word2vec を用いた自然言語の意味頻度分布解析

松浦弘樹 * 田中久美子 ** 内田誠一 *
(*九州大学, **東京大学)

1 はじめに

テキストデータの増大とその処理の必要性により, 言語を統計的に捉え, 数学的なモデル化を行う必要がある. 代表的なモデルである Zipf 則 [1] は, 単語の出現頻度を $f(r)$, r を頻度のランクとする分布がべき乗則 $f(r) \sim r^{-\alpha}$, 特に指数 α が 1 のべき分布に従うことを示す経験則である. しかし, 実際の単語の頻度分布では高頻度と低頻度の区間で異なる変化が生じており, 言語ごとの差異が多少は生じる.

ところで, 最近は単語の意味的な情報をベクトル空間で表す研究が注目されている. 意味を数値ベクトル化することにより, 単語や文字のように数学的なモデルとして表すことも可能になる.

本研究では, テキストの意味を頻度分布化する方法について検討する. 意味頻度分布の可能性として, Zipf 則のようなべき乗則が成り立つこと, 意味は単語よりも抽象度が増しているため, 言語によらず似た分布になることが期待される. この点を確認するために, 同一内容のテキストである英語と日本語の旧約聖書に対し, 意味頻度分布を作成し単語の頻度分布と比較した結果を示す.

2 意味頻度分布の推定

本研究では, 入力として旧約聖書のテキストを与え, 意味頻度分布を作成した. 具体的には, 意味頻度分布を作成する際に必要な意味情報の数値化には, word2vec[2] を用いて行った. word2vec とは, ニューラルネットワークの学習によって単語の意味を数値ベクトル化する手法である. 本研究では, 旧約聖書の英語 (総単語数:631404, 異なり単語数:8186), 日本語 (総単語数:871748, 異なり単語数:5349) をベクトル化すべく, それらを用いて学習を行った. ただし, これだけでは word2vec の学習には少ないために, 各言語の wikipedia を用いたネットワークの事前学習を行った.

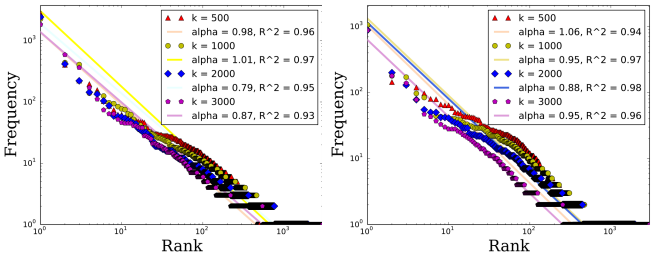
次に, 意味ベクトル化された単語集合に対して, k-means 法を用いてクラスタリングを行った. この際, クラスタ数 k は意味の量子化程度のパラメータとして重要である. そこで, テキストの異なり単語数から, クラスタ数については $k = 500, 1000, 2000, 3000$ の 4 通りを試した.

最後に, クラスタリングの結果から意味頻度分布を作成した. 頻度 $f(r)$ は各クラスタの出現単語数とし, r は各クラスタを頻度の大きいものから決定した.

3 結果

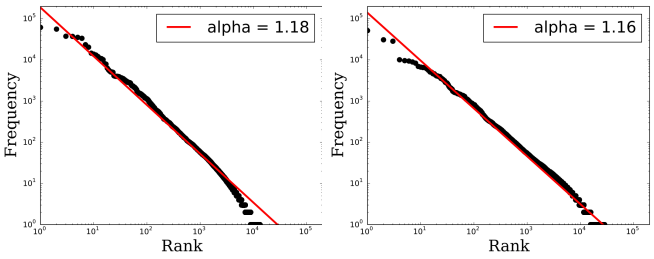
英語と日本語の旧約聖書を用いて, 前述した手法で意味頻度分布を作成し, 比較のために単語の頻度分布も作成した. 頻度分布は, 横軸に単語 (意味) のランク r, 縦軸に出現頻度 $f(r)$ をプロットした両対数グラフである. 同図には, 最小二乗法による Zipf 則への回帰の結果, パラメータ α, 意味頻度分布にはクラスタ数 k での結果と, 回帰結果の決定係数 R^2 の値も示している.

図 1 は意味頻度分布の結果であり, 図 1(a), 1(b) のどちらも, k が大きくなるに連れて分布はより直線に近づいていることがわかる. よって, 意味の頻度分布は単語と同様にべき乗則に従っている可能性がある. しかし, α の値が単語の頻度分布ほど 1 に近くないことから, Zipf 則が成り立つ分布ではないと考えられる.



(a) 英語 (b) 日本語

図 1: 意味の頻度分布



(a) 英語 (b) 日本語

図 2: 単語の頻度分布

また, 図 1 の R^2 が最も 1 に近い英語 ($k = 1000$), 日本語 ($k = 2000$) のグラフと, 図 2 を比べると, 意味頻度分布は単語の頻度分布よりも似た変化をしているように見える. 単語の頻度分布では, r の値が大きくなると, 図 2(a) では直線的な減少を続けるが, 図 2(b) では減少量が大きくなり直線的ではなくなる. 一方, 意味の頻度分布では, どちらも一度大きく下がり直線に落ち着くという変化をしており, 分布の変化に大きな差はない. しかし, 最小二乗法による回帰で導出した α の分布間での差異が, 単語の分布間の差異に比べて大きいため, 意味が単語よりも似ている分布であるとは言えない.

4 まとめ

テキストから意味の頻度分布を作成する方法として, word2vec を用いて意味ベクトル化を行い, k-means を用いてクラスタリングを行う方法を検討し, その手法に基づき実験を行った. 今後の課題として, より多くの言語やジャンルのテキストを用いて, 意味頻度分布がどのような分布になるのか, 今回言語間で意味頻度分布に差が出た理由, を調査することが挙げられる.

参考文献

[1] Zipf, G. K. "Human Behaviors and the Principle of Least Effort", An Intoroduction to Human Ecology. (1944).

[2] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. "Distributed Representations of Words and Phrases and their Compositionality", in Advances in Neural Information Processing Systems, (2013)