社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
2021-09-TL (2021-09)

[招待講演] 説明可能な人間としてのニューラルネットワーク

対照研究の新手法

村脇　有吾

京都大学 大学院情報学研究科　〒606-8501 京都市左京区吉田本町
E-mail: murawaki@i.kyoto-u.ac.jp

あらまし　本講演では、説明可能な AI (XAI) の分野で開発されている諸技術が計算人間科学に応用可能であることを主張する。XAI の文脈では、人工知能は人間とは異質な他者であり、ブラックボックスの中身を説明することで人間の信頼を勝ち取るべきものとみなされる。一方、計算人間科学においては、人間こそがブラックボックスであり、人工知能は人間の機能を近似する役割を果たす。したがって、説明手法は人間の説明に転用できるのではないだろうか。これが本講演の着想である。具体例として、ニューラルネットワークに基づく分類器が対照研究に応用できることを示す。まず 2 つの人間集団が書いたテキストを識別する分類器を訓練し、次に説明手法を適用することで分類器がどのように分類を行っているかを分析する。現在のニューラルネットワークの高い表現力を活かすと、文脈依存の語や長い表現を探索的に調査できることが大きな利点である。
キーワード　説明可能な AI、ニューラルネットワーク、分類器、対照研究

Neural Network as an Explainable Human

A New Approach to Contrastive Studies

Yugo MURAWAKI

Graduate School of Informatics, Kyoto University　Yoshida-honmachi, Sakyo-ku, Kyoto, 6060-8501, Japan
E-mail: murawaki@i.kyoto-u.ac.jp

Abstract　In this talk, I argue that techniques developed in the field of explainable AI (XAI) have potential applications in computational human sciences. In typical XAI scenarios, artificial intelligence is seen as a technological Other that is obliged to win human trust by explaining the black box. In computational human sciences, however, it is humans that are the black box, and artificial intelligence serves as an approximation of human functions. This observation motivates us to use explanation methods to explain humans. As a concrete example, I show that neural network-based classifiers can be applied to contrastive studies. We begin by training a classifier to discriminate texts written by two groups of humans and then apply an explanation method to analyze how it performs classification. A major advantage of this approach is that the high expressive power of modern neural networks allows us to investigate context-sensitive words and long expressions in an explorative manner.
Key words　explainable AI, neural networks, classifiers, contrastive studies

1. 人間こそがブラックボックス

人間が理解可能なように振る舞う機械学習システムに対する需要は機械学習分野では古くから認識されてきた [1] が、深層学習の進展にともない、説明可能な AI (XAI) という名前を与えられて脚光を浴びている [2]。機械学習の応用先として自動運転、医療、司法、雇用等に関わる重大な意思決定が射程に入ってきた結果、システムの透明性を確保し、人間の信頼を勝ち取ることが重要な課題として認識されるようになっている。こうした研究のなかから、解釈性と説明性が重要な概念として浮上してきた。両概念は区別なく用いられることも多いが、本発表では、[3] に従い以下のように切り分ける。解釈可能な機械学習は本質的に解釈しやすいモデル (例えば線形回帰) に注力するのに対し、説明可能な機械学習は人間には本質的に解釈困難なモデルに対して事後的に説明を与えることに取り組む。本発表で着目するのは後者である。

自然言語処理は機械学習の応用先の一つであり、XAI の興隆の影響を受けつつあるが、その動向には機械学習分野とのずれ

が見られる。講演者の見るところでは、機械学習分野で想定されるような重大な意思決定にテキストが用いられることが少ないからである。テキストはそこから誤りなく情報を抽出するのが難しいという点で信頼できない媒体である。その上、データ収集の仕組みを適切に設計すれば構造化データで代替しやすいという性質を持つ。このような理由から、自然言語処理における XAI [4] は、一般利用者に対する説明よりも、開発者自身の関心に答えることに向かいがちである。

このような傾向は自然言語処理の社会応用の未熟さを反映するものであるが、本発表では発想を変え、XAI の肯定的な転用方法を提案する。そもそも自然言語処理は工学的応用と科学的探究にまたがる分野であり、機械学習モデルによる (部分的) 再現を通じて人間の言語能力を解明することが目標の 1 つとなっている。XAI の一般的文脈では機械学習モデルは説明を通じて人間の信頼を勝ち取るべき異質な他者とみなされるのに対し、自然言語処理においては人間こそが解明されるべきブラックボックスである。この前提を踏まえて XAI の分野で開発された説明手法を見直すと、次のような発想に至る。機械学習モデルを人間の代理とみなし、そこから説明を得ることで、人間を説明することが可能ではないだろうか。また、この取り組みは自然言語処理に限定されるものではなく、計算人間科学一般に応用可能ではないだろうか。

2. 母語話者表現検出タスク

機械学習モデルを説明可能な人間とみなす取り組みの具体例として、講演者らによる研究事例 [5] を紹介する。講演者らはテキスト中の母語話者に特有の表現を検出するタスクを提案し、これを母語話者表現検出と名づけた。検出技術自体は言語非依存だが、対象言語として英語を想定する。

このタスクを提案した背景には、英語の熟練の第 2 言語 (L2) 話者であっても母語話者[1]の発話の理解に失敗するという観察がある [6], [7]。L2 話者同士が支障なく会話している場に母語話者が参入することで、たちまちコミュニケーションが破綻するというのである。特に [6] はデンマークでの事例に言及している。西ヨーロッパの人間はほぼ完璧な英語を話しているように日本語話者である講演者からは見えるが、それでも母語話者とのコミュニケーションに困難を覚えるというから驚きである。コミュニケーションを阻害する要因には発音もあるが、本発表では表現に着目する。[6] の例を借りると、ballpark figure は概算値を意味するアメリカ英語の熟語だが、L2 話者には理解しづらい。このような表現を母語話者表現とよぶことにする。

L2 話者が関わるコミュニケーションの失敗は、L2 話者の矯正により解消すべきであると一般に考えられている。英語の場合は巨大な語学教育産業が世界的に成立しており、商業的にもそれ以外の選択肢を考えづらい。この産業との結びつきを深めている自然言語処理でも事情は同じである。

(注1)：英語の国際的広がりは特に大英帝国の植民地支配の歴史を反映して複雑である。本発表では、母語話者英語として通常は矯正の対象とみなされない英米の標準英語を想定し、旧植民地において土着化した英語は考慮しない。

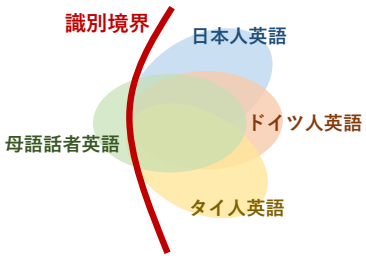

図 1　想定される母語話者英語と L2 英語の分布

しかし、講演者は、[6], [7] と同様に、適応すべきは母語話者であると提言する。熟練度が高く、L2 話者同士のコミュニケーションに支障がない場合であっても母語話者英語の理解が難しいという観測に基づくと、英語の分布は概略的には図 1 のようになっており、母語話者英語に特有の領域が存在すると推測できる。この図の中の多様な英語が重なり合う領域を少なくとも国際的なコミュニケーションの場における規範とすべきである。言い換えると、英米標準英語はもはや L2 話者が目指すべき規範ではなく、しばしば規範から逸脱する代物として捉え直すべきである。

この提言を実現するには、政治的課題と技術的課題を解決しなければならない。政治的には、数の上で母語話者を圧倒する L2 話者を糾合する必要があるが、本発表では深入りしない。技術的には、何が規範からの逸脱であるかが自明でないため、これを明らかにしなければならない。それがすなわち母語話者表現検出タスクである。国際的なコミュニケーションの場において ballpark figure のような母語話者表現を自動検出し、母語話者に矯正を促すことがこのタスクの主要な用途となる。

最後に、技術的課題をデータ駆動で解決するために、データについて触れておこう。具体的には、母語話者か L2 話者かという書き手の属性がメタ情報として付与された大量の英語テキストが必要となる。[5] では L2-Reddit コーパス [8] を用いた。Reddit は英語圏の掲示板的ウェブサイトであり、その一部では利用者が自身の所属国を申告している。[8] は所属国を使って母語話者、L2 話者を認定している。荒っぽい近似ではあるが、母語話者、L2 話者ともに数千万規模の文が手に入るため、大量の教師データを必要とすることで悪名高い深層学習を適用しやすいという強みがある。

3. ニューラル分類器を説明する

母語話者表現検出に XAI の説明手法を使うことが [5] の方法論的提案であるが、その前に他の可能性を検討してみよう。いま、文の集合があり、各文に書き手が母語話者か L2 話者かというメタ情報が付与されているとする。古典的には、母語話者集合、L2 話者集合のそれぞれに対して単語の頻度分布[2]を計算し、χ^2 検定や G 検定等を用いて差異に着目するという方法が採られてきた。あるいは、書き手の属性を目的変数、文の bag-of-words 表現を説明変数とする線形回帰モデルを作り、大

(注2)：発展編として、生の頻度分布を用いるのではなく、ベイズ統計の生成モデルを用いて分野等のずれを吸収する [9] という方向性も考えられる。

きな係数と紐づいた説明変数を調べるという手法も考えられる。

こうした手法の明らかな欠点は、単語タイプを手がかりとしているため、多義性や単語より長い表現を考慮できないことである。線形回帰モデルの説明変数に句を導入することも考えられるが、素朴に用いると組合せ爆発を起こすため、事前に句の候補を絞り込む必要がある。しかし、母語話者表現検出は探索的タスクであり、ballpark figure のような、事前に想定していなかった表現を発見することに意義がある。

線形回帰モデルは、現代的に見直すと、解釈可能な機械学習の一種ということになる。ここまでの議論で明らかなように、母語話者表現は複雑な言語現象であり、解釈可能なモデルでは太刀打ちできない。ball、park、figure という (ballpark が正書法上 1 語の扱いであることを度外視すれば) 何の変哲もない基本的な単語が組合せにより途端に L2 話者にとって難しくなるという現象が問題の核心である。

ここで、人間が母語話者表現をどう処理するか考えてみよう。人間であれば誰しも習熟可能とは限らないが、少なくとも母語話者と熟練の L2 話者の英語に長期間さらされた人間の多くは、「このテキストは母語話者っぽい」、「こちらは L2 話者に違いない」といった感覚を獲得するに至るであろう。注意すべきは、これは直感の一種であり、何を根拠にそのような推測を行ったかは当人にとっても自明ではなく、内省を必要とする。テキスト中のパターンと書き手の属性との連合を学習しているはずだが、その中身はまさにブラックボックスである。

この議論は人間を計算機に置き換えた場合も成り立つ。表現力は高いがブラックボックス的なモデルを一旦獲得したのちに、そのモデルに対して事後的に説明を与えるという方向性が有望だろう。そして、このブラックボックスモデルによる人間の代替をある程度の精度で実現してくれるのが、深層学習というブランド名とともに復活したニューラルネットワークである。具体的には、単語列で表現される文を入力とし、書き手の属性の予測を出力するニューラル分類器を設計し、教師データを用いて訓練すればよい。[5] では双方向 LSTM という古典的なアーキテクチャを用いて分類器を構築したが、近年は Transformer とよばれるより強力なアーキテクチャが普及している。

自然言語処理におけるニューラルネットワークは、単語を連続値ベクトルで表現することで汎化能力を得ただけでなく、単語同士の非線形な相互作用を捉える能力を持っている。したがって、ballpark は figure が後続することで特殊な語義を発火させるといった複雑な言語現象も、十分なデータがあれば捉えられる。

ニューラル分類器ができたら、次はその振る舞いを説明する。書き手の属性が母語話者であり、分類器も母語話者と予測した文は母語話者表現を含むことが期待される。したがって、そうした分類器の予測に強く貢献した入力系列中の部分列[3]を同定すれば良い。これは一般的な定式化であり、従来研究では感情分析モデルの分析を主な対象であった。映画批評を入力として

(注3)：説明手法の都合上は連続した部分列である必要はないが、簡単のために連続した列に対象を限定する。

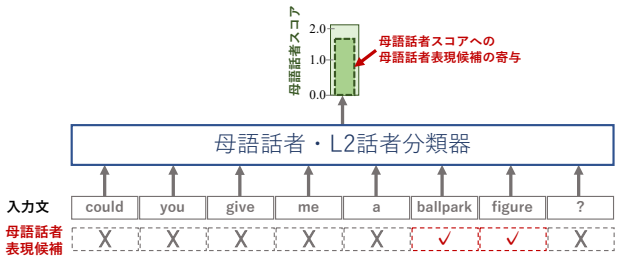

図 2　ニューラル分類器の説明に基づく母語話者表現検出

肯定か否定かを予測するモデルを作ったとき、ある予測を行うためにテキスト中のどの部分を手がかりとして使ったかを知りたいというのである。

この問題に対する解法の提案は多いが、[5] では文脈分解 (CD) とよばれる手法 [10] を採用した。いま母語話者表現候補として入力中のある部分列に着目したとき、CD は、図 2 のように、出力の母語話者スコアに対する当該部分列の寄与分を近似的に算出する。非線形操作を含むニューラルネットの計算は厳密には寄与分と非寄与分に分解できないため近似ではあるものの、CD は対抗手法とくらべて比較的人間の直感に沿った結果を返すことが報告されている。この手法を用いて、様々な母語話者表現候補に対して寄与分を計算し、寄与分が大きな候補を最終的に選び出す。候補列挙の方法としては、[5] は単純に n-gram の総当りを採用したが、CD の後継手法は階層クラスタリングを用いた候補の絞り込みを行っている。

L2-Reddit コーパスにこの手法を適用したところ、privy のように難易度の高い単語も検出されたが、例えば以下のように基本語からなる表現が目立った (下線部が検出された表現)。

> Good luck with it, but I think the location is going to be a hard sell.

当該表現は「人を説得して買ったり承諾させるのが難しいもの」というような意味であり、平易な単語しか使っていないはずだが、L2 話者には理解しづらい。

また、以下のように、検出範囲は人間の直感から外れているものの、文自体は母語話者らしい例も散見された。

> Its used for casual diving off New Jersey but is currently at quite a depth.

この文の後半部分は「海の深いところに沈んでいる」というような意味だが、やはり L2 話者には理解しづらく、自分で産出するとなるとなおさら困難である。定量評価を含む結果の詳細については [5] を参照されたい。

4. 実践的工夫

4.1 中立ラベルの導入

以上が基本アイデアとなるが、ニューラルネットワークに人間の代理をつとめさせるのはやはりそれほど簡単ではなく、いくつもの工夫が必要が欠かせない。最初の工夫が中立ラベルの導入である。

ここまでの議論では、書き手の属性である母語話者、L2 話者

という 2 種類のラベルのみを考えてきたが、図 1 で見た通り、両者には実際には大きな重複がある。つまり、2 種類のラベルのいずれかに分類するための手がかりを欠いた文がかなりの割合を占めており、分類器に 2 値分類を強いると、混乱した分類器が直感に反したスコアを返すようになりがちである。

この問題に対処するために、書き手の属性 (母語話者もしくは L2 話者) と文ラベルを分けて考え、後者に対して中立という第 3 のラベルを導入する。そして、補完ラベル [11] という概念を導入することで両者を紐づける。すなわち、書き手の属性の上での母語話者は文の補完ラベルでは L2 話者ではないに対応する。L2 話者ではないということは、文ラベルとしては母語話者もしくは中立である。同様に、書き手の属性の上での L2 話者は文の補完ラベルでは母語話者ではないに対応し、文ラベルとしては中立もしくは L2 話者である。

詳細は [5] に譲るが、分類器には 3 値の文ラベルに対応する要素数 3 のスコアベクトルを出力させ、このベクトルを補完ラベルの空間に射影するだけで、全体の枠組みに対する大きな変更なしに訓練が行える。元の 2 値分類では 60% 程度の分類精度であったが、中立ラベルを導入し、分類が難しい事例を中立ラベルに吸収させて無視することで、精度が 80% 程度まで向上した。書き手の属性自体が自己申告の所属国に基づく大雑把な近似であったことを考慮すれば、まずまずの精度である。

4.2 偽の手がかりとの戦い

ニューラル分類器は高い表現力を誇る一方で、まるで受験テクニックばかりを磨く学生のように、本質的とは言いがたい偽の手がかりを使って分類しがちであることが知られている。実際、自然言語処理における XAI は、偽の手がかりを分析することが主要な動機となっている。

母語話者表現検出においても、タスクの目的が言語的に特徴的な表現の検出である一方、実際に行っていることは書き手の属性の予測であり、両者にずれがあることは否めない。例えば、Bay Area のような米国の固有名詞が現れる文は母語話者 (米国) らしいと推測できるが、この表現自体は母語話者らしくはない。固有名詞のように前もって明らかな表現の場合は、前処理によって対処可能である。具体的には入力文に固有表現抽出を適用し、人名、地名等を特殊トークンで置き換えてしまえば、分類器はズルができなくなる。

前処理は強力な手段だが万能ではない。例えば、gun control and gun ownership はいかにもアメリカ的な話題であり、分類器は母語話者表現のように扱うが、その意味は L2 話者にとっても明らかであり、言語的に面白い表現ではない。このように普通名詞だけからなる表現は前処理では対処しにくい。

この問題への解法として講演者が期待しているのは分野敵対的訓練 [12] である。例えば、メタ情報として、書き手の属性以外にも、当該テキストのトピックが与えられているとする。gun control and gun ownership が登場する文はトピックとしてはアメリカ政治に分類されるといった具合である。このとき、入力文に対してトピックを予測する別のニューラル分類器を用意する。ただし、この分類器は書き手の属性の分類器との間で、入力から途中までのネットワークを共有するとする。この部分ネットワークを特徴抽出器とよぶ。分野敵対的訓練は、書き手の属性の分類器とトピック分類器を通常通りに訓練する一方で、トピック分類器が分類に失敗するように敵対的に特徴抽出器を訓練する。その結果、特徴抽出器が入力文から抽出する情報はトピックの違いに鈍感になり、特徴抽出器を共有する書き手の属性の分類器もその手がかりを使えなくなることが期待できる。[5] の発表時点では分野敵対的訓練からは芳しい結果を得られなかったものの、有望な手法として引き続き注目している。

5. 今後の展望

[5] は母語話者表現検出という個別具体的なタスクに話題を限定していたが、そこで示した方法論的提案は一般性を持つ。2 つの人間集団間の差異を探る研究は一般化すれば対照研究といえる。ニューラル分類器を説明するという手法は対照研究の他の課題に対しても適用可能だと講演者は考えている。XAI の説明手法全般に対象を広げ、テキスト媒体や分類問題といった制限も取り払えば、計算人間科学全般においてさらに広い適用先が見つかるだろう。

提案手法の大きな欠点は、タイプではなくトークンレベルの表現しか得られないことである。単語単位であれば数え上げるだけでデータの要約が可能であったのに対し、提案手法は単語同士の非線形な相互作用を捉える代償として、検出された表現の集約が難しくなっている。データ全体を俯瞰できるような手法の開発が急務であると講演者は認識している。

文　献

[1] A.A. Freitas, “Comprehensible classification models: A position paper,” SIGKDD Explorations Newsletter, vol.15, no.1, p.110, March 2014. https://doi.org/10.1145/2594473.2594475

[2] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, “XAI—explainable artificial intelligence,” Science Robotics, vol.4, no.37, pp.••–••, 2019. https://robotics.sciencemag.org/content/4/37/eaay7120

[3] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nature Machine Intelligence, vol.1, pp.206–215, May 2019.

[4] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, “A survey of the state of explainable AI for natural language processing,” Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pp.447–459, Association for Computational Linguistics, Suzhou, China, Dec. 2020. https://aclanthology.org/2020.aacl-main.46

[5] O. Harust, Y. Murawaki, and S. Kurohashi, “Native-like expression identification by contrasting native and proficient second language speakers,” Proceedings of the 28th International Conference on Computational Linguistics, pp.5843–5854, International Committee on Computational Linguistics, Barcelona, Spain (Online), Dec. 2020. https://aclanthology.org/2020.coling-main.514

[6] S. Hazel, “Why native English speakers fail to be understood in English and lose out in global business,” 2016. https://theconversation.com/why-native-english-speakers-fail-to-be-understood-in-english-and-lose-out-in-global-business-54436

[7] C. McCusker and R. Cohen, “Tower of babble: Nonnative speakers navigate the world of ’good’ and ’bad’ English,” 2021. https://www.npr.org/sections/goatsandsoda/2021/04/25/989765565/tower-of-babble-non-native-speakers-navigate-the-world-of-good-and-bad-english

[8] E. Rabinovich, Y. Tsvetkov, and S. Wintner, “Native language cognate effects on second language lexical choice,” Transactions of the Association for Computational Linguistics, vol.6, pp.329–342, 2018.

[9] A. Haghighi and L. Vanderwende, “Exploring content models for

multi-document summarization," Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp.362–370, Association for Computational Linguistics, Boulder, Colorado, June 2009. https://aclanthology.org/N09-1041

[10] W.J. Murdoch, P.J. Liu, and B. Yu, "Beyond word importance: Contextual decomposition to extract interactions from LSTMs," International Conference on Learning Representations, pp.••–••, 2018. https://openreview.net/forum?id=rkRwGg-0Z

[11] X. Yu, T. Liu, M. Gong, and D. Tao, "Learning with biased complementary labels," Proceedings of the European Conference on Computer Vision (ECCV), pp.68–83, 2018.

[12] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," Journal of Machine Learning Research, vol.17, no.1, pp.2096–2030, 2016.