第28回年次大会予稿

オープンデータのRDF化のための項目名のクラスタを使用した述語のサジェストに関する研究

A study on suggestion of predicates using the clustering of item names for making RDF of Open Data

陳博[1*], 渕田孝康[2], 泊大貴[1], 久永忠範[1]
Bo CHEN[1*], Takayasu FUCHIDA[2], Daiki TOMARI [1],Tadanori HISANAGA[1]

1 鹿児島大学大学院理工学研究科
Graduate School of Science and Engineering, Kagoshima University
〒890-0065 鹿児島市郡元1-21- 40
E-mail: chinhaku204@gmail.com
2 鹿児島大学大学院理工学研究科
Graduate School of Science and Engineering, Kagoshima University
〒890-0065 鹿児島市郡元1-21- 40
E-mail: fuchida@ibe.kagoshima-u.ac.jp
*連絡先著者 Corresponding Author

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. 鹿児島市で2016年7月にオープンデータがCSV等公開されている. 地方自治体のオープンデータは, データ形式・フォーマットの違いにより開示されても積極的な活用まで至っていないのが現状ではある. 本研究では, RDFに焦点を当てて, 述語の語彙共通化を行うため, オープンデータの項目名をクラスタリングし, 割り当てられたカテゴリを教師信号として入力し, 深層学習を行い, 述語のサジェストを提案する. また, 述語のサジェストの結果を示した.

In recent years, there has been increasing interest in Open Data. The utilization of Open Data has been promoted, and many organizations including the national government, local governments and other organizations are working on publishing and utilizing Open Data. Open data of Kagoshima City has been published in July 2016 as CSV format. Under Open Data of local governments, even if they are disclosed due to the difference in data format and the lack of linkage of data, there is no active utilization yet. In this research, we focused on the vocabulary that corresponds to the predicate of RDF form for Open Data, and proposed a method to suggest predicates by the clustering of item names for RDF of Open Data as teacher signals of Deep Learning. Then, we presented results of the predicate suggestion.

キーワード: オープンデータ, CSV,述語サジェスト,クラスタ, Word2Vec
Open Data, CSV, Suggestion of Predicate, Cluster, Word2Vec

1 はじめに

2013年6月に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，国や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．

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[1]．

本研究では，RDFに焦点を当てて，述語の語彙共通化を行うため，オープンデータの項目名をWord2Vec[2]でベクトルし，クラスタリングし，割り当てられたカテゴリを教師信号として，Deep Learningで学習し,述語のサジェストを提案する．また,述語の予測の結果を示した．

2 現在のオープンデータ状況

オープンデータの現状はどのようになっているか，オープデータの活用促進を行う取込みにについて調査した．

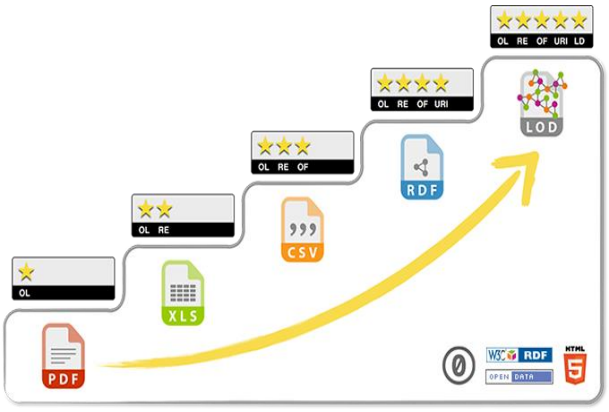

図1 オープンデータの5段階

2.1 オープンデータの5段階

オープンデータのため，Tim Berners Leeが図1に示す5つの段階とそれに値するデータ形式を提唱した[3]．

第1段階から第3段階は,オープンライセンスで提供されていたり,特定のアプリケーション,構造化されたデータとして公開されているデータ形式であり,第4段階から第5段階は,物事の識別にURIを利用していたり,他のデータをリンクしているデータ形式である．

2.2 政府におけるオープンデータ

政府は,2014年10月1日にデータカタログサイトDATA.GO.JP[4]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2020年04月現在，26,809のデータセットとして開示されている．

鹿児島市で2016年7月にオープンデータがCSV等公開されている[5]．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．

3 提案手法

本研究では，Tim Berners Lee が提唱したオープンデータの５つの段階のデータ形式(図1)で，機械判読可能である第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，オープンデータの項目名をクラスタリングし，割り当てられたカテゴリを教師信号として，Deep Learningで学習し,述語のサジェストを提案する.

3.1 オープンデータの収集

地方公共団体が公開するオープンデータを「DATA.GO.JP」のデータカタログサイトのデータベースサイトの地方公共団体の359のリンクサイトから約3万件のオープンデータを収集した．そのデータでCSV形式である14286ファイルを抽出した.

3.2 Word2Vecで単語ベクトル表現

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．Word2Vecの手法は図2に示すCBOWとskipgramの2種類である.

Word2Vecは単一の単語をベクトル化するが,複合語は単語ベクトル空間で表現できる．例えば，図3の単語ベクトルの計算イメージに示すように「鹿児島」と「県」の和で「鹿児島県」で表現できる.

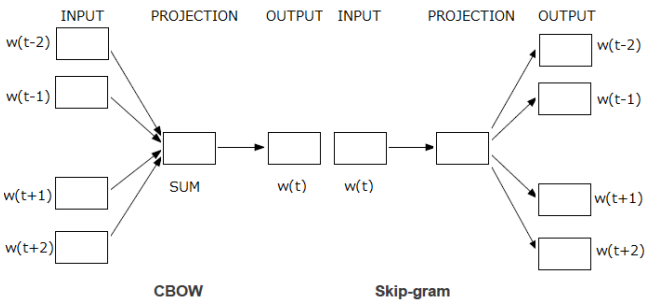

図2 Word2VecのCBOWとSkip-gramモデル

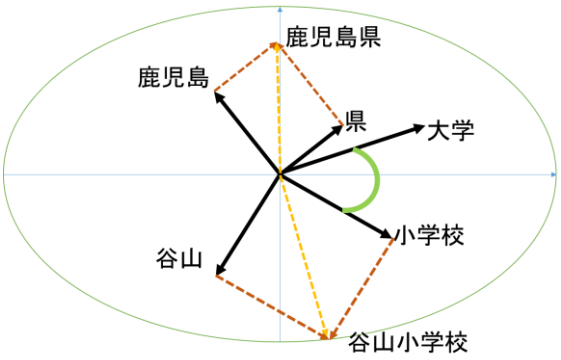

図3 ベクトル空間で単語の合成ベクトル

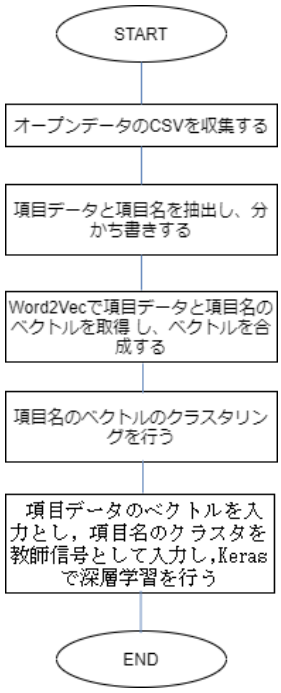

図4 アルゴリズムのフローチャート

3.3 述語のサジェストの手順

図4にアルゴリズムのフローチャートを示す．その詳細な手順は下記の通りである.

1. 自治体が公開しているオープンデータの csv ファイルを収集する.
2. 項目データと項目名を抽出し,分かち書きする．日本語の分かち書きとは文を語に分解し,それぞれをスペースによって区切る書き方である．Janome という Python の形態素解析エンジンを使用する.
3. Word2Vec で項目データと項目名のベクトルを取得し,図3に示すようにベクトルを合成する.
4. 項目名のベクトルのグラスタリグを行う.

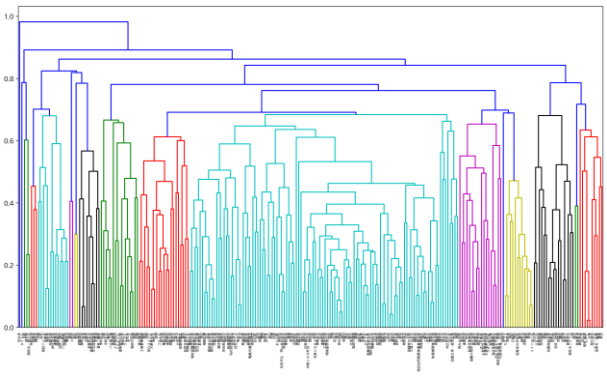

図 5 階層的クラスタリングによって得られたデンドログラム

表1 階層クラスタリングによる項目名クラスタ

item name	cluster	item name	cluster	item name	cluster
特典内容	11	住所	30	名称	31
表示内容	11	住所・場所	30	項名称	31
問合せ内容	11	店舗住所	30	目名称	31
活動内容	11	所在地	30	節名称	31
内容	11	工場所在地	30	細節名称	31

図 5 に階層的クラスタリングによって得られたデンドログラムを示す. クラスタとして, 50 を設定したときの項目名のクラスタの一部を表 1 に示す. 項目データのベクトルを入力とし, 項目名のクラスタを教師信号として入力し, 深層学習を行う. オープンソースの深層学習のライブラリとし, Keras を用いた. 図 6 に学習のデータの概念図を示す.

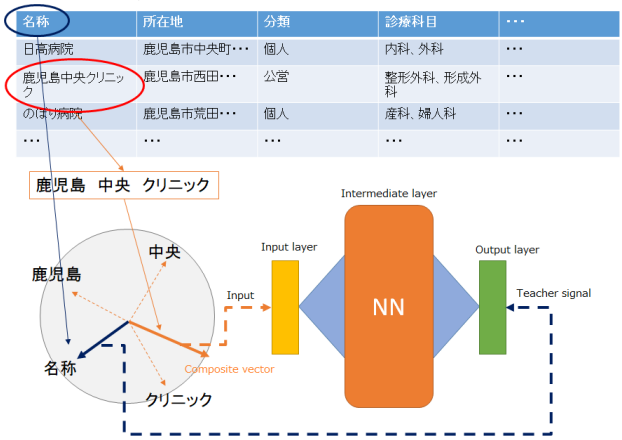

図6 データの流れ

手順2で抽出した項目名と項目データのペア（学習用のデータセット）数は 1, 245, 709である.

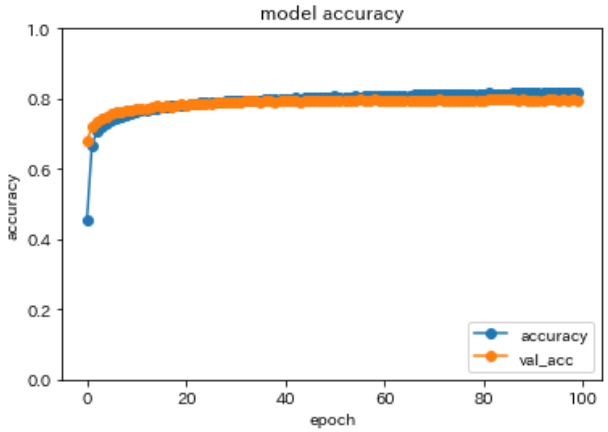

図 7 項目名のクラスタを使用し学習の精度

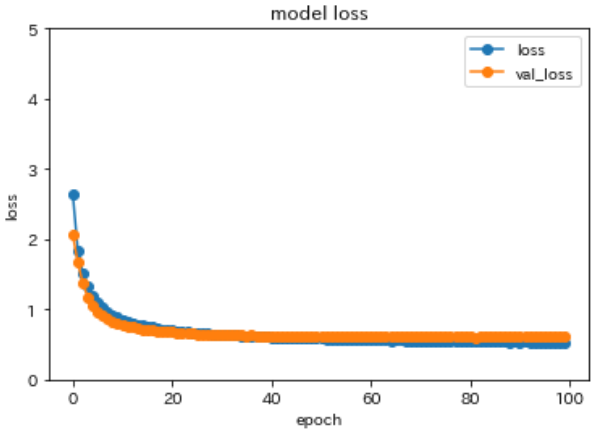

図8 項目名のクラスタを使用し学習の損失

4 実験と結果

項目名のクラスタをの使用した実験と LabelEncoder[6]より得た項目名のラベルを使用した実験の 2 つの実験を行った.

4.1 実験1

Word2Vec で得た項目データと項目名の 50 次元の 100, 000 データセットを使用した. このうち 80, 000 インスタンスをトレーニングセット, 20, 000 インスタンスをテストセットとして使用した.

50 次元のベクトルのデータセットを使用し, 100 回学習した. 図 7 は項目名のクラスタを使用して深層学習した精度を示し, 図 8 は学習の損失を示す. それぞれ

表2 深度学習で予測したクラスタの結果の一部

item data	predict cluster	list of the cluster's data
日高病院	24	機関名,実施機関,医療機関名,指定医療機関名…
鹿児島中央クリニック	23	施設名,事業所名,病院名,事業者名…
のぼり病院	23	施設名,事業所名,病院名,事業者名…
高橋	27	氏名,氏名漢字,フリガナ,備考
陳	27	氏名,氏名漢字,フリガナ,備考
中島	27	氏名,氏名漢字,フリガナ,備考
北寺尾三丁目	31	町名,地区名称,名前,名称…
西田1丁目	31	町名,地区名称,名前,名称…
鴨池1丁目	31	町名,地区名称,名前,名称…
鹿児島大学	23	施設名,事業所名,病院名,事業者名…

の縦軸は精度または損失を表し,横軸はエポックを表す．エポックとはデータセットを使用しニューラルネットワークでトレーニングする回数である．

表2に項目名と項目データを入力し,深層学習で予測したクラスタの結果の一部を示す．この結果から予測した結果はおおむね適切なクラスタをサジェストできていることが分かる．項目データが異なっていても,予測クラスタの結果は同じになる場合がある一方で,似た言葉でも異なる述語にサジェストされている場合でも見られた．これより,階層的クラスタリングにより一番良い項目名のクラスタを見つける方法と,クラスタの最も適切な数はどれかを検討する必要がある．

4.2 実験2

前述した手順4で，項目名のクラスタリングのクラスタを使用せずに，代わりにpythonのLabelEncoderで割り当てられたラベルを教師信号として学習する実験を行った．LabelEncoderは語彙からカテゴリデータを取り扱う時ラベルを0～n-1の値に変換できる．ここでnはクラスの種類数である．LabelEncoderを使用し，語彙からラベル(0～n-1)を変換する例を図9に示す．

図10と図11に実験2の結果を示す．図7の精度と図10の精度を比較すると，項目名のクラスタを使用した場合の学習の精度はLabelEncoderを使用した結果より高いことが分かった．

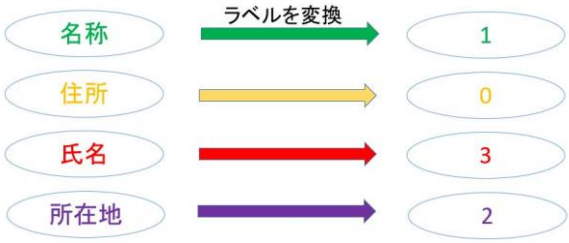

図 9 LabelEncoder でラベルの変換の例

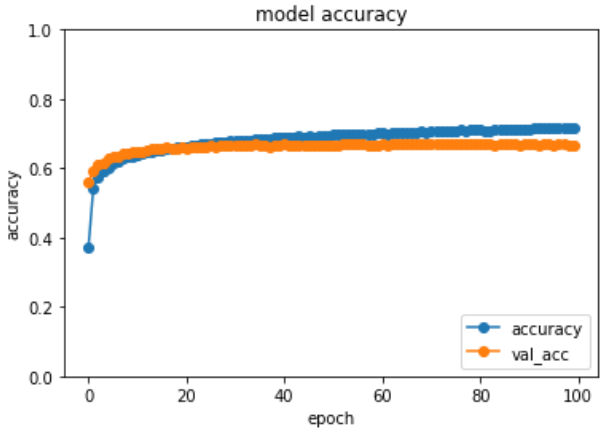

図 10 項目名のクラスタを使用せず学習精度

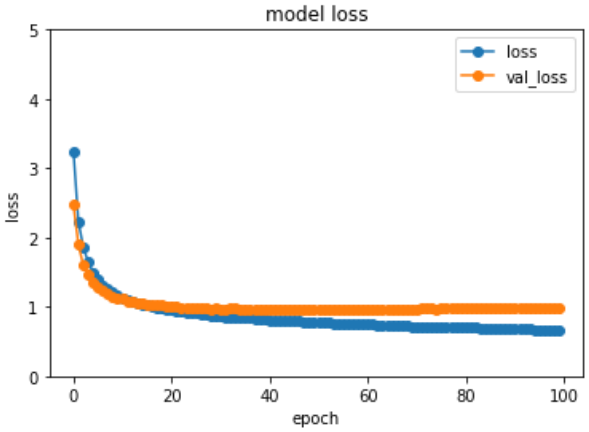

図 11 項目名のクラスタを使用せず学習**損失**

5 まとめと今後の課題

本研究では，実際に公開されているオープンデータを活用して，RDF形式に焦点を

当てて，述語の語彙共通化を行うため，オープンデータの項目名をクラスタリングし，割り当てられたカテゴリを教師信号とし，深層学習を行い,述語のサジェストの提案を行うシステムを作成した.

今回は単一の単語に対する述語サジェストを行ったが，今後は単一の列に対するサジェストについての実験を行う予定である．また，述語をサジェストしたクラスタを活用し，簡便にRDF化できるシステムを提供していきたい.

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] 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

[2] Word2Vec,https://en.wikipedia.org/wiki/Word2vec

[3] Tim Berners Lee :5つ星オープンデータ,http://5stardata.info/ja/

[4] DATA.GO.JP サイト,https://www.data.go.jp/

[5] 鹿児島オープンデータ,https://www.city.kagoshima.lg.jp/ict/opendata.html

[6] LabelEncoder, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html