１　はじめに

陳博　渕田孝康久永忠範能登大輔郭崇

（鹿児島大学　理工学研究科)

自治体オープンデータ向けの単語ベクトルの学習法の研究

近年，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，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．2016年7月に鹿児島市においてオープンデータ[1]がCSVなどの形式で公開された．このような環境整備は進めているものの，利用する側である市民や民間企業からすればオープンデータを活用しやすい環境が整っているとは言い難い．そこで，オープンデータを現在よりも幅広い層が活用できる策を講じる必要がある．

本研究では，自治体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]を用いて，単語ベクトルを学習する方法を提案する．

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近似性を測る手法を提案する．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．

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．流れとして下記の手順である．

1. 自治体が公開しているオープンデータのcsvファイルを取得する．
2. csvファイルから指定した列のデータを抽出する．
3. 各列のデータが分かち書きされ（区切る），1行に１つの単語が並べられるデータを取得する．
4. word2vecを用いて，単語ごとに辞書から単語ベクトルを求める．
5. 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．

４　実験結果

今回の実験では,鹿児島市のオープンデータ　[1] “避難所(4-1\_hinannjyo.csv)”を使用した．また，単語総数157,339,847,単語種類269,399を使用した．結果を以下の表に示す．

表1 cbowの結果



表2 skipgramの結果



表1と表2にcbow，skipgram手法で100次元，200次元，300次元のベクトル空間で，各次元ごとに順位，単語，距離を表す．表1より学習手法の違いでの結果が異なることが分かった．また,“---for Core---”はコアの語彙から“---for All---”は全語彙（単語辞書）から学習した結果となる．コア語彙からの出力の結果には共通している語彙があるが，全語彙からの出力は非常に相違がある．そして，cbow手法では小学校や公民館等オープンデータの項目に近い単語が多く抽出されたが，skipgram手法では地名等の固有名詞や独特の単語が抽出されたことがわかる．

5　まとめ

自治体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，単語ベクトルを学習する方法の提案および実験を行った．今後，単語ベクトルを求める精度を向上するため，改良策を検討する．また，鹿児島市のオープンデータだけでなく，ほか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，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/