国特有の単語に着目した多言語のニュースによる国家間関連性分析

葛野 航希[1] 横井 健[1]
[1] 東京都立産業技術高等専門学校
kouki_kuzuno-public@yahoo.co.jp takeru@metro-cit.ac.jp

概要

各国が注目しているニュースをイベントレベルで分析する手法が存在した．しかし，既存の手法では国家間の潜在的な結びつきは取りこぼす恐れがある．そこで，各国特有の単語をニュース記事から Wikifier を用いて抽出し CrRv にて単語重要度を付与し，各国家間の関連を定量化する手法を提案する．5 カ国のニュース記事と特有の単語を用いて国家から国家への関心スコアを算出し，スコアに影響する単語を抽出した．

その結果，政治などの大きなイベントに関連する単語だけでなく，文化的，経済的な単語が高い寄与度を示し，各国間の恒常的な結びつきを明らかにすることができた．

1 はじめに

各国のニュースメディアが報道する内容には，その国自身のニュースだけでなく，関連する他国のニュースが扱われる場合がある．これに着目した研究として，Xi Chen らは複数言語の多数の国家のニュース記事を対象とし，イベントレベルでのニュースの多様性や共時性から世界各国がどのようなイベントに興味を持っているかと各国の関連を分析している [1].

一方で，各国に関連のあるトピックへの深い理解の為には，ニュース記事のテキストを構成する単語レベルでの分析も重要である．国家間の微細な結びつきは突発的で大きなイベントに現れるとも限らず，イベントレベルでの分析手法では取りこぼす恐れがある．そこで本研究では，「他国のアイデンティティを構成する上で重要な単語が，自国のニュース記事に頻繁に出現している状態」を国家間の意味的な関連とし，これを取得することを目的とする．著者らは以前，国ごとに特有の単語とカテゴリ，そして単語に対する重要度を含んだ単語辞書を利用し，各国の関連を取得する手法を提案した [2]. この手法では，Wikipedia-en[1)] の見出し語と記事カテゴリを単語辞書の作成に用いた．国家間の関連の取得に複数の国の英語のニュース記事を用いた．しかし，英語のニュース記事と単語のみを対象としていたため，分析対象が英語圏の視点に限定されるという課題があった．

そこで，本研究では複数の言語のニュース記事と国ごとに特有の単語を用いて，その国にとって重要な単語に着目しながら，国家間の関連を明らかにする．特に，その国家間の潜在的な関連を形成する上で眶となる単語を取得することを目指す．

2 提案手法

まず，本研究の提案手法の概略図を図 1 に示す．図 1 に示すように，提案手法は大きく分けて (a) から (g) の手順から成る． 最初に提案手法で利用するデータを手順 (a) と (b) で取得する．手順 (a) では各国の公用語で書かれた記事を収集し，手順 (b) では各国特有の単語 (以下、特有語と呼ぶ) の辞書を，他国家の公用語での呼称も含めて Wikidata[2)] を基に構築する．次に，各国のニュース記事に出現する特有語の検出と出現頻度の集計を手順 (c) で行う．このとき，単語重要度算出の前処理として，自国の記事における，自国および他国の特有語の出現頻度を分けて集計する．さらに，手順 (c) で集計した特有語の出現頻度を用いて，手順 (d) で単語に重要度を付与する．重要度を付与する具体的な方法は以下の通りである．

1. ニュース記事を Doc2Vec[3] でベクトル化し，対象とする国家数のクラスタにクラスタリング
2. クラスタの記事の中で一番多い出版された国で

1） https://en.wikipedia.org
2） https://www.wikidata.org

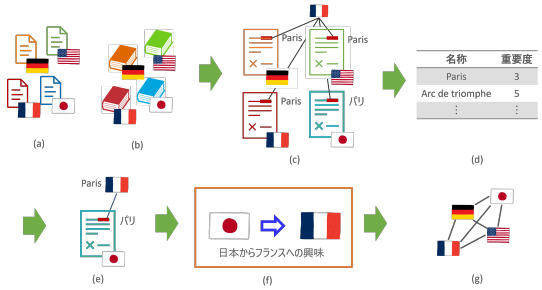

図 1　提案手法の概略図

クラスタ内の記事にラベル付け

3. 出版された国が A 国，国ラベルが A 国の記事を A 国に対する専門的な記事，それ以外の記事を一般的な記事とし，CrRv[4] を用いて重要度を付与

手順 (e) では手順 (f) で抽出したデータと手順 (d) で付与した単語重要度を用いて，以下の式で国 A から国 B へのどの程度関心を持っているかの関心スコア，$InterestScore_{A\to B}$ を計算する．このとき，A 国のニュース記事を D_A，B 国特有の単語辞書に含まれる単語集合を T_B, T_B に含まれる単語を t_B，t_B が D_A で出現した回数を $C_A(t_B)$ とし，CrRv による t_B の単語重要度を $W(t_B)$ とする．

さらに，記事量の差による影響を排除するため，$InterestScoreA \to B$ を D_A に含まれる単語の総数 N_{total} で割ることで正規化し，$InterestScore^{norm}_{A\to B}$ とする．

$$InterestScore_{A\to B} = \sum_{t\in T_B} (C_A(t_B) \times W(t_B)) \quad (1)$$

$$InterestScore^{norm}_{A\to B} = \frac{InterestScore_{A\to B}}{N_{total}} \quad (2)$$

最後に，手順 (g) にて全ての国家間の組み合わせに対して (e) および (f) の処理を適用し，全体の関係性と関係性に影響している単語を明らかにする．

3　実験

本節では，本研究で使用したデータと実験の方法について記す．この実験では，対象とする国家のニュース記事から提案手法の関心スコアを計算した後，スコアに影響している主な単語を確認する．

3.1　使用したデータ

データは主に各国の各言語で収集したニュース記事と，各国特有の単語である．

まず，本研究で使用したニュース記事の情報を以下に記す．なお，記事の収集には mediacloud[5] を利用した．

- 収集したニュース記事の対象国
  - アメリカ，日本，フランス，ドイツ，ロシア
- 収集したニュース記事の言語
  - 英語，日本語，フランス語，ドイツ語，ロシア語
- 収集したニュース記事の出版期間
  - 2024/08/01 から 2025/07/31
- 件数
  - 5 各国の記事で各 10,000 件（計 50,000 件）
  - 収集した期間内でなるべく均等になるようにサンプリング

次に，Wikidata から収集した単語の数，存在する各言語のラベルの数を表 1 に示す．また，表 2 に Wikidata から単語を収集した際の条件を示す．対象とした国は記事を収集した国と同様の 5 カ国であり，QID はそれぞれ Q30（アメリカ）Q17（日本），フランス（Q142），Q188（ドイツ），Q159（ロシア）である．

表 1　収集した単語の言語表記数

	en	ja	fr	de	ru
usa	222,978	44,458	113,224	99,013	55,420
jpn	63,304	66,331	25,286	11,978	30,877
fra	157,965	16,147	159,087	75,914	29,377
deu	101,160	11,915	86,710	98,684	20,068
rus	63,653	4,482	12,265	19,432	128,513

表 2　収集した単語の Wikidata プロパティとカテゴリ

カテゴリ	プロパティ	利用クラス/値
行政地区	P150 (行政区画)	-
人物	P27 (国籍) P106 (職業)	Q33999 (俳優) Q639669 (音楽家) Q36180 (作家) Q82955 (政治家)
自然地理	P17 (国)	Q8502 (山) Q8503 (河川) Q23397 (湖) Q46169 (国立公園)
企業	P31 (分類) P159 (本社所在地) P17 (国)	Q4830453 (企業)

3.2 実験方法

まず，記事から各国の特有語を抽出し出現頻度を取得する際には，Wikifier[6] を利用した．Wikifier は PageRank[7] を利用して，文脈をに基づいて文章中の単語と Wikipedia 概念をエンティティリンキングする手法である．Wikifier で各国の記事内の単語をエンティティリンキングした後，抽出した概念（単語）を収集した単語データを利用して各国の特有語であり，特定のカテゴリの単語にフィルタリングを行った．

次に，フィルタリングを行った特有語に CrRv で単語重要度を付与する．比較として TF-IDF[8] も利用して単語重要度を付与し，関心スコアを算出した．

3.3 実験結果

まず，図 2 と図 3 に単語重要度付与手法 CrRv ベースと TF-IDF の関心スコアのヒートマップを示す．図 2 では上から順にドイツ → アメリカ，ロシア → アメリカ，フランス → ロシアが関心スコアが高く，図 3 では上から順にロシア → 日本, ロシア → フランス, ロシア → ドイツが関心スコアが高いことがわかる．これら 6 つの関係の関心スコアに影響を与えた上位 5 単語をそれぞれ表 3 と表 4 に示す．また，単語単体で計算した国 A から国 B への関心スコアを，全体の関心スコアに占める割合 (寄与度) として表に示す．二つの表のどちらも地名や国名がほとんどを占めており，表 3 と表 4 の二つで寄与度の分布が大きく異なっていることが確認できる．

4 考察

この章では，実験結果を踏まえた考察を示す．まず，図 2 と図 3 では，関心スコアの高い関係が大きく異なっている． 表 3 と表 4 の詳細を見ると，CrRv 上位の単語，特に 1 位の単語の寄与度が大きく関心スコアの大部分を占めるのに対し，TF-IDF では寄与度が低く分散している． これは CrRv ではその国特有の専門的な単語に大きな重みを付与するために，特定の単語の寄与度が大きくなるのに対し，TF-IDF では CrRv よりも幅広い一般的な単語が関心スコアに影響を与えているからだと考えられる．

次に，表 3 に挙がっている関心スコアに大きく影響している単語を見ていくと，全体的には国名や地名，各国の大統領と首相の Donald Trump と Vladimir

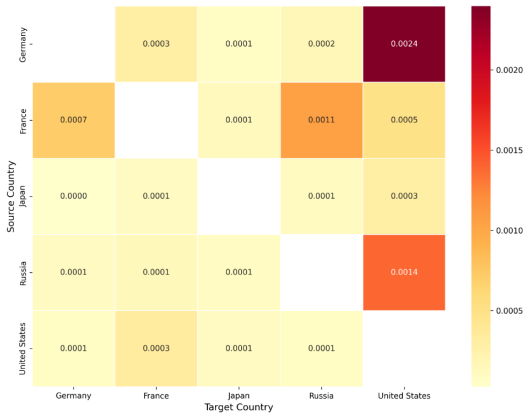

図 2 CrRv での関心スコアのヒートマップ

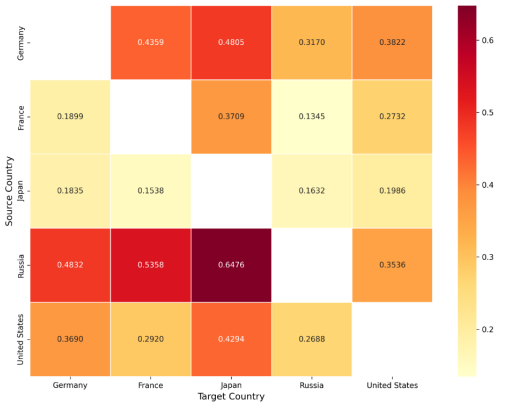

図 3 TF-IDF での関心スコアのヒートマップ

表 3 CrRv ベースにて影響を与えた上位 5 単語

国ペア	順位	単語 (概念)	寄与度 (%)
DEU ↓ USA	1	Washington，D.C.	86.18
	2	United States	9.90
	3	Donald Trump	1.28
	4	Michael Jackson	0.44
	5	Thomas Mann	0.22
RUS ↓ USA	1	United States	73.50
	2	Washington，D.C.	16.30
	3	Donald Trump	4.05
	4	Metro-Goldwyn-Mayer	2.48
	5	Michael Jackson	0.72
FRA ↓ RUS	1	Russia	97.94
	2	Saint Petersburg	0.65
	3	Vladimir Putin	0.63
	4	Moscow	0.42
	5	Gérard Depardieu	0.20

表 4 TF-IDF ベースにて影響を与えた上位 5 単語

国ペア	順位	単語 (概念)	寄与度 (%)
RUS ↓ JPN	1	Japan	0.39
	2	Tokyo	0.09
	3	Iga (伊賀)	0.02
	4	Shigeru Ishiba	0.01
	5	Osaka	0.01
RUS ↓ FRA	1	France	4.00
	2	Roman Polanski	1.36
	3	Paris	1.32
	4	Emmanuel Macron	0.84
	5	Marine Le Pen	0.08
RUS ↓ DEU	1	Germany	1.88
	2	Berlin	0.91
	3	Olaf Scholz	0.50
	4	Friedrich Merz	0.50
	5	Munich	0.39

Putin があり，政治的な関心が高いことがわかる．さらに個別の関係を見ると，ドイツからアメリカでは Thomas Mann （トーマス・マン）という第一次世界大戦前後に活躍し，ナチス政権時代にドイツからアメリカに亡命した作家が挙がっている．これは分析の対象としたニュース記事が 2024 年から 2025 年であったことと，トーマス・マンが 2025 年に生誕 150 周年であったことが重なり，これが記事内で言及されていたからだと考えられる[3]．また，ロシアからアメリカでは Metro-Goldwyn-Mayer (MGM) というカリフォルニア州に所在する映画製作会社があり，MGM は 2021 年に Amazon に買収され Amazon MGM Studio と改名された後，2025 年初頭には映画シリーズ「007」のクリエイティブコントロール権を得ている[4]．また，フランスからロシアでは Gérard Depardieu というフランスの俳優だがロシア国籍を取得している人物が確認できる．

次に，表 4 の単語でも全体的に国名と地名が大きな割合を占めているが，ロシアから日本では Iga（伊賀）や Osaka（大阪），ロシアからドイツでは Munich のように首都だけでない地名が確認でき，CrRv とは違った側面の関心が得られていると考えられる．

これらのことから，C 本論文で示した手法では政治的関心や国際的な大きなニュースだけでなく，国家間の文化的，経済的な関心を得ることができていると考えられる．

3） https://mann2025.de/about/

4） https://www.aboutamazon.com/news/company-news/amazon-mgm-studios-james-bond

5 おわりに

本研究では，多言語のニュース記事に対して，記事に出現する各国家特有の単語に重要度を付与したものを利用することで，国家間の関連性を取得することを目的とした．

多言語のニュース記事を分析する手法として記事の中の要素を Wikidata にリンクさせ，ニュース記事に含まれる各国特有の単語に出現頻度の差を利用して単語重要度を付与し，それを利用することで国から国への関心を取得した．また，その関心に影響している有意な単語と単語がどの程度影響しているかを取得することができた．

将来の展望として，トピック単位での分析との併用や，時系列での分析が挙げられる．

参考文献

[1] Xi Chen, Scott A. Hale, David Jurgens, Mattia Samory, Ethan Zuckerman, and Przemyslaw A. Grabowicz. Global news synchrony and diversity during the start of the covid-19 pandemic. In **Proceedings of the ACM Web Conference 2024**, pp. 2639–2650, 2024.

[2] 葛野航希, 横井健. 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の検討. 情報処理学会全国大会講演論文集, 第 86 回, p. 921–922, 2024.

[3] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In **Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32**, ICML'14, p. II–1188–II–1196. JMLR.org, 2014.

[4] 滝川真弘, 山名早人. 特定分野における単語重要度計算手法の提案と短文からの著者専門性推定への適応. 情報処理学会研究報告 (IPSJ SIG Technical Report), Vol. 2017-NL-233, No. 15, pp. 1–6, 2017.

[5] Hal Roberts, Rahul Bhargava, Linas Valiukas, Diara Jen, Momin M. Malik, Catherine S. Bishop, Emily B. Ndulue, Aashka Dave, Justin Clark, Bruce Etling, Rob Faris, Akshat Shah, Jen Rubinovitz, Alexis Hope, Catherine D'Ignazio, Fernando Bermejo, Yochai Benkler, and Ethan Zuckerman. Media cloud: Massive open source collection of global news on the open web. In **Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)**, Vol. 15, pp. 1034–1045, 2021.

[6] Janez Brank, Gregor Leban, and Marko Grobelnik. Annotating documents with relevant wikipedia concepts. In **Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD)**, 2017.

[7] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[8] Gerard Salton, Edward A. Fox, and Harry Wu. Extended

boolean information retrieval. **Communications of the ACM**, Vol. 26, No. 11, pp. 1022–1036, 1983.