Word2Vec を利用した観光地の分類手法

中村 みなみ* 乙武 北斗** 吉村 賢治**
(福岡大学 *大学院工学研究科 **工学部)

1 はじめに

自治体等が観光地を巡るルートを生成するにあたって、ルートにストーリー性を持たせたいという希望がある。そこで、本稿では観光地の案内文をもとに、ストーリー性を持つ観光地のグループを求める手法を提案する。文書分類の先行研究としては、TF/IDF を利用した手法[1]やDoc2Vecを利用した手法[2]等が挙げられる。本稿では Word2Vec を用いた、TF/IDF の精度向上を目的とする。

2 提案手法

2.1 ベクトル空間の改善

TF/IDF で出てきた単語を Word2Vec でベクトル化してクラスタリングを行う。今回、クラスタリングには bayon[3]を使用した。

(1) TF/IDF で抽出した名詞に対して Word2Vec を用いてベクトル表現を求める(200 次元)。
(2) ベクトルされた単語を bayon でハードクラスタリングをする。この結果同一のクラスタに属する単語は一つの単語として扱い、案内文をクラスタリングする際に一つのベクトル軸とする。
(3) (2)で次元削減されたベクトル空間で各案内文のハードクラスタリングを行う。
(4) クラスタの中心ベクトルとの類似度を用いて再度案内文のソフトクラスタリングを行う。

2.2 ストーリー性の定義

本研究では、共通の人物や出来事と関連がある観光地を巡るルートにはストーリー性があると考える。共通の人物や出来事は案内文から TF/IDF でキーワードとして抽出されるはずであり、ベクトル空間法を用いた文書の類似性をストーリー性の第一次近似として利用する。

3 実験

3.1 実験内容

実験で使用したデータは以下の通りである。

(1) Word2Vec の学習に用いたデータ
Wikipedia の「神社」「天満宮」「文化財」「日本の国宝」「観光」「観光地」のカテゴリを含んでいる記事約 9700 記事(約 81MB)
(2) 分類対象のデータ
インターネットから収集した京都観光地寺院 110 件、京都観光地他 10 件計 120 件の案内文

今回の実験では、Word2Vec で次元削減してクラスタリングした結果と、次元削減をせず TF/IDF の結果のみでクラスタリングをしたものを比較する。

また、各クラスタに属するか否かはクラスタの中心ベクトルとの類似度が 0.1 より高いものとする。

3.2 実験結果と考察

全体的な実験結果を表 1 に示す。表 1 の平均要素数は各クラスタに属する観光地の平均数であり、次元数はクラスタリングをする際のベクトルの軸の数である。

表 1 全体結果

	従来手法	提案手法
クラスタ数	28	27
平均要素数	4.61	5.78
次元数	3559	3191

全体としてはクラスタの数は減り、それぞれのクラスタに属する観光地の数が増加している。これは「子ども」と「子供」という同じ単語で表記の揺れが生じた単語や「オープン」と「開館」等意味的に類似している単語が従来の手法では違う単語として処理されていたが、Word2Vecを利用したことで 1 つの単語として処理されることになったためと推測できる。しかし、大幅な次元削減がみられない原因としては以下の問題点が考えられる。

実験の問題点

実験の問題点として以下の 4 つが考えられる。

(1) 学習データの不足
次元数は削減できているが、軸となる単語 3559 のうち 349 の単語はベクトル化されていなかった。ベクトル化されていないという事は学習データに出現しない単語であったということである。学習データを更に増やす必要がある。
(2) 観光地案内文の内容
案内文の中には近隣の観光地を含む観光ルートを提案しているものがあり、類似度を上げる原因になっている。このような文を削減することでより精度の良いクラスタリングができると考えられる。
(3) 人名の分割
現在の手法では人名を「姓」「名」で分割している。人物に関しては「姓」「名」で分けず 1 つの単語として処理するという改良が考えられる。
(4) 形態素解析の精度
形態素解析の辞書に登録されていない人名が有るために形態素解析の精度が低下している。これらを登録して形態素解析の精度を改善することでクラスタリング精度の改善になると考えられる。

4 まとめ

本稿では、Word2Vec を用いて次元削減をし、案内文をストーリー性のあるクラスタに分類する手法を提案した。今後は 3.2 で示した問題点について改善していきたい。

参考文献

[1] Willi Richert Luis Pedro Coelho 著 斎藤 康毅訳 オライリー・ジャパン 実践機械学習システム
[2] Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. (http://arxiv.org/pdf/1405.4053v2.pdf)
[3] fujimizu/bayon Tutorial_Japanese (https://github.com/fujimizu/bayon/wiki/Tutorial_Japanese)