１　はじめに

陳博\*　久永忠範\*　能登大輔\*　増田翼\*　泊大貴\*\*　満園大輔\*\*　渕田孝康\*

（鹿児島大学　\*大学院理工学研究科　\*\*工学部)

Word2Vecを用いたオープンデータの述語提案手法の研究

2013年にG8で合意されたオープンデータ憲章を皮切りに，世界的にオープンデータへの関心が高まりつつある。オープンデータの活用が推進されていて，国や自治体がオープンデータの公開・活用に取り組んでいる。鹿児島市でも2016年7月にオープンデータ[1]がCSV等公開されているが，このような形では「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」という活用までには至っていないのが現状である。

本研究ではオープンデータ向けRDFの述語に注目し，語彙の共通化を行うために，Word2Vec[2]を用いたオープンデータの述語サジェスト手法を提案する。いくつかのコーパスの差異を比較し，述語ベクトルと単語ベクトルを組み合わせて述語のサジェストを行う。

２　外部状況

オープンデータの公開と活用は日本が国策として推進する重要な課題の一つである。現在RDFとして公開されているデータも名前空間の共通化についてほとんどなされていない。RDFのデータモデルは，リソース情報を主語，述語，目的語の3つの要素で表す。最近，様々な方法で名前空間が提案され，それらの名前空間の中で定義されているクラスやプロパティを述語として利用されている。しかし，実際に市民や民間企業側にはそのような形の名前空間理解し，述語を使ってRDFを作成するのは非常に困難である。また，行財政，国土交通省，経済産業省等多くの省庁から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。全国のデータを複数利用し,データの整備を整えるシステムが開発されていないのが現状である[3]。

３　提案手法

本研究では，5つ星オープンデータ「4」の第4段階のRDF形式に焦点を当てて，述語の語彙共通化を行うため，word2vecで学習した結果で単語の類似度を測り，オープンデータの述語のサジェストを提案する。word2vecとは，単語の意味や文法を捉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。その中でパラメーターに関してword2vecの学習データの次元数100，200，300の３つとword2vecの手法cbowとskipgramの2種類である。

いくつかのコーパスの差異を比較し，word2vecを用いて，をそれぞれ学習し，抽出した語彙を活用し，述語のサジェストを行う。

* 1. 実験１

オープンデータを扱うのに適した辞書を作成したうえでどのようなコーパスを使用するのが適切かを調査することを目的とし，wikipediaの辞書と自治体の辞書の間にどの程度違いが見られるのかを実験した。

まず，コーパスに含まれる名詞で，出現頻度が高いものを上位1000個抽出した。図1のように名詞と出現回数が各名詞ごとに表示されたファイルになった。左側にwikipediaからの辞書の抽出結果を表示した。右側に自治体からの辞書の抽出結果を表示した。

次に，抽出した名詞1000個のベクトルをそれぞれの辞書から抽出した。

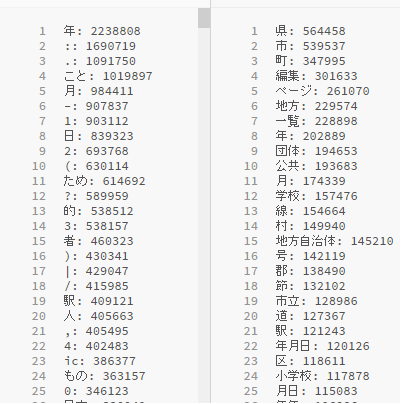


図1.出現頻度上位の名詞

それから，自治体コーパスでの頻出名詞1000個のそれぞれの辞書でのベクトルをcsvファイルに出力し，それぞれの辞書で単語ベクトルのカテゴライズを行った。カテゴライズにはR言語を使用した。R言語は統計処理を行う機能が多いスクリプト言語である[5] 。

さらに，各クラスタの平均ベクトルを求め,そのベクトルをMDS法で二次元化し，二次元座標に示した。Wikipedia辞書のクラスタは，クラスタ14と15のみが独立し，その他のクラスタは似たクラスタとなり，自治体のほうはクラスタ10と13，15のみが独立し，その他のクラスタは似たクラスタになった。多くのクラスタが似た座標に集中していたことが分かった。wikipediaの辞書のクラスタの可視化の結果を図2に，自治体の辞書のクラスタの可視化の結果を図3に示す。

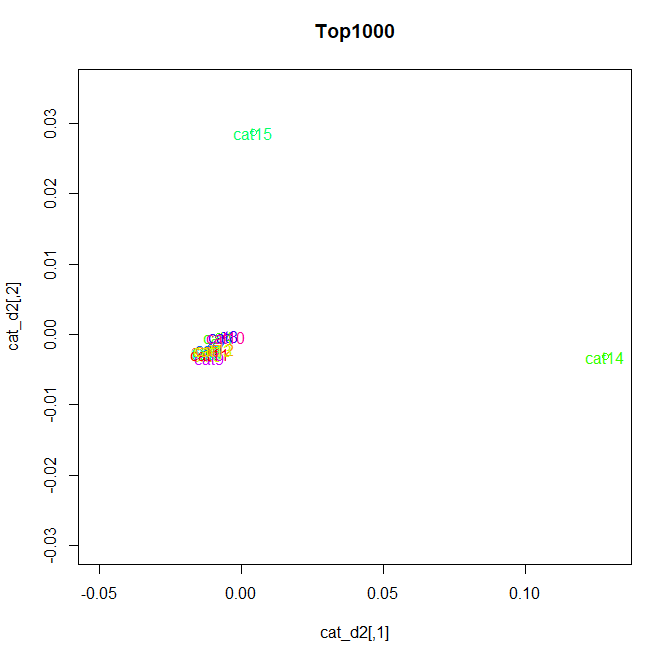


図3. wikipedia辞書のクラスタの可視化

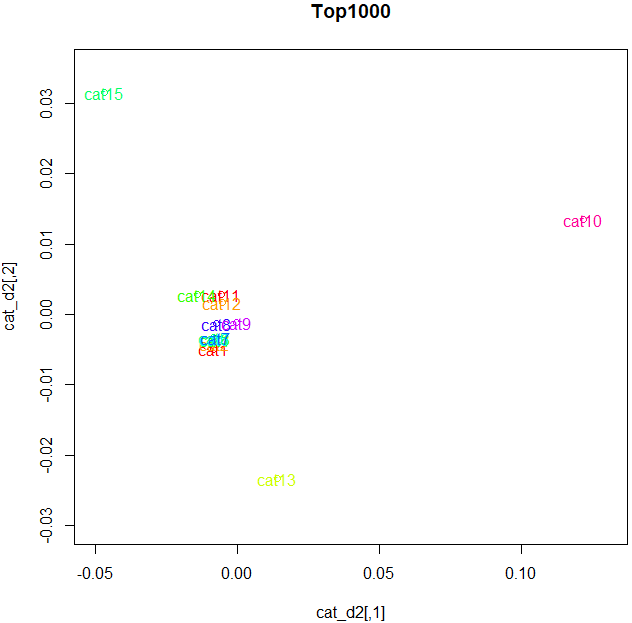


図2.自治体辞書のクラスタの可視化

最後に，各辞書のクラスタ間の類似度を計算し，辞書によってどの程度単語の分類のされ方に違いところを定量的に求めた。自治体ホームページからの辞書のクラスタに含まれる単語数をn，wikipediaからの辞書のクラスタに含まれる単語数をm，各辞書のクラスタの中に含まれる共通の単語数をsとすると,類似度rは，



で表す。式により，二つのクラスタ間に共通の単語が一つもない場合は類似度が0となり，クラスタ間の単語がすべて一致した場合は類似度が1となる。



図4.クラスタ間の類似度算出結果

* 1. 実験2

word2vecを用いて，ベクトルを学習する方法で実験を行った。与えられたcsvファイルの指定した列の全データから，その列のデータを表現するに最も相応しい述語を対象語彙から抽出する。手順として，まず，公開しているオープンデータのcsvファイルを取得する。次に，csvファイルから指定した列のデータを抽出する。それから，各列のデータが分かち書きされ，1行に１つの単語が並べられるデータを取得する。最後,word2vecを用いて，単語ごとに辞書から単語ベクトルを求める。コア語彙及びすべての単語辞書それぞれの中から，最も上位１０単語を表示する。



表1.wikipediコーパスと自治体コーパスで学習した結果

表１に自治体コーパスとwikipediコーパスをそれぞれ学習し，実験した結果である。学習した結果によると，コーパスが異なっても，多くの同語彙が出たことが分かった。

４　まとめ

いくつかのコーパスの差異を比較し，wikipediaからの辞書と自治体からの辞書の間にどの程度違いが見られるのかを実験した。また，オープンデータ向けRDFの述語に注目し，語彙の共通化を行うために，word2vecを用いて，複数のコーパスを使用し，単語ベクトルを学習し，実験を行った。今後，鹿児島市のオープンデータだけでなく，全国のオープンデータを使用していきたい。word2vecで抽出した語彙を活用し，述語ベクトルを組み合わせて，RDF形式に変換していく方法を検討する。

謝辞

本研究はJSPS科研費JP16K00421の助成を受けたものです。

参考文献

1. 鹿児島市オープンデータhttps://www.city.kagoshima.lg.jp/jousys/opendata.html
2. Word2Vec　 https://deepage.net/bigdata/machine\_learning/2016/09/02/word2vec\_power\_of\_word\_vector.html
3. 著者 庄司 昌彦 “国内における活用環境整備”, 情報処理学会論文誌, vol.54, no-12, pp. 1244 - 1247
4. 5つ星オープンデータ http://5stardata.info/ja/
5. プログラマーのためのR言語入門 – Qiita

<https://qiita.com/stkdev/items/6aba2c1db2fa056170ae>