１　はじめに

陳博\*　久永忠範\*　泊大貴\*　永住拓視\*\*　 濱平陸\*\* 渕田孝康\*

（鹿児島大学　\*大学院理工学研究科　\*\*工学部)

オープンデータの述語サジェストのための

Word2Vecを用いた項目名のクラスタリング

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が，データ形式の統一がなされていないため，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。オープンデータを連携するためにはRDFの述語を統一することが必要である。

本研究では，適切な述語をサジェストするために，公開されているオープンデータの項目名を抽出し，Word2Vec[1]で得られた単語ベクトルを使用し，階層的クラスタリングを行った。さらに，クラスタリングの条件を変えることで，生成されるクラスタの適切さを定量的に評価した。

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。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，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[2]。

３　提案手法

本研究では，5つ星オープンデータ[3]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，

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。また，クラスタリングの条件を変えることで生成されるクラスタの適切さを定量的に評価する。

* 1. 実験１

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。公開されているオープンデータCSVファイルの項目名を抽出し，Word2Vecで得られた項目名の合成ベクトルを使用して階層的クラスタリングの実験を行った。

まず，全国のオープンデータより収集したCSVファイルから，項目名を抽出する。

次に，項目名が分かち書きされ，Word2Vecで得られたベクトルを抽出し，それぞれの合成ベクトルを求める。

それから,項目名の合成ベクトルを用いて，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。カテゴリ数を40に設定し，割り当てられたカテゴリ結果を表1に表す。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[4] 。

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。項目名のクラスタの可視化の結果を図1に示す。図1の実験条件はカテゴリ数が40を設定した。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。



表1.割り当てられたカテゴリの一部結果

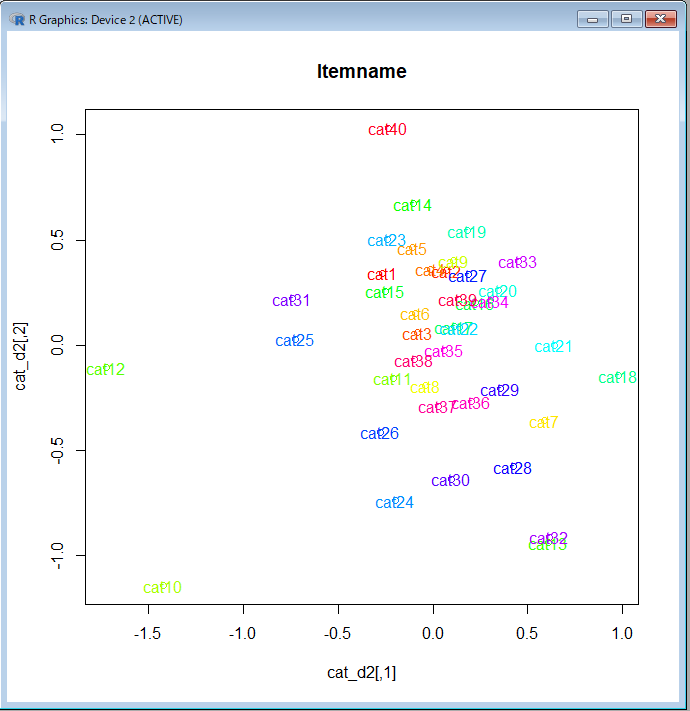


図1.項目名のクラスタの可視化

* 1. 実験2

実験1のカテゴリ数および結果が正しいのかどうか評価する必要である。そこで，クラスタリングの条件を変えることで，生成されるクラスタの適切さを定量的に評価するために，DCSE(Degree of Category for Similarity Evaluation)を提案する。DCSEとは，ひとつのクラス内に含まれる単語のペアの類似度の分布を利用し，カテゴリを評価する手法である。

DCSEのフローチャートを図2に示す。まず，カテゴリ内に含まれるペア単語の類似度を求める。

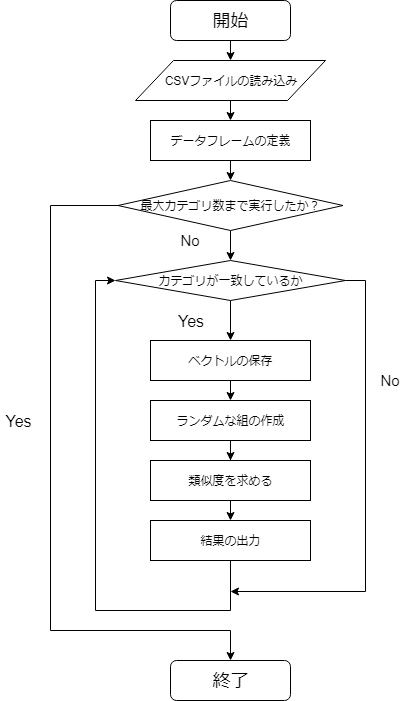
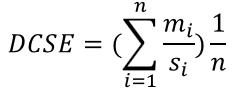


図2. DCSEのフローチャート

次に，得られた類似度の分布ををエクセルでグラフ化する。一つのカテゴリ分布の平均mと標準偏差sを求める。それから，その全てのカテゴリの平均を評価値(DCSE)として，使用する。カテゴリの最大数をn，カテゴリ平均類似度をm，カテゴリ標準偏差をsとし，評価値(DCSE)を

　　　　　　　　　　　(1)



で計算する。

カテゴリ数が40～100，200，250を設定し，クラスタリングした結果のDCSEを図3に示す。結果によると，DCSEはカテゴリ数が増えるとともに下がることなく，上昇していくことが分かった。なぜなら，m/sにおける平均は，カテゴリ数が多くなるにつれ，一つのカテゴリの平均は大きくなり，ばらつきも少なくなるからだと考えられる。

さらに，カテゴリの類似度を使用し，評価する。計算式(E)は

　　 (2)



で表す。ここで，αは重みであり，Bはカテゴリ数が増えると減っていく値（検証中）を表す。適当にαを変え，実験を行

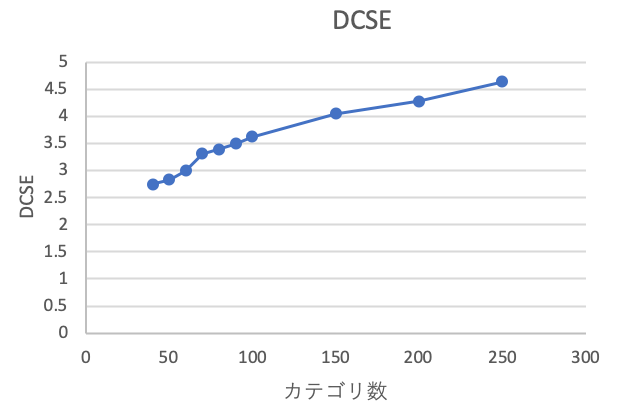


図3.クラスタリングしたDCSE結果

う。

４　まとめ

公開されているオープンデータCSVファイルの項目名を抽出し，Word2Vecで得られた単語ベクトルを用い，階層的クラスタリングを行った。さらに，クラスタリングの条件を変えることで，生成されるクラスタの適切さをDCSEで定量的に評価した。また，クラスタの適切さにほかの評価手法も検討していく。今後，割り当てられたカテゴリを教師信号として活用し，Deep Learningで学習し，述語のサジェストを行い，RDF形式に変換していく方法を検討する。

参考文献

1. Word2Vec　 https://deepage.net/bigdata/machine\_learning/2016/09/02/word2vec\_power\_of\_word\_vector.html
2. 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
3. 5つ星オープンデータ http://5stardata.info/ja/
4. プログラマーのためのR言語入門 – Qiita

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