# 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

京都大学
大学院情報学研究科
村脇 有吾

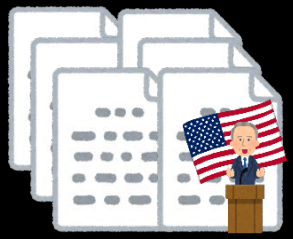

言語と思考研究会/MAPLL-TCP 2021 基調講演 @ online　　2021年9月18日

# 村脇 有吾 MURAWAKI Yūgo

- 表の仕事: 自然言語処理 (NLP) 全般
- 裏テーマ: 言語学の中でこれまでNLPと接点の少なかった分野に計算集約統計的手法を導入
- 最近の対外発表:
  - Umakoshi+ (2021). Japanese zero anaphora resolution can benefit from parallel texts through neural transfer learning. EMNLP Findings. (to appear)
  - Ueoka+ (2021). Frustratingly easy edit-based linguistic steganography with a masked language model. NAACL
  - Harust+ (2020). Native-like expression identification by contrasting native and proficient second language speakers. COLING
  - Murawaki (2020). Latent geographical factors for analyzing the evolution of dialects in contact. EMNLP

# 対照研究 + 説明可能なAI (XAI)

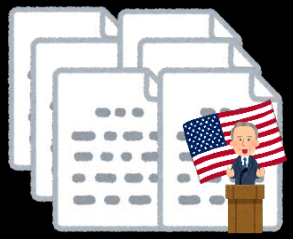
2つの集団間の差異をテキストを通じて明らかにしたい



VS

# 対照研究 + 説明可能なAI (XAI)

2つの集団間の差異をテキストを通じて明らかにしたい

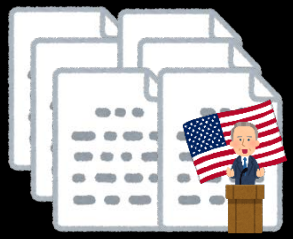


VS

- 人間がテキストを読んで感じたことを記述
- 深い洞察が可能
- 科学的観点からは、検証を欠いた仮説どまり

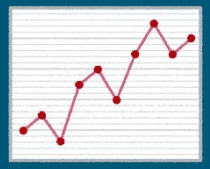
# 対照研究 + 説明可能なAI (XAI)

2つの集団間の差異をテキストを通じて明らかにしたい



VS

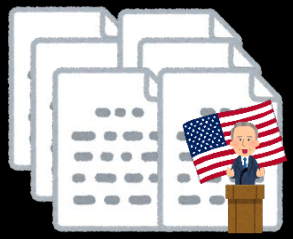
- 人間がテキストを読んで感じたことを記述
- 深い洞察が可能
- 科学的観点からは、検証を欠いた仮説どまり



- 統計的手法を駆使した計量的分析
- 表層的手がかり (e.g. 単語頻度) に終始
- 検証可能性を持つ科学的手法

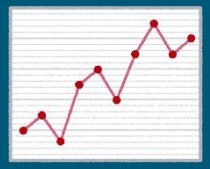
# 対照研究 + 説明可能なAI (XAI)

2つの集団間の差異をテキストを通じて明らかにしたい



VS

- 人間がテキストを読んで感じたことを記述
- 深い洞察が可能
- 科学的観点からは、検証を欠いた仮説どまり

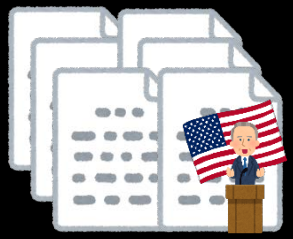


- 統計的手法を駆使した計量的分析
- 表層的手がかり (e.g. 単語頻度) に終始
- 検証可能性を持つ科学的手法

- ブラックボックス的なニューラル分類器 + 説明手法
- 従来よりも深い文脈を捉えられる
- 検証可能性を持つ科学的手法

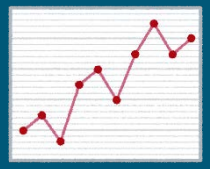
# 対照研究 + 説明可能なAI (XAI)

2つの集団間の差異をテキストを通じて明らかにしたい



VS

- 人間がテキストを読んで感じたことを記述
- 深い洞察が可能
- 科学的観点からは、検証を欠いた仮説どまり



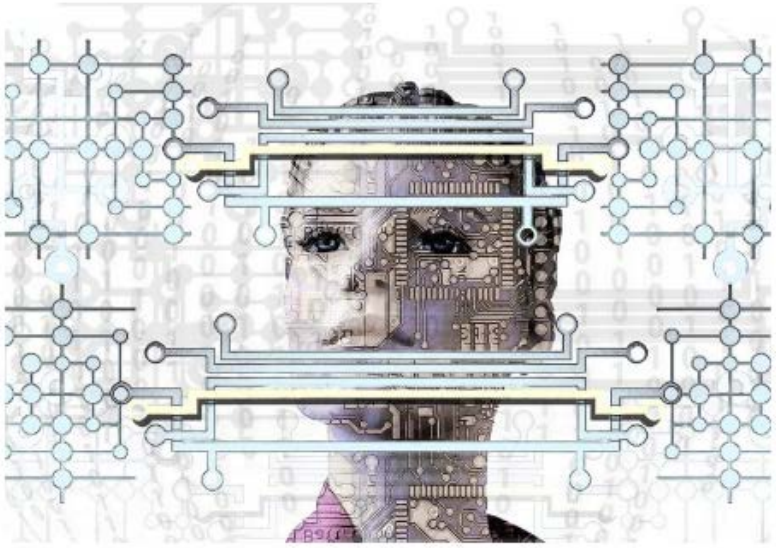
- 統計的手法を駆使した計量的分析
- 表層的手がかり (e.g. 単語頻度) に終始
- 検証可能性を持つ科学的手法

- ブラックボックス的なニューラル分類器 + 説明手法
- 従来よりも深い文脈を捉えられる
- 検証可能性を持つ科学的手法

# XAIの背景: AIの浸透と透明性への要求

## After Uber, Tesla incidents, can artificial intelligence be trusted?

**Posted by Andrew Careaga**
On April 9, 2018



In light of recent accidents with autonomous and semi-autonomous vehicles, will people put their trust in artificial intelligence? Missouri S&T researchers are digging for answers.

Given the choice of riding in an Uber driven by a human or a self-driving version, which would you choose?

https://news.mst.edu/2018/04/after-uber-tesla-incidents-can-artificial-intelligence-be-trusted/

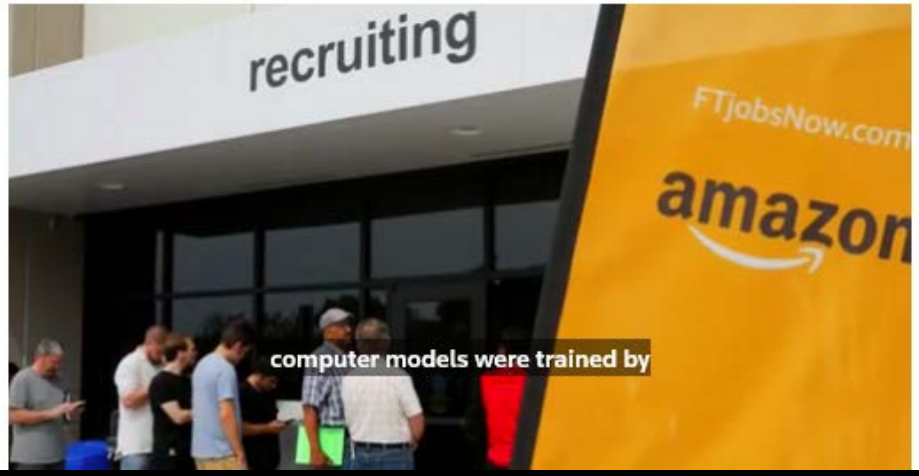
# XAIの背景: AIの浸透と透明性への要求

OCTOBER 11, 2018 / 8:04 AM / UPDATED 3 YEARS AGO

## Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.




https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

# XAIの背景: AIの浸透と透明性への要求

## AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

by **Karen Hao** January 21, 2019

IAN WALDIE/GETTY IMAGES

**AI might not seem to have a huge personal impact if your most frequent brush with** machine-learning algorithms is through Facebook's news feed or Google's search rankings. But at the Data for Black Lives conference last weekend, technologists, legal experts, and community activists snapped things into perspective with a discussion of America's criminal justice system. There, an algorithm can determine the trajectory of your life.

https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/

# 社会におけるAI: 異質な他者

- 内部で何をしているのかわからない不気味なブラックボックス
- にもかかわらず重大な意思決定を自動化しようとしている
- 人間の信頼性を勝ち取るための透明性の確保が急務

# NLP (計算言語学) におけるAI: 人間の代理 (proxy)

- 人間の言語能力のAIによる再現を通じた解明が研究目標 (の一つ)
- 人間こそがブラックボックス
- AIの方がはるかに内部の振る舞いを説明しやすい

- 人間がテキストを読んで感じたことを記述
- 深い洞察が可能
- 科学的観点からは、検証を欠いた仮説どまり

# XAIにおける解釈性と説明性

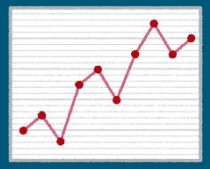
[Rudin+, 2019] inter alia

- 解釈可能な (interpretable) 機械学習
  - 本質的に人間に解釈しやすいモデルを利用
    - 線形回帰、決定木等
- 説明可能な (explainable) 機械学習
  - 本質的に人間には解釈困難なモデルを事後的に説明

# XAIにおける解釈性と説明性

[Rudin+, 2019] inter alia

- 解釈可能な (interpretable) 機械学習
  - 本質的に人間に解釈しやすいモデルを利用
    - 線形回帰、決定木等
- 説明可能な (explainable) 機械学習
  - 本質的に人間には解釈困難なモデルを事後的に説明



- 統計的手法を駆使した計量的分析
- 表層的手がかり (e.g. 単語頻度) に終始
- 検証可能性を持つ科学的手法

解釈可能な機械学習

- ブラックボックス的なニューラル分類器 + 説明手法
- 従来よりも深い文脈を捉えられる
- 検証可能性を持つ科学的手法

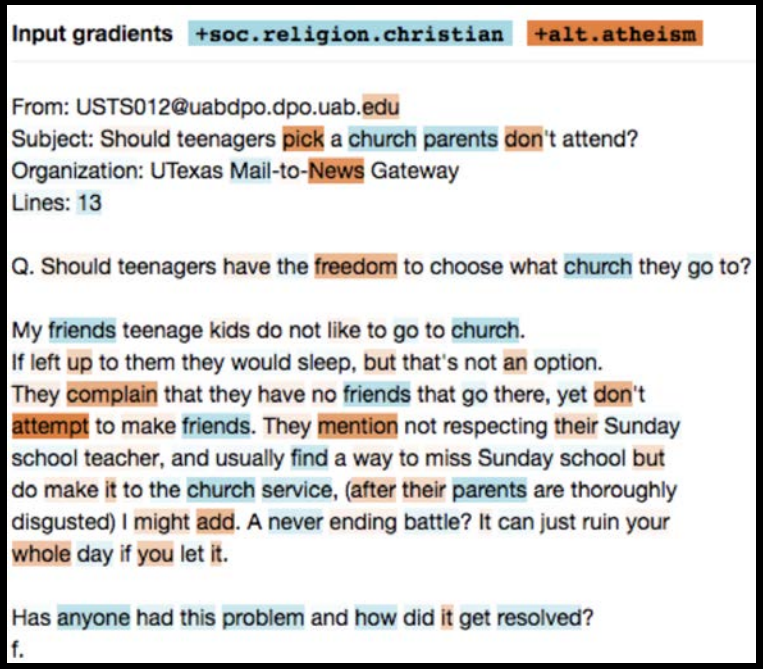
説明可能な機械学習

# NLPにおける説明手法 1/3

- 機械学習分野の後を追って、2017年頃から説明手法の研究が流行
- 主な用途: ニューラルモデルの振る舞いを開発者が分析
  - 人間の信頼性の確保がXAIの主目的だったはず
  - そもそもテキストは重大な意思決定にはあまり用いられない

# NLPにおける説明手法 2/3

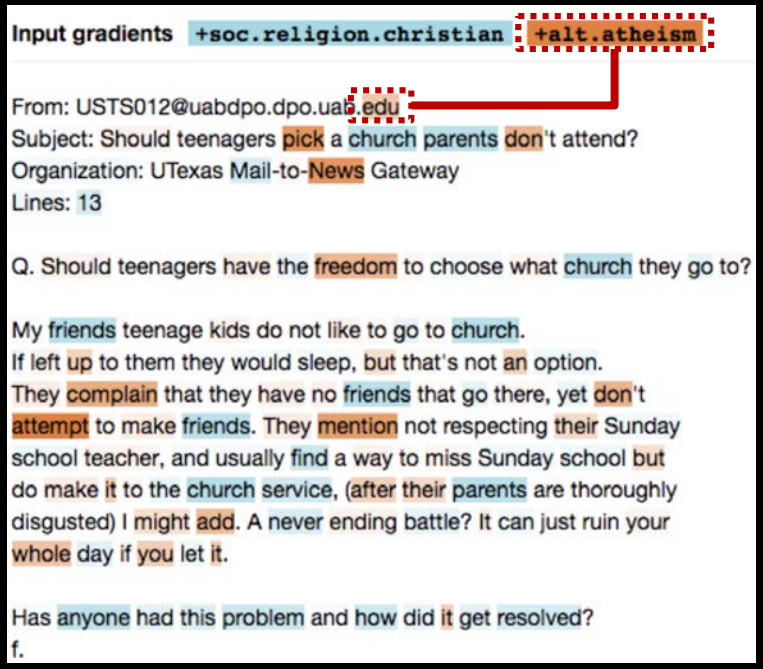
個別事例内の
入出力間の対応の
局所化に基づく説明




[Danielevsky+, 2021]

# NLPにおける説明手法 2/3

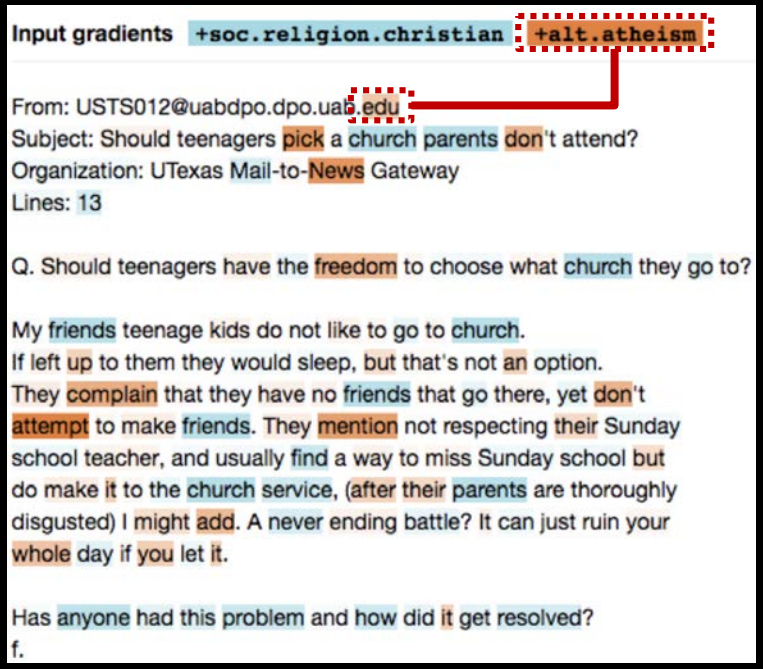
個別事例内の
入出力間の対応の
局所化に基づく説明




[Danielevsky+, 2021]

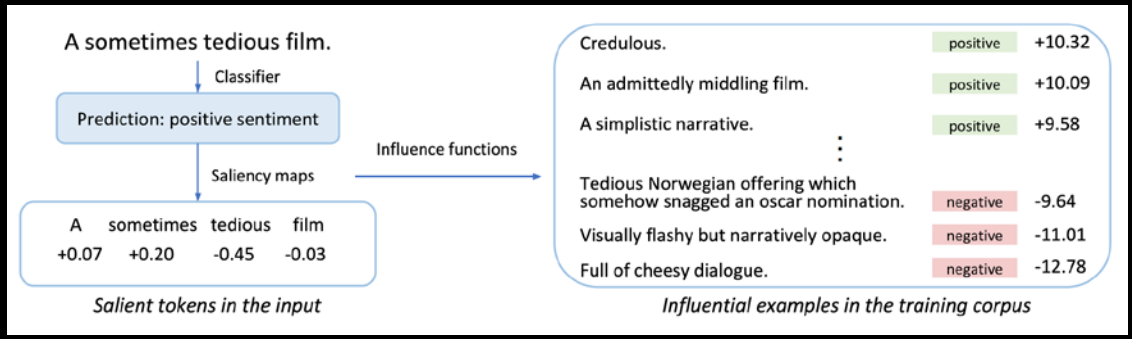
# NLPにおける説明手法 2/3

個別事例内の
入出力間の対応の
局所化に基づく説明




[Danielevsky+, 2021]

他の事例に基づく
説明




[Han+, ACL 2020]

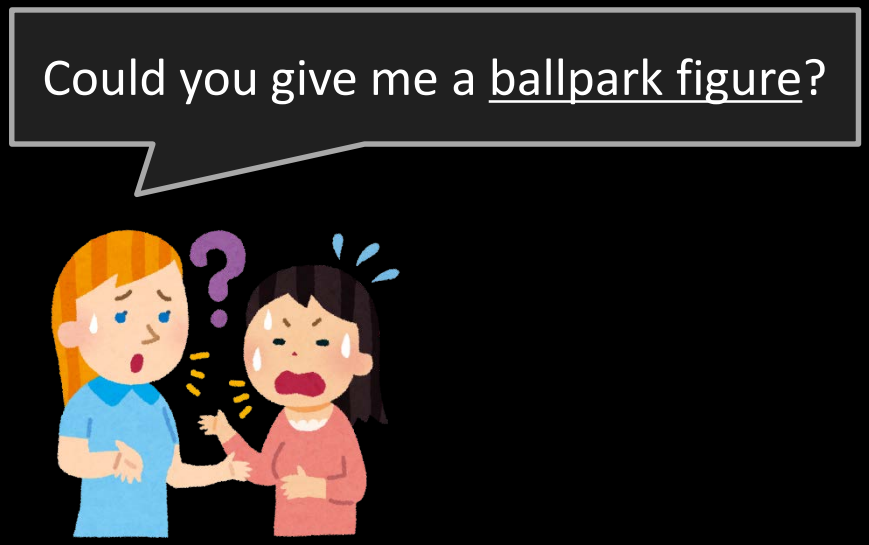
# NLPにおける説明手法 3/3

- 現状の課題を発見する用途が中心
  - 偽の (spurious) 手がかりへのモデルの依存
  - 訓練データ中のノイズ

- 本講演での提案: 説明手法は人文社会学的な応用が可能
  - [Harust+, 2020] の母語話者表現検出タスクを概念実証例として

# 母語話者表現検出タスク

[Harust+, 2020]



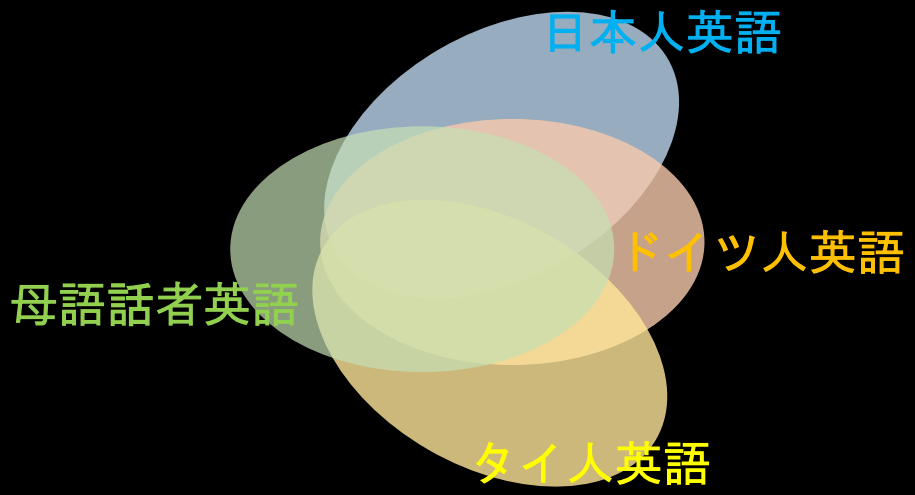

- 母語話者に特有の表現を検出するタスク
- 応用
  - 英語母語話者の国際的コミュニケーションへの適応促進
  - 熟練の第二言語 (L2) 話者のさらなる学習支援
  - 母語・L2言語獲得の違いの探究

# 国際的コミュニケーションにおける障害としての英語母語話者

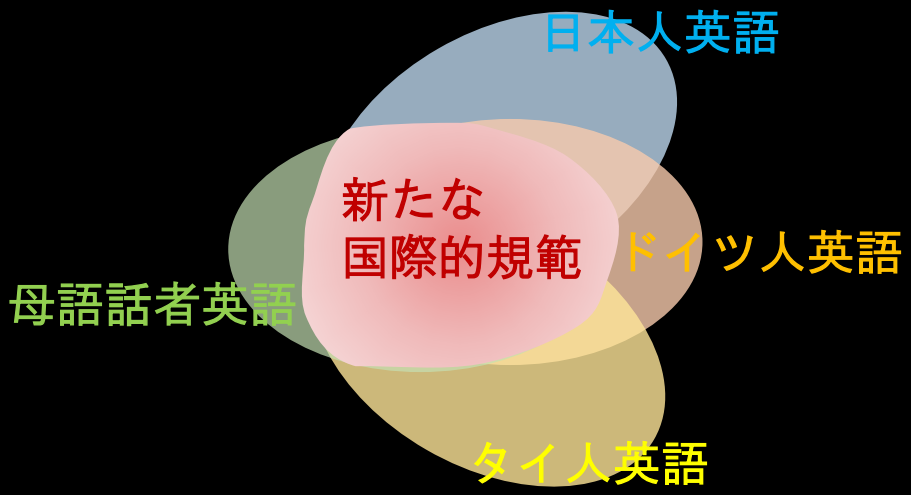
- 英語は国際的コミュニケーションにおける共通言語の地位を確立
- 英語母語話者の発話はL2話者には理解が困難という観察
  - L2話者同士のコミュニケーションでは支障がない熟練度であっても
- コミュニケーションの失敗はL2話者の矯正によって解消すべきなのか?
  - 語学教育産業$$は自然言語処理の重要な応用先だけど…

https://theconversation.com/why-native-english-speakers-fail-to-be-understood-in-english-and-lose-out-in-global-business-54436

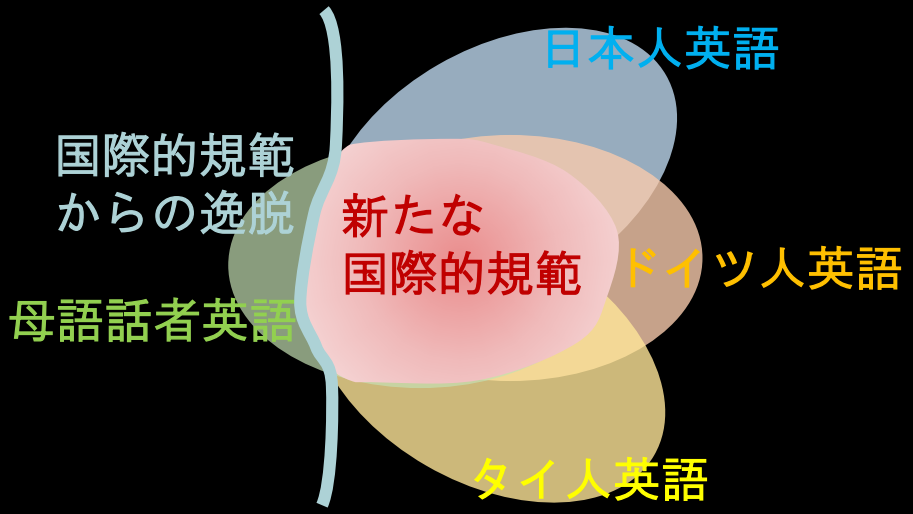
# 提案: 国際的コミュニケーションにおける新たな規範の創出



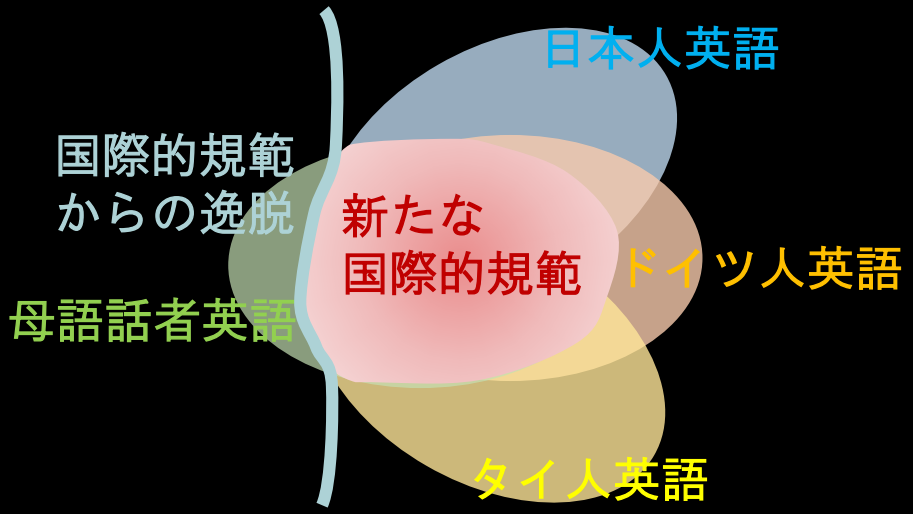
# 提案: 国際的コミュニケーションにおける新たな規範の創出



# 提案: 国際的コミュニケーションにおける新たな規範の創出



# 提案: 国際的コミュニケーションにおける新たな規範の創出

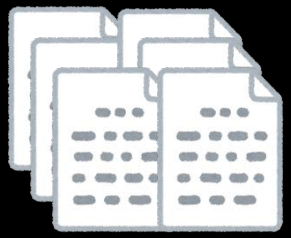



- 政治的課題: 数の上で母語話者を圧倒するL2話者を糾合する必要

  技術的課題: 何が母語話者表現 (国際的規範からの逸脱) かを明らかにする必要

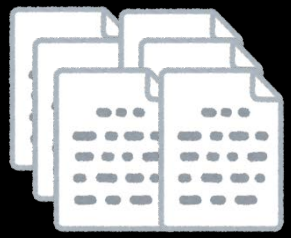
# 母語話者表現検出は対照研究

2つの集団間の差異をテキストを通じて明らかにしたい



母語話者が書いた
テキスト

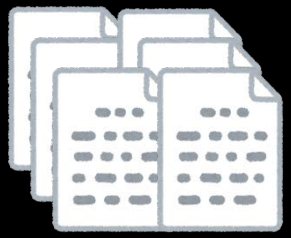
VS



L2話者が書いた
テキスト

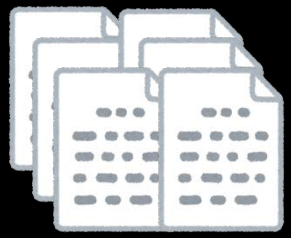
# 母語話者表現検出は対照研究

2つの集団間の差異をテキストを通じて明らかにしたい



母語話者が書いた
テキスト

VS

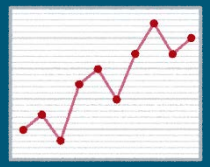


L2話者が書いた
テキスト

L2-Redditコーパス [Rabinovich+, TACL2018]

- Reddit: 英語圏の大規模オンライン掲示板群
- 一部では利用者が所属国 (≒母語話者 or L2話者) を申告
- 数千万文規模

# 解釈可能な機械学習では母語話者表現検出は困難

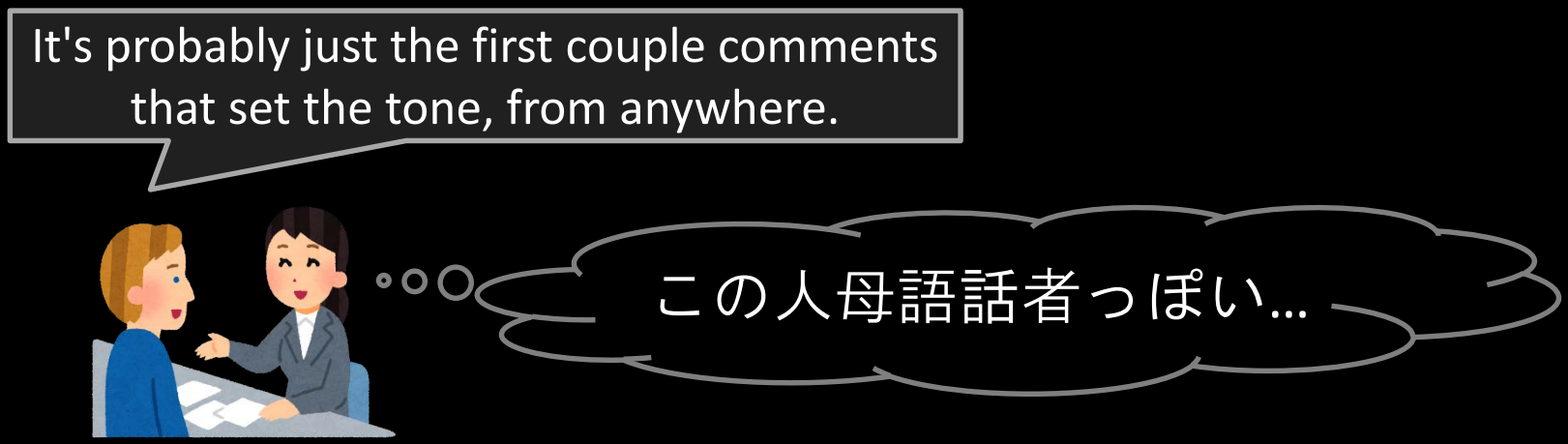


- 統計的手法を駆使した計量的分析
- 表層的手がかり (e.g. 単語頻度) に終始
- 検証可能性を持つ科学的手法

解釈可能な機械学習

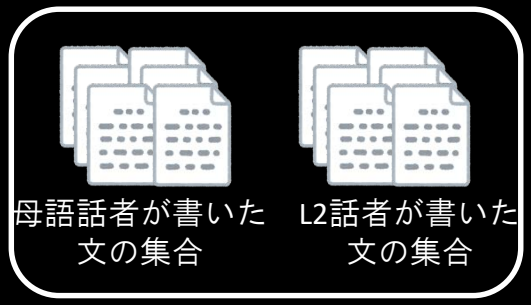
- 従来手法の例
  - 母語話者とL2話者の単語頻度分布の違いを検定
  - 文のbag-of-words表現による線形回帰
- 多義性や単語より長い表現が扱えない
  - ballpark figureは簡単な単語だけからなるが、L2話者には理解困難

# 人間の母語話者判定は ブラックボックス的

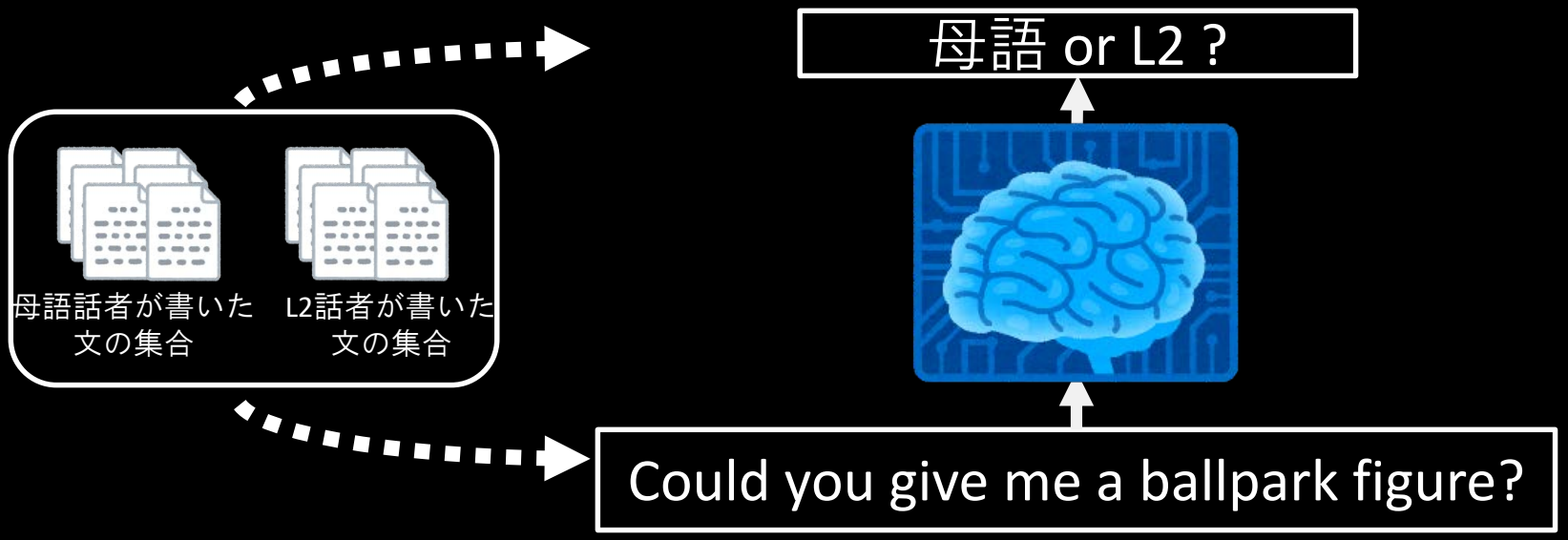



- 母語話者とL2話者の英語に長期間さらされた人間は話者の属性と表現パターンの連合を獲得する
- この連合は当人にとってもブラックボックス的であり、内省を必要とする

# 提案手法: 文単位のニューラル分類器 + 説明手法に基づく表現への絞り込み

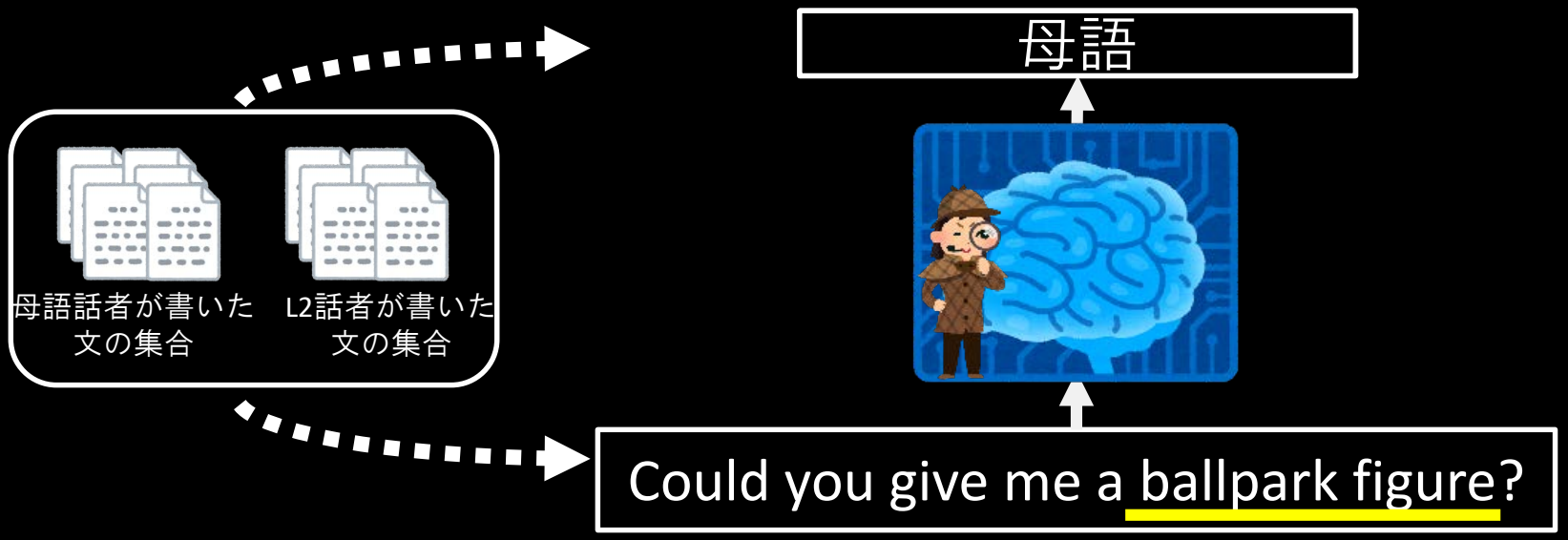


# 提案手法: 文単位のニューラル分類器 + 説明手法に基づく表現への絞り込み




- Step 1: 母語・L2を文単位で予測するニューラル分類器を訓練

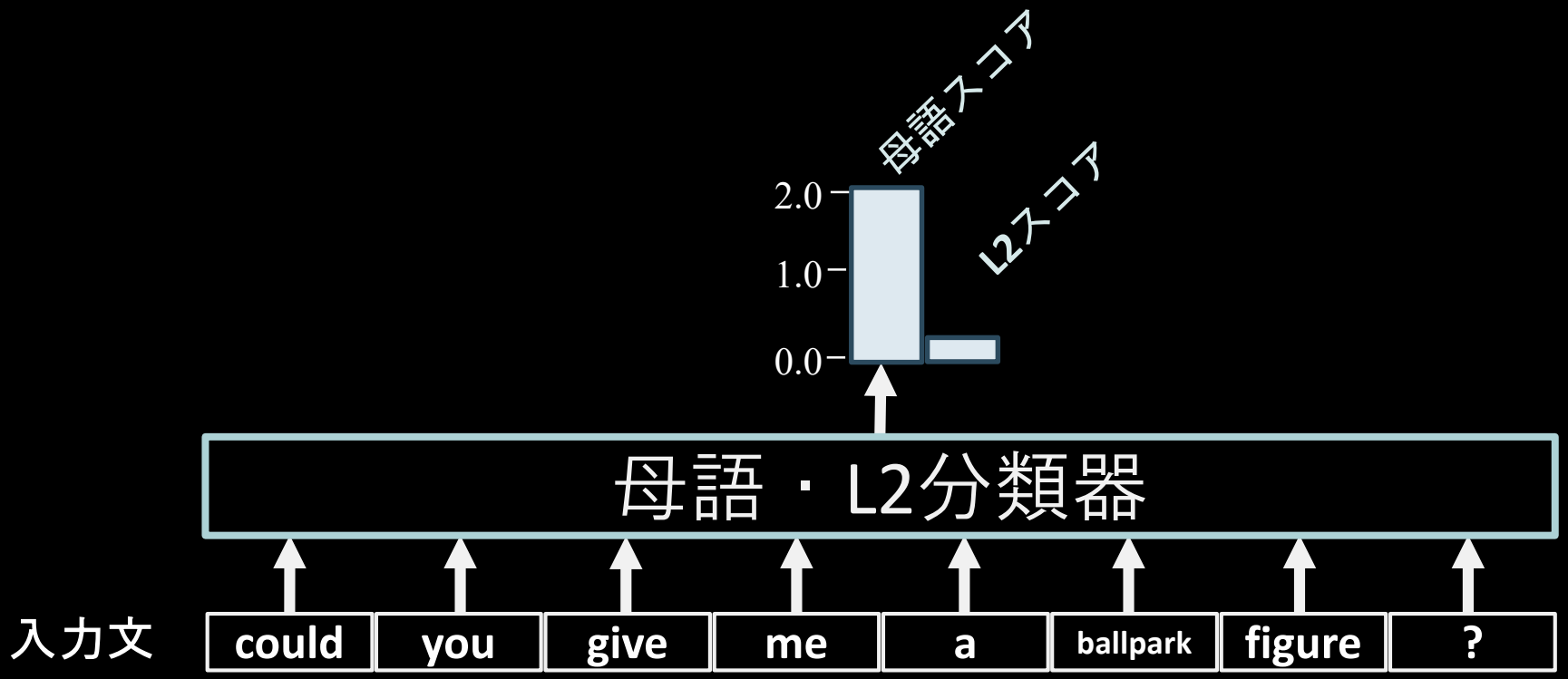
# 提案手法: 文単位のニューラル分類器 + 説明手法に基づく表現への絞り込み




- Step 1: 母語・L2を文単位で予測するニューラル分類器を訓練
- Step 2: 書き手が母語話者 & 分類器が母語と予測した事例について、予測に強く貢献した入力系列中の部分列を同定

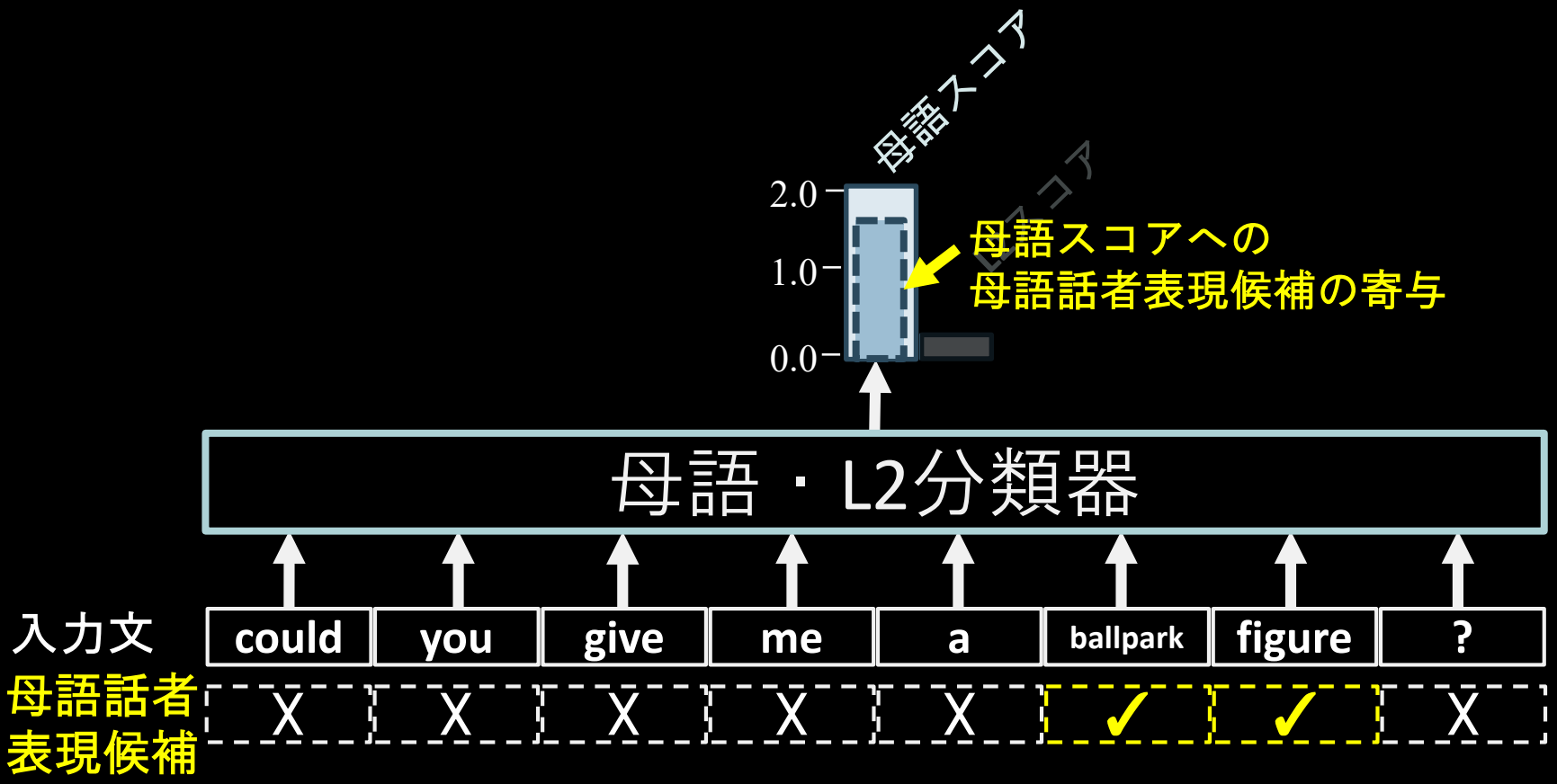
- ブラックボックス的なニューラル分類器＋説明手法
- 従来よりも深い文脈を捉えられる
- 検証可能性を持つ科学的手法

説明可能な機械学習



- ブラックボックス的なニューラル分類器 + 説明手法
- 従来よりも深い文脈を捉えられる
- 検証可能性を持つ科学的手法

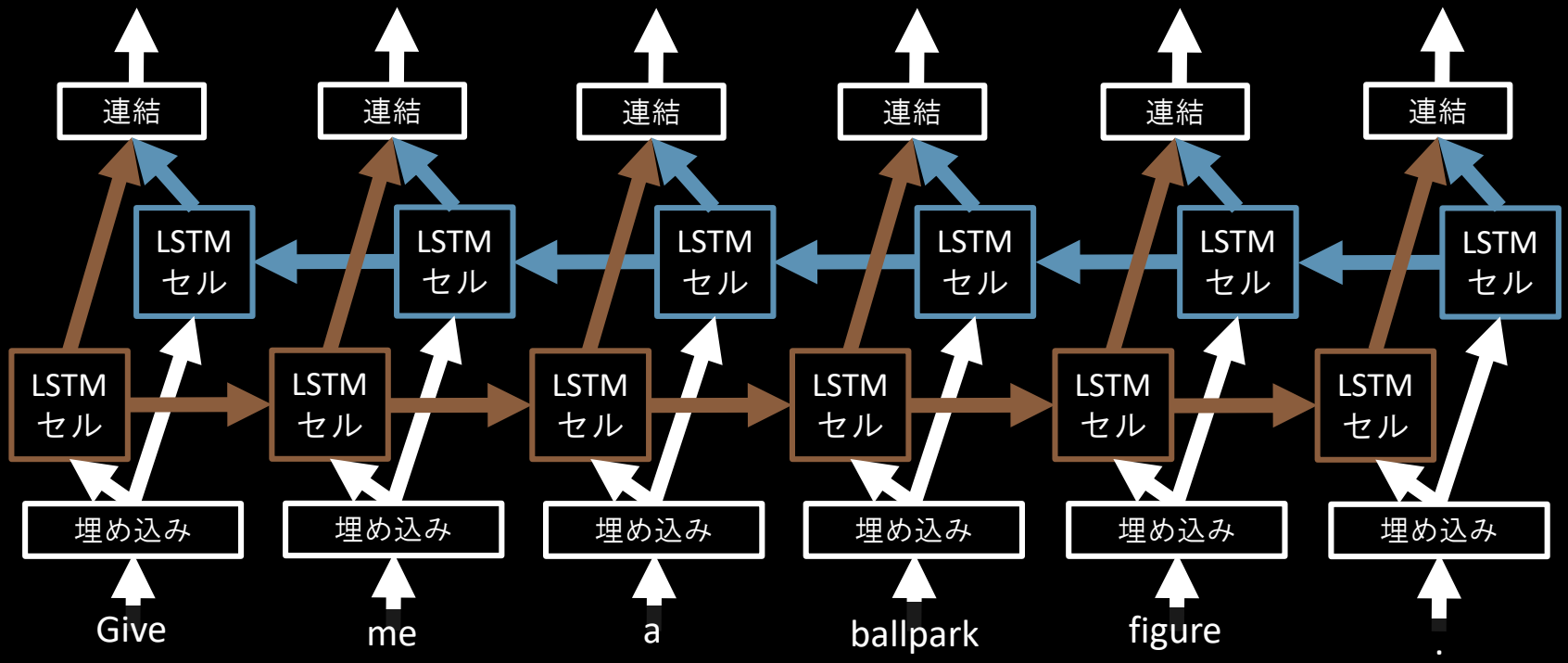
説明可能な機械学習



# ニューラル分類器 1/2

- 単語を連続値ベクトル (埋め込み) で表現することによる汎化
- 単語同士の非線形な相互作用を認識可能
  - ballparkはfigureが後続することにより「概算」の語義が発火

# ニューラル分類器 2/2




- [Harust+, 2020] は分類器を双方向LSTM [Graves+, 2005] を用いて構築
- 最近はTransformer [Vaswani, 2017] ベースの手法が有力

# 説明手法: 文脈分解 (contextual decomposition)

[Murdoch et al., 2018]

| Attribution Method | Heat Map |
|---|---|
| Gradient | used to be my favorite not worth the time |
| Leave One Out (Li et al., 2016) | used to be my favorite not worth the time |
| Cell decomposition (Murdoch & Szlam, 2017) | used to be my favorite not worth the time |
| Integrated gradients (Sundararajan et al., 2017) | used to be my favorite not worth the time |
| Contextual decomposition | used to be my favorite not worth the time |

Legend: Very Negative / Negative / Neutral / Positive / Very Positive

- ニューラル分類器の入力から出力までの計算をすべて、(1) 注目する入力部分列の貢献と (2) 入力の残りの部分の貢献に近似的に分解
- 他の手法とくらべて人間の直感とより整合的

# 母語話者表現検出の結果 1/3

難易度の高そうな単語に由来する事例はそれほど多くない

- They started with a wildly lopsided pitching match-up and kept it respectable.
- Seems to be privy to things that normal bloggers are not.

# 母語話者表現検出の結果 2/3

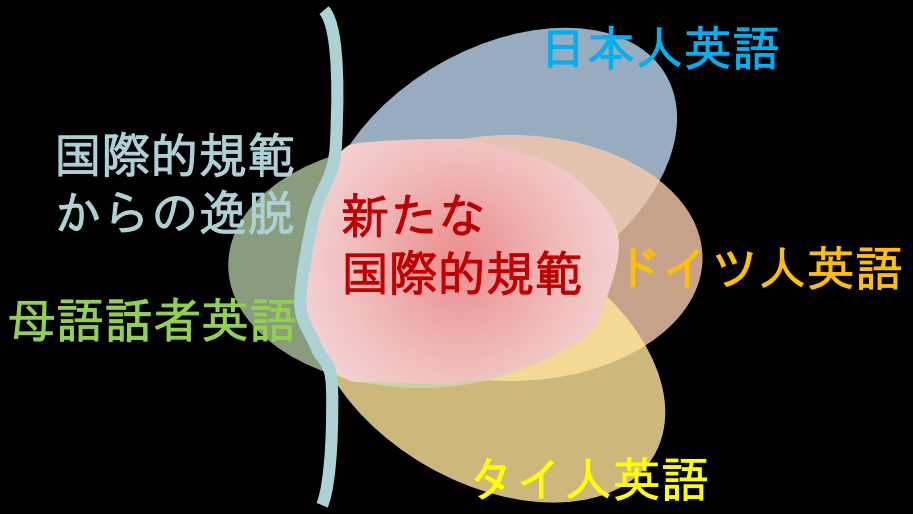
単語そのものは簡単でも全体としてL2話者には難しそうな事例

- Good luck with it, but I think the location is going to be a hard sell.
- Do you find that the objectionable ones get better as the glasses and bottles sit out a bit?
- The ATM at my old work would occasionally toss out bills stuck together.

# 母語話者表現検出の結果 3/3

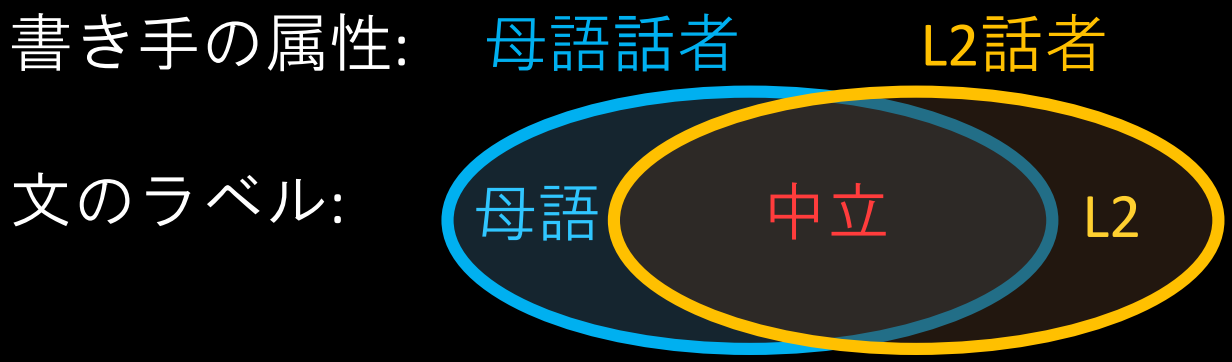
- 検出事例の半分弱は良い
- 残り半分は確かに母語話者らしいが表現として難しくないものや解析誤りなど
  - No, its pretty damn conclusive, that gun control and gun ownership have no effect on a nations violence rate.

# 中立ラベルの導入 1/2




- 母語話者、L2話者の英語の分布には大幅な重複があり、分離不可能
- そこで重複領域を表す中立ラベルを導入

# 中立ラベルの導入 2/2



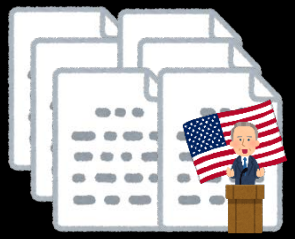

- 補完ラベルからの学習 [Yu+, 2018]
  - 書き手の属性: 母語話者
    = 文のラベル: 母語 or 中立
    = 文の補完ラベル: not L2
  - 書き手の属性: L2話者
    = 文のラベル: L2 or 中立
    = 文の補完ラベル: not 母語

# 偽の (spurious) 手がかりへの対策

- Bay Areaは母語話者表現?
  - 前処理で固有名詞を特殊トークンで置き換えてしまう
    - Bay Area ⇒ [LOC]
    - Angela Merkel ⇒ [PSN]
- gun control and gun ownershipは普通名詞なので前処理による対策は難しい
  - 母語話者とL2話者との間の話題の偏りが手がかりとして利用できそう

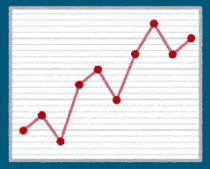
# 再掲: 対照研究 + 説明可能なAI (XAI)

2つの集団間の差異をテキストを通じて明らかにしたい



VS

- 人間がテキストを読んで感じたことを記述
- 深い洞察が可能
- 科学的観点からは、検証を欠いた仮説どまり



- 統計的手法を駆使した計量的分析
- 表層的手がかり (e.g. 単語頻度) に終始
- 検証可能性を持つ科学的手法

解釈可能な機械学習

- ブラックボックス的なニューラル分類器 + 説明手法
- 従来よりも深い文脈を捉えられる
- 検証可能性を持つ科学的手法

説明可能な機械学習

# まとめと今後の展望

- 対照研究の新手法: ニューラル分類器 + 説明手法
- ニューラル分類器、説明手法の開発は日進月歩
- 今後の課題:
  - 偽の手がかりへの対策
  - より小規模なテキストへの適用
  - データ全体を俯瞰する手法の開発