5W-08

ニュース記事に含まれる固有表現を用いた国家間の関連性分析に関する検討

葛野　航希 †
東京都立産業技術高等専門学校 †

横井　健 ‡
東京都立産業技術高等専門学校 ‡

1　はじめに

昨今、情報化によってインターネット上で大量の情報が飛び交い、その中の情報を分析することは、情報科学の分野においても重要である。その中で、一国家について調査する際にも一国家の情報だけでなく、関係する他国とのつながりも含めることも重要となっている。

国家間の関連性分析を行う従来の研究では、各国の報道機関が Web 上に公開しているニュース記事のクラスタリングの誤分類と記事のトピックから、各国にどのトピックで関連性があるのか分析する手法が存在する [1]。

しかし、[1] では関連性を求めるためのトピック選定とトピックと国との関連性の判断は手動であり、トピック以外の単語については考慮されていなかった。そこで、本研究では各国のニュース記事に存在する国特有の固有表現に着目し、ニュース記事の誤分類を利用することでそれらに重要度を付与することから国家間の関連性の強度を求めることを目標とする。

2　TF-IDF、相互情報量を拡張した単語重要度

[2] では単語重要度の計算方法として従来手法の TF-IDF を特定分野のコーパス D_p, 一般分野のコーパス D_n のドキュメント集合から、それぞれに対して単語の出現回数 TF 値 を求め、単語重要度を計算する方法が提案されている。また、相互情報量を組み入れることによってたまたま使われた単語の重要度を低く設定しながら、特定分野でのみ使われる単語の重要度を高く設定できる手法も提案されている。

3　提案手法

本節では、本研究で行った手法について述べる。まず、英語版 Wikipedia[*1] の一国家に関連するカテゴリからページタイトルを取得し、それを国特有の固有表現辞書とする。地名の辞書作成には Simplemaps[*2]を利用する。

次に、ニュース記事を Doc2Vec によってベクトル化し、 k-means 法を用いてクラスタリングを行う。その際のクラスタ数は記事の発行国数とする。ラベルには、ラベリング済でないクラスタについて、各クラスタ内で発行国の割合が一番大きい発行国をラベルを割り当てる。その後、割り当てたクラスタをラベリング済とし、これを繰り返すことでラベリングを行う。

さらに [2] の計算手法を用いて、各国の辞書に含まれる固有表現の単語に重要度を付与する。D_p に発行国とラベルが同一の記事、D_n に発行国とラベルが異なる記事、単語集合 T に発行国の辞書を割り当て、計算する。

そして、一国家の一記事の中に含まれる別国の辞書の単語の出現頻度を調べ、重要度との積の合計を取ることでその記事スコアとする。一

Inter-National Relations Analysis Focusing on Proper Nouns Included in News Article

† Koki Kuzuno, Tokyo Metropolitan College of Industrial Technology
‡ Takeru Yokoi, Tokyo Metropolitan College of Industrial Technology

[*1] https://en.wikipedia.org
[*2] https://simplemaps.com

国家の記事スコアの合計が一定以上のものをカウントし、発行国の記事全てに対する出現確率を別国への興味スコアとする。

4 実験

本節では、本研究で行った実験方法とその結果について記す。

4.1 実験方法

まず、固有表現の辞書の作成は 1) 人名 2) 組織名 3) 企業名 4) 地名を対象とした。国家名と取得できたその国家に関連する単語の数を表 1 に示す。

次に、アメリカ、イギリス、カナダの Web で公開されているニュースサイトから記事を収集した。記事は Doc2vec によってベクトル化した後、クラスタ数は 3 とし k-means 法によってクラスタリングを行った。

一つの発行国とラベルが同一の文書を D_p、同一でないの文書を D_n とし、[2] の 2 つの手法による単語重要度を順に計算した。このとき、パラメータは 2 つの計算方法ではそれぞれ $\alpha = 0.2$、$\alpha = 1.1, \beta = 2.7$, 閾値 k は 7500 と 0.1 とした。最後に閾値を超えるものをカウントし、発行国のドキュメント集合に対する出現確率を求めた。

表 1　国別固有名詞辞書の単語数

国家名	単語数
USA	719,283
GBR	553,301
CAN	135,689

4.2 実験結果

TF-IDF を拡張した計算法での興味スコアの順位と相互情報量を拡張した計算法での興味スコアの順位をそれぞれ表 2、表 3 に示す。どちらの表を見てもアメリカ、イギリスが相互に順位が高く、先進国としてメディアが大きく報道していると考えられる。しかし、カナダへの興味スコアの順位が低く、これは国際的な影響力が他二国よりも小さいためだと考えられる。また、表 1 を見ると単語数が他二国より少ないため、他二国の記事に出現する単語が少なかったためであるとも考えられる。

表 2　TF-IDF を拡張した計算法での興味スコア

発行国 \ 対象	can	usa	gbr
can	×	4	2
usa	6	×	1
gbr	5	3	×

表 3　相互情報量を拡張した計算法での興味スコア

発行国 \ 対象	can	usa	gbr
can	×	3	4
usa	5	×	2
gbr	6	1	×

5 まとめ

今回、ニュース記事の誤分類を利用し固有表現に重要度を付与することから国家間の関連性の強度を求めることは一部達成できた。しかし、各国の単語数の差によって影響が出ているとも考えられ、今後注視する必要がある。

謝辞

本研究は JSPS 科研費 19K12110 の助成を受けたものである。

参考文献

[1] 生島良太, 横井健, 「ニュース記事を用いた国家間の関連性の要因分析」, 信学技報, Vol. 118, No. 210, pp. 41-46, 2018.

[2] 滝川真弘, 山名早人, 「特定分野を対象とした単語重要度計算手法の提案と Twitter における専門性推定への適応」, FIT2016, RD-001, 第 2 分冊, 2016.