ヶ丘6丁目12,099-265-6642
教育施設,特別支援学校,武岡台養護学校,,小野町2760,099-282-0440,
教育施設,特別支援学校,皆与志養護学校,,皆与志町1782-1,099-238-5
教育施設,特別支援学校,鹿児島大学教育学部附属特別支援学校,,下伊
教育施設,特別支援学校,鹿児島養護学校,,吉野1丁目42-1,099-243-01
教育施設,特別支援学校,鹿児島聾学校,,下伊敷1丁目52-27,099-228-2

図4 鹿児島市オープンデータ 教育施設(特別支援学校)

Figure 4 Kagoshima-city Open Data Education facility(Special Support School)

"--- for Core ---"は、コアの語彙から"---for all---"は、以前学習した約27万種類の単語(これから全語彙という)から分かち書きした名称のそれぞれのベクトル和の語彙に近い値の単語を上位からそれぞれ並べたものである。

表1(a), 表1(b)の見方は、CBoW, Skip-gram 共に100次元, 200次元, 300次元のベクトル空間で、各次元ごとに左から順位, 単語, コサイン距離となっている。

CBoWのコア語彙における100次元, 200次元, 300次元の1位は、同じ「施設」が選択され、上位10位までの単語は、多少前後するもののそれほど差異がなく、同じような単語を抽出している。

また特色として、抽出された全語彙からの単語も一般的な単語が占めている。但し、コア語彙と全語彙の抽出された単語を比較すると一致するものがない。オープンデータで揭示されるデータのベクトル和と抽出された語彙のコサイン距離をみるとコア語彙の値より全語彙の値が大きく、語彙の近似性が見られる。これは、メジャーな名前空間の語彙でもデータで表示される単語とのコサイン距離は、だいぶかけ離れているのがわかる。

Skip-gramのコア語彙における100次元, 200次元, の1位は、「設置者」, 300次元は「担当者名」となっている。それぞれの10位までの単語を比較すると3つの次元に共通な単語は4つしかなく、次元によって抽出される単語にば

らつきがあることがわかる。全語彙においては、上位に抽出される単語は、順位は違うが同じような単語が抽出されているが、下位になると地名や学校名など固有名詞が抽出されている。CBoWと同様に抽出された単語のコサイン距離は、コア語彙よりも全語彙の方が大きく、コア語彙のようなメジャーな名前空間に登録されている単語とデータの単語のベクトル和は、だいぶかけ離れている傾向にあることがわかる。

教育施設(特別支援学校)											
cbow-100				cbow-200				cbow-300			
--- for Core ---				--- for Core ---				--- for Core ---			
1施設	0.423162	1施設	0.367637	1施設	0.343528	2法人	0.359411	2区	0.349766	2区	0.328814
3区	0.357639	3法人	0.318465	3法人	0.300566	4関連施設	0.299427	4設置者	0.290957	4設置者	0.27101
5施設関連	0.299427	5関連施設	0.272064	5関連施設	0.262328	6設置者	0.299099	6施設関連	0.272064	6施設関連	0.262328
7活動	0.250364	7設立日	0.250798	7設立日	0.256028	8設置日	0.250258	8設置日	0.249244	8事業年度	0.250201
9事業年度	0.248341	9活動	0.24276	9設置日	0.244288	10設立日	0.243779	10事業年度	0.234775	10管理者	0.231793
--- for All ---				--- for All ---				--- for All ---			
1学校	0.88494	1学校	0.865475	1学校	0.86202	2養護	0.732174	2養護	0.67753	2養護	0.662967
3中学校	0.702103	3中学校	0.657467	3教育	0.628492	4教育	0.695108	4教育	0.639034	4中学校	0.624465
5小・中学校	0.690864	5小・中学校	0.615993	5校	0.587145	6短期大学	0.680631	6校	0.613343	6小学校	0.582875
7盲学校	0.676451	7幼稚園	0.607639	7幼稚園	0.58138	8幼稚園	0.674764	8小学校	0.601674	8小・中学校	0.570489
9校	0.664794	9短期大学	0.592289	9公立	0.553341	10小中学校	0.656468	10公立	0.584	10学院	0.543373

表1(a) CBoW 教育施設(特別支援学校)

Table 1(a) CBoW Education Facility (Special Support School)

教育施設(特別支援学校)											
skipg-100				skipg-200				skipg-300			
--- for Core ---				--- for Core ---				--- for Core ---			
1設置者	0.607099	1設置者	0.49454	1担当者名	0.459572	2施設	0.592706	2設置者	0.452466	3設置日	0.579089
3設置日	0.579089	3設立日	0.48058	3関係者	0.445517	4法人	0.565163	4担当者名	0.478723	4設置日	0.445339
5対象者	0.54922	5利用者	0.472055	5対象者	0.441637	6関連施設	0.540788	6事業年度	0.466217	6事業年度	0.441278
7施設関連	0.540788	7年	0.463541	7施設	0.437573	8建築面積	0.540685	8関係者	0.45971	8設立日	0.427645
9関係者	0.535848	9敷地面積	0.457293	9担当者役職	0.425937	10設置位置	0.532763	10施設	0.456743	10駐車場種別	0.424096
--- for All ---				--- for All ---				--- for All ---			
1学校	0.896645	1学校	0.865928	1学校	0.852272	2高等	0.827913	2養護	0.804265	2養護	0.789891
3養護	0.826171	3高等	0.773952	3高等	0.760493	4幼稚園	0.819672	4盲学校	0.754616	4盲学校	0.703006
5中学校	0.811766	5中学校	0.75097	5中学校	0.702933	6盲学校	0.799812	6幼稚園	0.732936	6幼稚園	0.683023
7教育	0.790643	7教育	0.707925	7五ヶ別府	0.679682	8小・中学校	0.790133	8公立	0.70543	8教育	0.670602
9短期大学	0.7782	9生徒	0.694759	9立教女学院	0.658033	10私立	0.773525	10五ヶ別府	0.693791	10千葉敬愛	0.652872

表1(a) Skip-gram 教育施設(特別支援学校)

Table 1(a) Skip-gram Education Facility (Special Support School)

CBoWとSkip-gramで抽出された単語を比較するとコア語彙においては、CBoWの方がオープンデータの項目に近

い単語を抽出しており、Skip-gram は、一般的な単語が多いが、固有名詞や独特の単語が抽出されている。

上記の実験結果から CBoW と Skip-gram を比較すると CBoW で抽出された単語の方が、よりオープンデータの項目に近い単語を抽出することができる。

以上は、鹿児島市ホームページに掲載されているオープンデータの教育施設（特別支援学校）のデータをもとに実験をおこなったが、他にも施設情報の官公庁・公共施設、教育施設、地域福祉施設、地域包括支援センター、子育て支援施設、高齢者福祉施設、障害者福祉施設等、20のデータも同様に教育施設（特別支援学校）と同じような傾向の結果になった。

今回は鹿児島市の施設情報に限定してこのような実験を行ったが、同様な実験を全国のオープンデータサイトを活用して、施設情報だけでなく、あらゆる分野のデータを活用して、データ項目に対してどのような単語が抽出されるかを検証してみる必要があると思う。

そして現在は、メジャーな名前空間はコア語彙のみを活用しているが、いろいろなメジャーな名前空間の語彙も取り込んでなるべく項目に近い共通語彙を抽出できるようにしたいと思う。

4. 結論

本研究は、RDF 形式の述語を共通化することによってオープンデータの連携を図ることを目的にしている。今回の実験で、実存するオープンデータを活用して述語の共通化を図るため Word2Vec というニューラルネットワークの仕組みを活用したが、IPA の提示するコア語彙に一致するものが少なかった。今後も述語の共通化を図るための方策を探るために、実際の RDF の述語に当たる語彙を、人的な視点でどのようなものを抽出するかアンケート調査を行う予定である。各自治体のオープンデータを参考にいろいろなデータを抽出しそのデータに対して、Word2Vec で抽出された上位 10 個の単語のうちそのデータの項目としてどれが一番適しているかのアンケートを行い、メジャーな名前空間に属する単語との比較を行い語彙の近似性を測りたいと思う。それをもとにオープンデータを簡易に RDF 化できるレコメンド機能などを提供していきたい。

図 5 は、RDF 生成の流れと活用である。オープンデータを RDF 化し、その RDF 化されたデータの連携と今後の新たな情報の創出を示した図である。今後のオープンデータ活用の促進を図ることを念頭にこれらの流れが円滑に行うことができるように研究を行いたい。

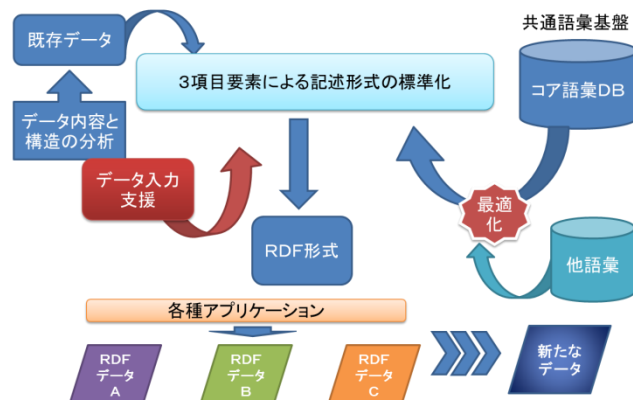


図 5 RDF 生成の流れと活用

Figure 5 Flow and utilization of RDF generation

謝辞 MS-Word のテンプレートファイルの作成にご協力頂いた皆様に、謹んで感謝の意を表する。

参考文献

- [1] RDF (Resource Description Framework) <https://en.wikipedia.org/wiki/RDF>
- [2] Word2Vec <https://en.wikipedia.org/wiki/Word2vec>
- [3] kuromoji: <https://www.atilika.com/ja/products/kuromoji.html>
- [4] DATA.GO.JP サイト, <http://www.data.go.jp/>
- [5] 世界最先端 IT 国家創造宣言 http://japan.kantei.go.jp/policy/it/2013/0614_declaration.pdf
- [6] 情報技術推進機構 (IPA) <https://www.ipa.go.jp/>
- [7] Tim Berners-Lee : Open Data のための 5 つ段階 <http://5stardata.info/ja/>
- [8] CBOW と Skip-gram https://en.wikipedia.org/wiki/Word2vec#Skip_grams_and_CBOW