# ニューラル分類器の予測の解釈に基づく集団に特徴的なテキスト表現の抽出: アメリカ人を例に

渡邉 幸暉　村脇 有吾　黒橋 禎夫
京都大学大学院情報学研究科
{k-watanabe, murawaki, kuro}@nlp.ist.i.kyoto-u.ac.jp

## 概要

本研究では、ある集団に特徴的なテキスト表現をその他の集団との比較により抽出する手法を提案する。提案手法は、2 種類のテキストを識別する分類器を訓練したうえで、ニューラルネットの説明手法に基づいて、分類器の予測に貢献する入力テキスト中の表現を特定する。具体例としてアメリカの文化や歴史を背景とした表現に取り組み、データセットの構築や評価実験を行う。

## 1　はじめに

人々の言語使用は、国や社会階層といった自身が属する集団の文化を反映していると考えられる。その性質を解明することで、例えば異文化コミュニケーションの支援といった応用が期待できる。

この研究課題への有力な取り組みは、異なる集団が産出したテキスト同士の比較を通じて差異を解明するというもので、対照研究あるいは比較研究とよばれる [1]。こうした対照研究は、主に社会言語学を中心としたいわゆる文系の研究者によって進められてきた。例えば、ウェブ、新聞記事、聖書といった様々な言語資料を比較することで、異なる集団が用いるテキスト表現の違いを明らかにできる [2]。

従来研究のうち人手による事例分析は、客観性を担保するのが難しい。また、対象テキストが数十件から数百件程度にとどまるなど、規模にも限界がある [3]。コーパスに基づく分析は客観性の問題を改善するが、人手によるアノテーションを必要とする場合はやはり規模に限界がある。自動分析を行う場合も、単語のような断片的で言語的に表層的な手がかりしか扱えないという課題が残る。

本研究ではこれらの対照研究の課題を解決するために、説明可能な AI を応用した手法を提案する。

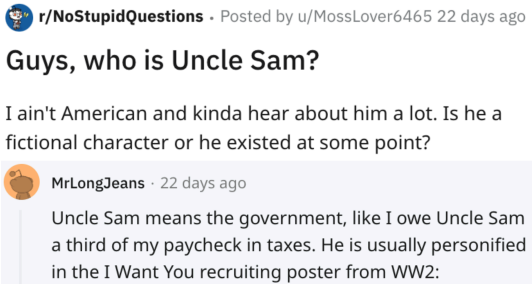


**図 1**　Reddit でのやりとりの一部
出典 : https://www.reddit.com/r/NoStupidQuestions/comments/z7h6fs/guys_who_is_uncle_sam/

提案手法はデータ駆動型の取り組みにより客観性と規模性を確保するとともに、近年のニューラルネットの強力な文脈処理能力を活用することで、従来の自動分析では扱うのが難しかったような長いテキスト表現を探索的に抽出する。ニューラルネットは高い性能の代償としてその振る舞いがブラックボックス化することが問題視されているが [4]、その解決策として説明可能な AI の研究が盛んに進められている [5][6]。本研究のキーアイデアは、その成果を応用し、ニューラルネットが異なる集団による 2 種類のテキストをどのような手がかりを用いて識別したかを解釈することを通じて、テキスト間の差異の原因を表現レベルにまで絞り込むことである。

提案手法の具体的な適用事例としてアメリカ人に着目する。アメリカは第二次世界大戦以来超大国とされ [7]、特に冷戦後は唯一の超大国であるとされている [8]。国際社会におけるアメリカの存在感は著しく、文化、経済、科学技術などの分野で他の国々がアメリカを追随する形で世界が発展している。

こうした背景のもと、アメリカ国内のコンセンサスが世界のコンセンサスであるかのような言動が見られることがある。具体的には、アメリカ以外の国の人がいる場でも、アメリカ人にしか理解できないような表現が使われてしまう。図 1 は、英語圏のオ

ンライン掲示板 Reddit で、アメリカ人ではないという投稿者がよく耳にするという “Uncle Sam” という人物について尋ねている様子を示す。この例では、“Uncle Sam” は特定の人物を指す言葉ではなく、アメリカ合衆国政府を意味する熟語であることが説明されている。

このようにアメリカの文化や歴史を背景とする表現を本研究ではアメリカニズムとよぶ (表 A.1)。アメリカニズムを抽出して提示することにより、アメリカ以外の国の人に対して意図せずアメリカニズムを使用してしまう事態を回避し、より円滑なコミュニケーションが可能になることが期待される。

本研究では、アメリカ人が書いた文と他の英語圏の国の人が書いた文を比較することで、アメリカニズムを抽出する。クローリングによるデータセット構築、Reddit が SNS であることを利用した対象の効果的な絞り込み、BERT に基づく強力なニューラル分類器の訓練を行ったうえで、説明手法の適用により表現の抽出を行う。抽出例を人手で分析したところ、実際にアメリカニズムが抽出できていることが確認できた。

## 2 関連研究

提案手法の基本的なアイデアは Harust ら [9] に由来する。Harust らは英語表現のなかには母語話者に特徴的なものがあると仮定し、英語母語話者によるテキストと第二言語話者によるテキストを比較することでそのような表現を抽出した。まず、Reddit の投稿から作成した話者国籍別の英文データセット [10] を利用して、入力されたテキストが英語母語話者のものか第二言語話者のものかを識別する分類器を作成した。この分類器に対して、説明手法の 1 つである contextual decomposition [11] を適用し、分類器の出力スコアに対する入力テキスト中の各表現の貢献を近似的に求めた。このスコアを当該フレーズの母語話者らしさとみなし、スコアの高いフレーズを母語話者に特徴的な英語表現の候補としている。この分類器の訓練には complementary-label learning [12] を応用した手法が使われており、通常の 2 つのラベルに加え、「どちらとも言えない」という特殊ラベルが導入されている。

Harust らは言語的な興味から英語母語話者と第二言語話者を比較したのに対し、本研究は文化的な観点からアメリカ人とその他の英語圏の人を対照する。また、Harust らは分類器を LSTM を用いて構築したのに対して、本研究はより強力な BERT を用いるなど、様々な点で手法に改良を加えている。

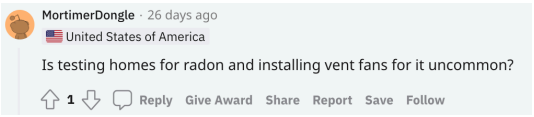


**図 2** Reddit で国籍を表明するタグが使用されている例
投稿者がアメリカ国籍であることが表明されている
出典 : https://www.reddit.com/r/AskEurope/comments/z81amu/comment/iycx57x/

**表 1** 構築したデータセットの国籍の内訳

| 国 | 投稿数 (単位: 百万) |
|---|---|
| アメリカ (国名のタグで特定) | 38 |
| アメリカ (州名・地名のタグで特定) | 14 |
| イギリス | 37 |
| カナダ | 8 |
| オーストラリア | 4 |

## 3 データセットの構築と絞り込み

### 3.1 データセットの構築

Harust ら [9] は話者国籍別の英語テキストとして Reddit に由来するデータセット [10] を用いた。このデータセットは、Reddit 内の一部のコミュニティで投稿者が自らの国籍をタグ (flair、図 2) で表明していることを国籍判別の手がかりとして利用している。原著者らの目的は、母語識別タスク (例えば英文の書き手がドイツ人であることを判別するなど) だが、アメリカやその他の英語圏の国が収録されており、本研究の目的にも利用できる。

このデータセットには投稿本文とその国籍ラベルのみが収録されている。しかし、例えば、Reddit が SNS であり、投稿者同士が関係を持っていることを利用すれば、より高品質にアメリカニズムが抽出できるかもしれない。こうした動機に基づき、本研究では独自に Reddit のクローリングを行った。

本研究では、アメリカと、比較対象として英語圏のその他の 3 カ国を対象とした。先行研究 [10] と比較すると、国籍判別の手がかりをを拡大している。投稿者が用いるタグは自由記述だが、先行研究は国名のみに着目していた。これに対し、本研究ではアメリカ国内の州名やその他の地名[1)] も利用した。こうしたタグの使用者はより内向きでアメリカニズムを表出する可能性が高いと期待できる。

データセット構築結果を表 1 に示す。州名や地名

1) Frorida や New York City など。New York City を略した NYC というものもあった。

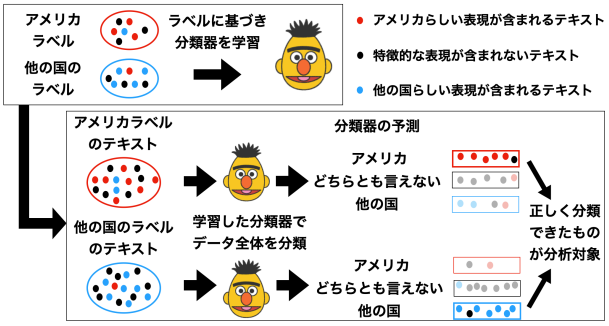


**図 3**　分析対象の絞り込みの手順

のタグも利用してことにより、アメリカ国籍の投稿者の投稿を多く収集できていることが分かる。

### 3.2　分析対象の絞り込み

一般に、ある集団の話者が産出したテキストに必ずその集団に特徴的な表現が出現するとは限らない。この問題に対処するために、Harust ら [9] は、母語話者と第二言語話者という 2 つのラベルに加えて、「どちらとも言えない」という特殊ラベルを導入し、特徴的な表現が出現しないテキストを吸収させていた。この目的のために complementary-label learning という機械学習手法を応用していた。

Harust ら [9] はこの 3 値分類器に説明手法を直接適用していたが、本研究ではまず 3 値分類器を使って分析対象となるデータを絞り込んだうえで、改めて 2 値分類器を訓練する (図 3)。処理段階を段階的に行うことで、例えば次節で示すように、Reddit の特性を利用したデータの質の改善が容易になる。

### 3.3　投稿者同士の関係の利用

Reddit の SNS としての特性は、データの質の改善に貢献し得る。例えば、アメリカ人同士での交流が中心の投稿者は、他の国の人との交流が多いアメリカ人よりもアメリカニズムを表出する可能性が高い。そこで、投稿者が自分と異なる国籍の投稿者に対する返信の割合を手がかりとして利用する。具体的には、投稿者に対して国際性スコアを付与する。

$$\text{投稿者の国際性スコア} = \frac{\text{その投稿者が自身と異なる国籍の人に返信した数}}{\text{返信先のうち、タグによって国籍が分かっている数}}$$

この国際性スコアは、絞り込み用の分類器に追加の特徴量として与える。

実験では、表 1 のアメリカと、アメリカ以外の国をまとめたその他の国の識別を行った。分類器としては BERT [13] を用い、各ラベル 900 万文ずつを

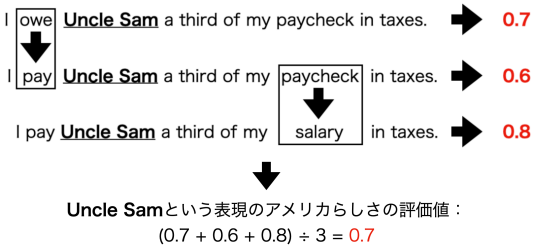


**図 4**　sampling and occlusion algorithm の実行例
四角で囲われた部分がランダムに変更された文中の語を表す。
ここでは簡単のため 3 文に対する出力の平均としている。

使って訓練を行った。結果を表 A.2 に示す。国際性スコアを付加することで、「どちらとも言えない」に分類された割合が減った[2)]にも関わらず、正しく分類できた割合が増えている。したがって、国際性スコアは絞り込みの精度向上に貢献していると判断できる。

図 1 や表 A.1 で示したアメリカニズムの具体例について、絞り込み前後での登場頻度の変化を表 A.3 に示す。特に登場回数が数千程度の例については、絞り込みにより、アメリカラベルのテキストにより偏って出現するようになったことが確認できる。

## 4　分類器の説明手法を使った各表現のアメリカらしさの評価

### 4.1　分類器の説明手法の利用

本研究で用いる分類器の説明手法は、分類器の出力に対して、入力テキストのどの部分が強く貢献したかを明らかにする。本研究では、絞り込み後のデータを用いて、アメリカとその他の国を識別する 2 値分類器を訓練している。したがって、分類器がアメリカラベルを予測した際、その予測に強く貢献した入力文中の表現が、アメリカらしい表現、すなわちアメリカニズムであるとみなせる。

テキストの分類器の説明手法は複数提案されているが、本研究では sampling and occlusion (SOC) algorithm [14] を利用する。ある表現を分析対象としたとき、原文中のその表現をパディングトークンで置き換えて再分類を行うと、もとの出力スコアからの差分をその表現の貢献と見なせる。ただし、このスコアの差分はその表現の前後の文脈によって変化する。SOC は、この文脈依存性を取り除くために、文脈中の語をランダムに変更した場合を考慮して、当該表現の評価値を求める (図 4)。実験では、1 つ

2)　「どちらとも言えない」に分類された割合はハイパーパラメータによって調整できる。

の入力に対して 20 文のサンプリングを行った。

### 4.2 分析対象となる表現の取り出し

分類器の説明手法を利用するためには、分析対象となる表現を指定する必要がある。Harust らは入力文中ののすべての 5-gram までを分析の対象としていたが、これでは明らかに意味をなさない表現が対象に含まれる、長さが 6 以上の表現は分析の対象にできないなどの問題点がある。

そこで、本研究では文に対する句構造解析を適用し、その結果を用いて表現を取り出す。具体的には、句構造解析によって得られた構文木の各頂点について、その子孫となる単語列を一つの表現として利用する。Harust らの手法と比べて、明らかに意味をなさない表現が含まれない、長さが 6 以上のものも取り出せるなどの利点がある[3)]。

### 4.3 既知のアメリカニズムを用いた分析

アメリカニズムの抽出は探索的なタスクであり、定量的な評価自体が挑戦的な課題である。そのため、まずは簡易的な評価として、既知のアメリカニズムを含む文に対する提案手法の振る舞いを調査した。

説明手法を適用する 2 値分類器としては BERT を用い、アメリカとその他の国の各ラベル 900 万文を用いて訓練した。注目する表現に対して、そのアメリカラベルへの貢献度から他の国のラベルへの貢献度を引いたものを評価値として用いる。この評価値が正であればアメリカらしい表現、負であれば他の国らしい表現であると評価されたことになる。

結果を表 A.4 に示す。上の 3 例は図 1 や表 A.1 で見たアメリカニズムの例、一番下は他の国のラベルに含まれるイギリスに特徴的な表現の例である。実際の評価値はアメリカニズムの例に対して正、イギリスに特徴的な表現で負となっていることが確認できる。

### 4.4 抽出結果の人手評価

次に、提案手法が抽出した表現を人手で評価した。3.2 節で説明したデータセットのうち 250 万文に対して、4.2 節で述べた手順で分析対象表現を列挙し、各表現に対して説明手法を適用して評価値を求めた。各表現の頻度を集計し、10 回以上出現する表現を対象としたところ、約 127,800 個の表現を得た。このうち頻度上位 1%から 100 個の表現をランダムに選んで人手で評価した。長い表現のうちどの部分を抽出すべきかは判断が難しいため、アメリカらしい表現が含まれていれば正解と判定した。なお、本論文の著者は全員が日本に住む日本語話者であるため、見逃したアメリカニズムがある可能性は否定できない。

調査の結果、100 個中 51 個がアメリカらしい表現であった。このうち、アメリカ国内の固有名詞が 2 個 (Buick や Grunhub)、アメリカ英語の綴り (イギリス英語の favour や honour に対する favor や honor など) を含む表現やアメリカ的な言い回し (home improvement store や、イギリス英語の grocery shop に対する grocery store) が 35 個であった。小切手の意味の check にはアメリカ英語であり、イギリス英語では cheque と綴るが、check という単語自体はイギリス英語でも用いられる。提案手法は writing a check という、check の語義が特定できる形の表現を抽出しており、ニューラルネットの高い文脈処理能力がうかがえる。51 個のうち残り 14 個が、アメリカの文化や歴史を色濃く反映した、狭義のアメリカニズムと判断できるものであった (表 A.5)。この結果は、探索的用途において提案手法がアメリカニズムの効率的発見に利用可能であることを示唆する。

正解と判定しなかった 49 個のうち 7 個は college を含む表現だった。college とよばれる機関自体は英語圏に広く存在するが、Reddit 上ではアメリカの機関の存在感が大きく、他の国での用法を頻度上圧倒した可能性がある。

## 5 おわりに

本研究では、ある集団に特徴的なテキスト表現をテキストの分類器の説明手法を活用して抽出する手法を提案し、その具体な適用事例として、アメリカ人に特徴的な表現であるアメリカニズムを抽出した。分析対象をアメリカに特徴的な表現が含まれるものに絞り込むことで、アメリカニズムの出現頻度に大きな偏りが発生して分析がしやすくなることを示し、実際の説明手法の適用で良好な結果が得られることを確認した。また、提案手法を実行した結果、評価値が高かったものからアメリカニズムを発見することに成功した。今後は、別の対象研究への提案手法の適用などについても検討したい。

3) 例えば “I owe Uncle Sam a third of my paycheck in taxes.” という文の場合、5-gram までで表現を取り出した際の候補は 45 個になるのに対し、句構造解析を用いると 19 個となる。

# 参考文献


[1] 井上優. 対照研究について考えておくべきこと. 一橋日本語教育研究, Vol. 3, pp. 1–12, 2015.

[2] 河正一. 社会言語学的調査の状況 ―言語行動に関する日韓対照研究を中心に―. 計量国語学, Vol. 31, No. 8, pp. 572–588, 2019.

[3] 佐々木倫子. 言語の対照研究と言語教育. 日本語科学, Vol. 3, pp. 127–134, 1998.

[4] Yavar Bathaee. The artificial intelligence black box and the failure of intent and causation. **Harvard Journal of Law & Technology**, Vol. 31, No. 2, pp. 890–938, 2018.

[5] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. **ACM Comput. Surv.**, Vol. 51, No. 5, pp. 1–42, 2018.

[6] Venessa Buhrmester, David Münch, and Michael Arens. Analysis of explainers of black box deep neural networks for computer vision: A survey. **Mahine Learning and knowledge extraction 2021**, pp. 966–989, 2021.

[7] William T. R. Fox. The Super-Powers; The United States, Britain, and the Soviet Union—Their Responsibility for Peace. New York: Harcourt, Brace and Company., 1944.

[8] Ian Bremmer. These are the 5 reasons why the U.S. remains the world's only superpower, 2015. https://time.com/3899972/us-superpower-status-military/.

[9] Oleksandr Harust, Yugo Murawaki, and Sadao Kurohashi. Native-like expression identification by contrasting native and proficient second language speakers. In **Proceedings of the 28th International Conference on Computational Linguistics**, 2020.

[10] Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. Native language cognate effects on second language lexical choice. **Transactions of the Association for Computational Linguistics**, Vol. 6, pp. 329–342, 2018.

[11] W. James Murdoch, Peter J. Liu, and Bin Yu. Beyond word importance: Contextual decomposition to extract interactions from LSTMs. In **International Conference on Learning Representations**, 2018.

[12] Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. In **Advances in Neural Information Processing Systems**, 2017.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, 2019.

[14] Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In **International Conference on Learning Representations**, 2020.

# A　参考情報

**表 A.1**　アメリカニズムの例

| アメリカニズム | 由来 | 意味 |
|---|---|---|
| Frisco | 不明 | サンフランシスコ市の異名 |
| How the turn tables. | アメリカのあるホームドラマで登場したセリフ | 形勢が逆転する様 |
| Kanaka | ハワイ語で「人」 | ハワイ人 |
| Rip Van Winkle | アメリカの童話の登場人物 | 長期間保存しておく |

**表 A.2**　絞り込み分類器の学習結果

| | 「どちらとも言えない」に分類された割合 | 「どちらとも言えない」に分類されなかったもののうち正しく分類できた割合 |
|---|---|---|
| テキストのみ | 70.5% | 86.1% |
| テキスト + 投稿者の国際性スコア | 69.7% | 87.0% |

**表 A.3**　データの絞り込みによる、データ中のアメリカニズムの登場回数の変化 (括弧内は残存率)

| | アメリカ | | その他の国 | |
|---|---|---|---|---|
| | 絞り込み前 | 絞り込み後 | 絞り込み前 | 絞り込み後 |
| 全データ数 (単位：百万文) | 105 | 17 (16%) | 97 | 28 (29%) |
| Uncle Sam | 2,101 | 853 (41%) | 768 | 107 (14%) |
| Frisco | 1,109 | 801 (72%) | 115 | 24 (21%) |
| How the turn tables. | 150 | 15 (10%) | 112 | 8(7%) |
| Kanaka | 12 | 5 (42%) | 4 | 1 (25%) |
| Rip Van Winkle | 84 | 12 (14%) | 25 | 3 (12%) |

**表 A.4**　Reddit 上の実際の投稿文に対する、説明手法の適用結果

| 投稿文 | 評価対象とする表現 | 評価値 |
|---|---|---|
| I owe Uncle Sam a third of my paycheck in taxes. | Uncle Sam | 0.72 |
| Any “must eats” in Frisco? | Frisco | 0.89 |
| Who else gonna Rip Van Winkle this coin for a few years? | Rip Van Winkle | 0.31 |
| Yes, that is London. | London | -0.84 |

**表 A.5**　分類器の説明手法による評価値が高かった表現から発見されたアメリカニズム

| 表現 | 意味 | 由来・文化 / 歴史的背景 |
|---|---|---|
| strict keto / people who do keto / be keto | ケトン食療法を厳守する人 / ケトン食療法をする人 / ケトン食療法をする | ケトン食療法 (keto) はアメリカの医師が開発した食事療法 |
| the same skin color / with skin color / on their skin color | 同じ肌の色 / 肌の色で / 彼らの肌の色で | アメリカ国内では肌の色 (人種) による扱いの違いが頻繁に問題になる |
| in the democratic party | アメリカ民主党において | the democratic party はアメリカ民主党を表す |
| have the electoral college | アメリカ大統領選挙人団を持つ | the electoral college はアメリカ大統領選挙人団という意味 |
| getting impeached | アメリカ連邦議会での弾劾を受ける | アメリカ連邦議会の弾劾制度 (impeachment) から |
| a violation of the first amendment | アメリカ合衆国憲法修正第 1 条に反している | the first amendment はアメリカ合衆国憲法修正第 1 条を指す |
| a high deductible plan | アメリカ国内の医療保険の形態の 1 つ | |
| legal in my state | 自分の住む州では合法 | アメリカ合衆国では州ごとに法律が定められていることによる |
| our state flag | 自分の州の州旗 | アメリカ合衆国では州ごとに州旗が定められていることによる |
| a dbag | 嫌な奴、ろくでなし | douchebag (医療器具) に由来するアメリカのスラング |