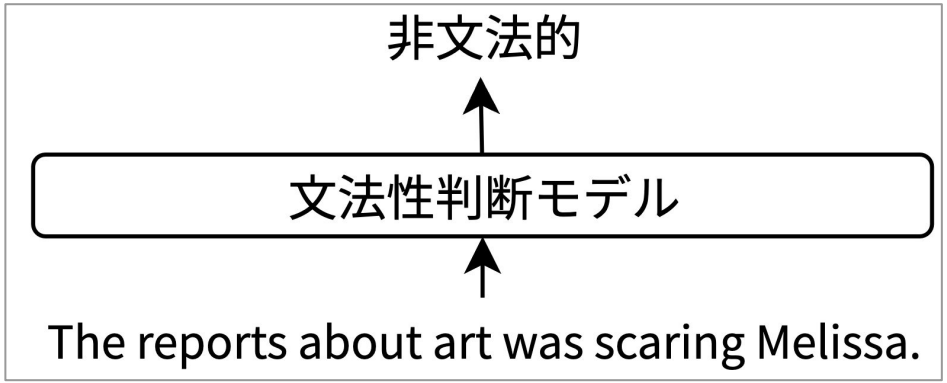
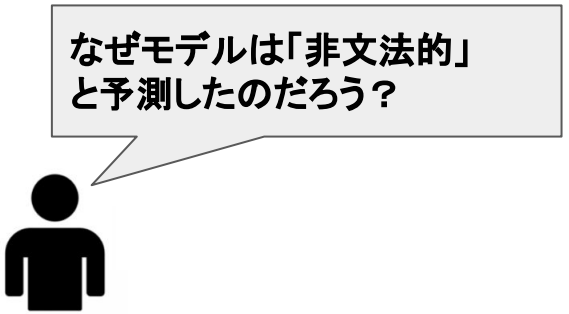
k近傍事例を用いたニューラルモデルの予測における定量的な解釈

五藤 巧, 出口 祥之, 上垣外 英剛 , 渡辺 太郎 (奈良先端科学技術大学院大学)

ニューラルモデルにおける予測根拠の解釈の必要性

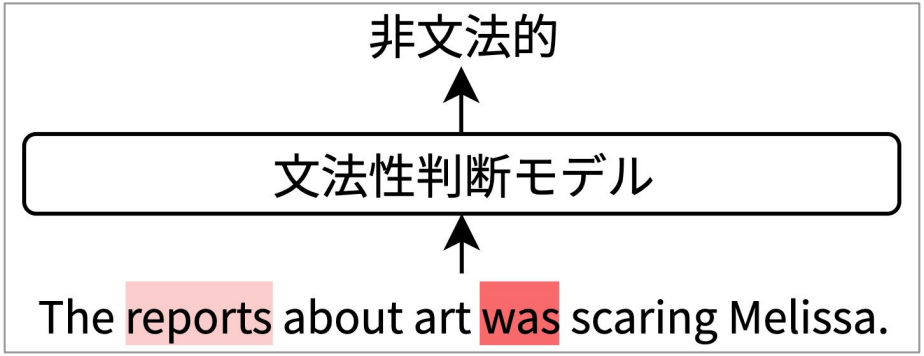
- **予測根拠の解釈**：モデルがどのような根拠で予測に至ったのかを知る
  - エラー分析：予測がなぜ誤ったのかを根拠を通じて知る
  - 信頼性の向上：ユーザに予測結果に加えて根拠まで提示する





先行研究：既存の解釈手法

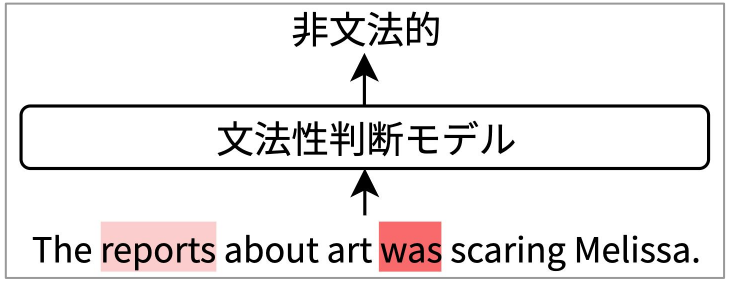
- **特徴量帰属**：どの単語が予測に貢献するかを定量化
  - 下図のように，単語に対する貢献度を示すヒートマップを提示

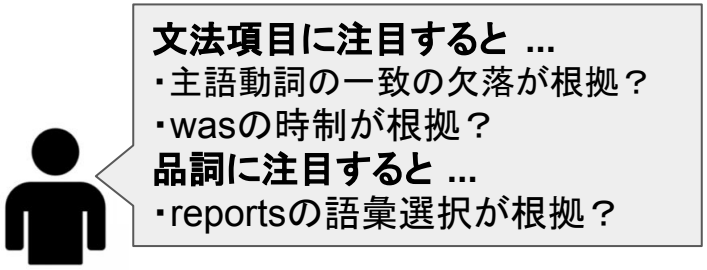


- **マルチタスク学習**
  対象タスクに 関連するサブタスクを同時に学習し，
  サブタスクの予測を解釈として使用

既存の解釈手法の課題

- 特徴量帰属：人間が解釈に介在する必要があり高コスト
  - 貢献度に基づいて，人間がタスク独自の観点を加えて解釈する必要あり





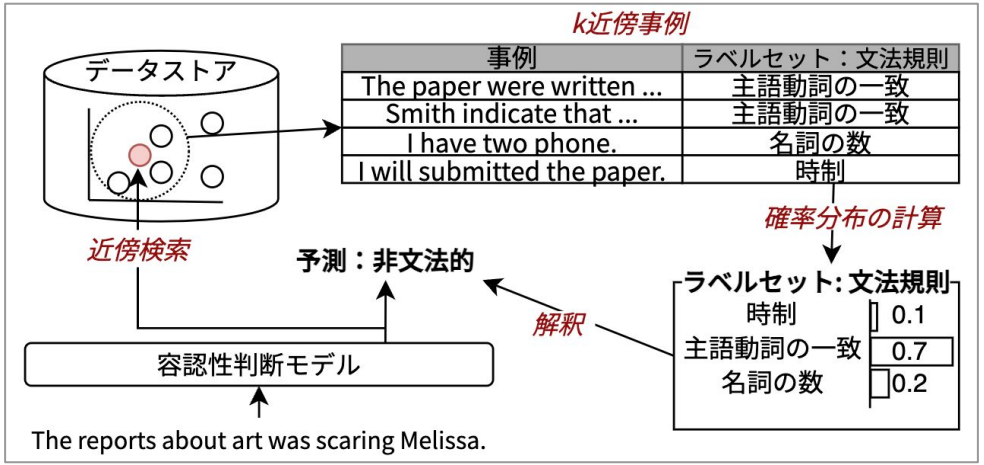
  - 解釈に人間の主観的なバイアスが混入する可能性
    - 理想的な結論につながるように，都合よく解釈してしまうかもしれない
    - 人間的には主語動詞の一致が根拠だが，モデルもそう思っているかは不明

- マルチタスク学習：サブタスクの学習によって元の予測結果が変化
  ＃どの観点で解釈するかを後から変えられない

提案法：k近傍事例に基づく定量的な解釈

- 人間が従来行う解釈を定量的に・客観的に提示したい
- モデルの埋め込み表現に基づくk近傍事例を利用
  - 時制を根拠に「非文法的」と予測するなら，同じく時制が根拠となる事例が入力事例と高い類似度になるはず
  - → どのような事例が埋め込み空間で類似するかに応じて根拠を定量化

- k近傍事例に基づいて解釈を確率分布として提示
  - 人間の判断が介在しないため低コスト・客観的
  - 複数事例に基づく「傾向」を提示可能



提案法：k近傍検索

- 対象タスクで学習されたモデル　Enc用いてデータストア　を$\mathcal{S}$作
  - 事例の埋め込み表現をキー，解釈ラベルをバリューとする辞書形式

$$\mathcal{S} = \{(\mathrm{Enc}(\boldsymbol{x}), c) \mid (\boldsymbol{x}, c) \in \mathcal{D}\}$$

事例の埋め込み表現　　事例と解釈ラベルのペア

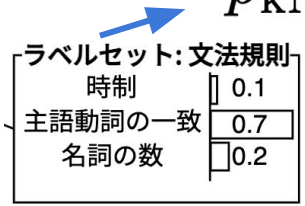
事例	ラベルセット：文法規則
The paper were written ...	主語動詞の一致
Smith indicate that ...	主語動詞の一致
I have two phone.	名詞の数
I will submitted the paper.	時制

$\mathcal{D}$の例

- 入力事例$\boldsymbol{x}'$を同様に埋め込み，
  k近傍事例$\mathcal{K} \subseteq \mathcal{S}$の距離に応じて解釈の分布を計算

2乗ユークリッド距離

$$p_{\mathrm{kNN}}(c_i|\boldsymbol{x}') \propto \sum_{(\boldsymbol{k},v)\in\mathcal{K}} \mathbb{1}_{v=c_i} \exp\left(\frac{-\|\boldsymbol{k} - \mathrm{Enc}(\boldsymbol{x}')\|_2^2)}{\tau}\right)$$



k近傍事例

温度パラメータ
大きいほど個々の事例が均一に影響

実験：容認性判断タスク

- 入力文が文法的に正しいかどうか判定するタスク
  - 言語モデルの統語知識を測定する目的
  - 言語理解ベンチマークGLUEにおけるCoLA

- 「文法的かどうか」に対する予測を次のラベルセットから解釈
  - 言語学における文法現象：BLiMPの12分類もしくは67分類の体系
  - 誤りタイプ：文法誤り訂正分野で定義される体系
    - VERB，NOUNなどの品詞情報と，VERB:SVAなどの一部文法項目

主語動詞の一致

非文法的な文：Janice is left by Samantha.
解釈ラベルセットとそのラベル BLiMPの12分類：argument structure BLiMPの67分類：passive_1 誤りタイプ：VERB

文法的な文：Janice is approached by Samantha.
解釈ラベルセットとそのラベル BLiMPの12分類：ACCEPTABLE（ダミーラベル） BLiMPの67分類：ACCEPTABLE（ダミーラベル） 誤りタイプ：CORRECT（ダミーラベル）

実験：提案法の適用手順と実験設定

用いるデータセット

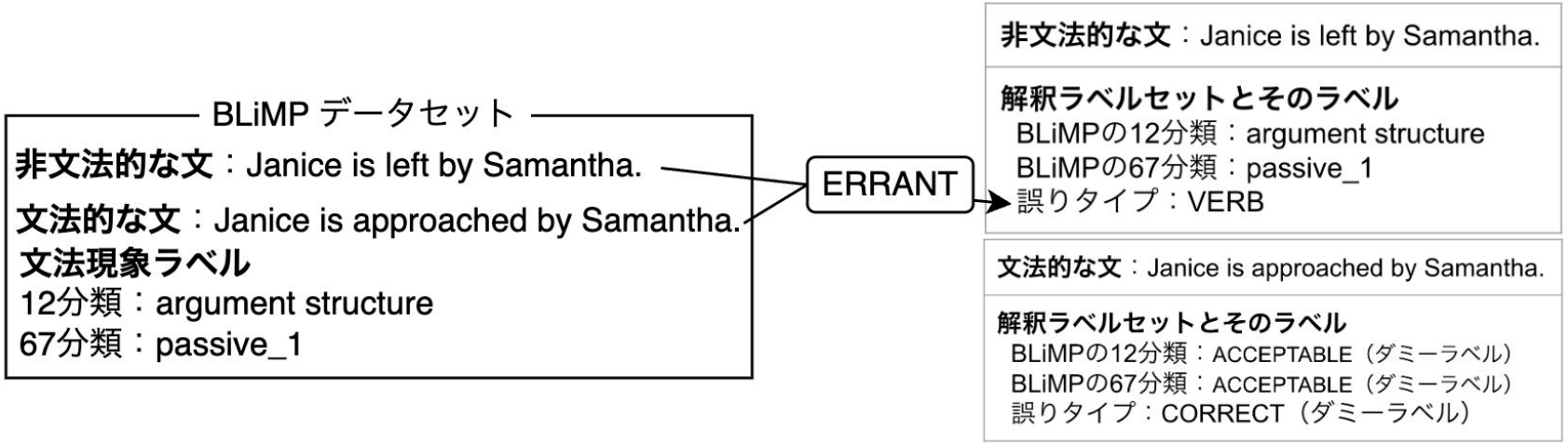
- CoLA：単文と文法的かどうかを示すラベル付きデータ
  → 学習に用いる
- BLiMP：表層に微差を持つ，文法的な文と非文法的な文のペアデータセット
  → 95:5に分割し，95%をデータストア，5%を評価に用いる

1. **対象タスクの学習**
   CoLAの学習データ8,551件を使用
   モデルはbert-base-cased，[CLS]に対応する表現で学習
   CoLA 評価データでMatthew´s Corr は 51.4

妥当な学習結果
BERTの原論文では52.1

実験：提案法の適用手順と実験設定

2. **データストアの構築**
   学習済みモデルを用いてBLiMP95%分割の各文を符号化
   文ペアなものを単文とその解釈ラベルとして扱う
   BLiMPの12分類・67分類についてはペアに付与されるラベルを使用
   誤りタイプはERRANTを用いて自動的に付与



実験：評価方法

- 提案法が提示する確率分布がどの程度妥当かを評価
  - BLiMPの5%分割を入力
- 不確実性キャリブレーションに基づく評価
  - 推定確率と実際の正解率が一致するかどうか
  - 70%の確率で予測した事例を集めれば，その中の70%が実際に正解するべき
- 尺度：期待キャリブレーション誤差（ECE）
  - 事例をその予測確率 (0.0, 0.1], (0.1, 0.2], . . . (0.9, 1]に応じてグループ化
    - 正解率計算のための正解ラベルはBLiMP5%分割に付与されているラベル

グループ内事例数

$$\mathrm{ECE} = \sum_{i=1}^{10} \frac{n_i}{N} \left| \underbrace{\mathrm{conf}_i}_{\text{グループ内平均予測確率}} - \underbrace{\mathrm{acc}_i}_{\text{グループ内平均正解率}} \right|$$

実験：k近傍検索時の設定

- 検索設定は多様に考えられる

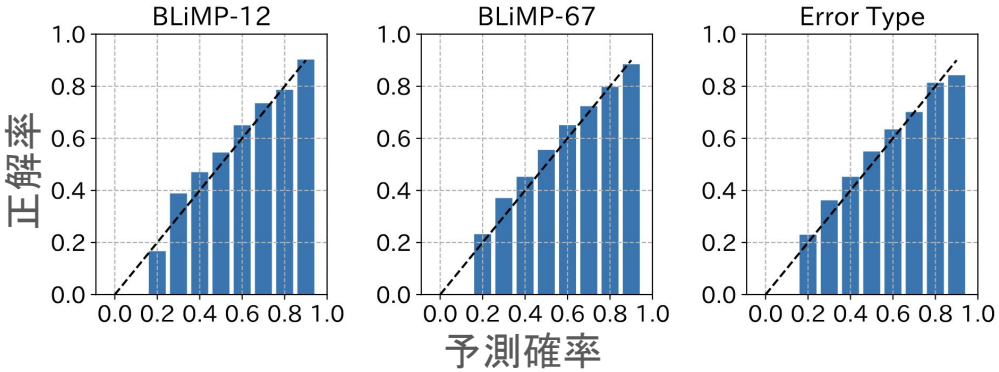
設定	説明	候補
K	k近傍事例をいくつ検索するか	{8, 16, …, 512, 1024}
T	個々の事例が与える均一性を制御する温度パラメータ（大きいほど均一に影響） $p_{\mathrm{kNN}}(c_i \mid \boldsymbol{x}') \propto \sum_{(\boldsymbol{k},v)\in\mathcal{K}} \mathbb{1}_{v=c_i} \exp\left(\frac{-\|\boldsymbol{k}-\mathrm{Enc}(\boldsymbol{x}')\|_2^2)}{\tau}\right)$	{0.001, 0.01, ..., 100, 1000}
層	何層目の[CLS]に対応する表現か	{1, …, 12}
FFN入出力	Transformerブロックのfeed forward層への入力表現と出力表現のどちらか	{入力表現，出力表現}

実験結果：最適な表現と評価結果

- ラベルセットによって最適な検索設定が異なる

解釈ラベルセット	ECE	K	τ	層	表現
BLiMP 12分類	0.0127	32	1.0	2	FFN入力
BLiMP 67分類	0.0114	16	1.0	4	FFN出力
誤りタイプ	0.0116	16	1.0	4	FFN出力

- 確率分布の信頼性は高い：ECEの各グループで予測確率と正解率は乖離しない



実験結果：ケーススタディ

非文法的との予測に対する解釈

- BLiMP12分類と誤りタイプでは
  主語動詞の一致が最も高い確率に

- BLiMP67分類では
  distractor_agreement_relational_noun
  が最も高い確率であり，
  モデルが名詞mothersと
  動詞doesの一致が欠落していること
  を根拠に予測したことが解釈可能

クエリ: The mothers of Cheryl does bake.	
BLiMP-12	確率（%）
subject_verb_agreement	42.72
ACCEPTABLE	40.53
argument_structure	11.26
npi_licensing	5.48
BLiMP-67	確率（%）
distractor_agreement_relational_noun	54.24
ACCEPTABLE	27.35
npi_present_2	4.84
intransitive	4.69
transitive	4.67
principle_A_domain_2	4.22
誤りタイプ	確率（%）
VERB:SVA	46.14
CORRECT	27.35
OTHER	12.79
ADV	4.84
VERB	4.67
NOUN	4.22

実験結果：ケーススタディ

文法的との予測に対する解釈

- 「非文法的と推論する余地をどのような観点で残しているか」を解釈可能

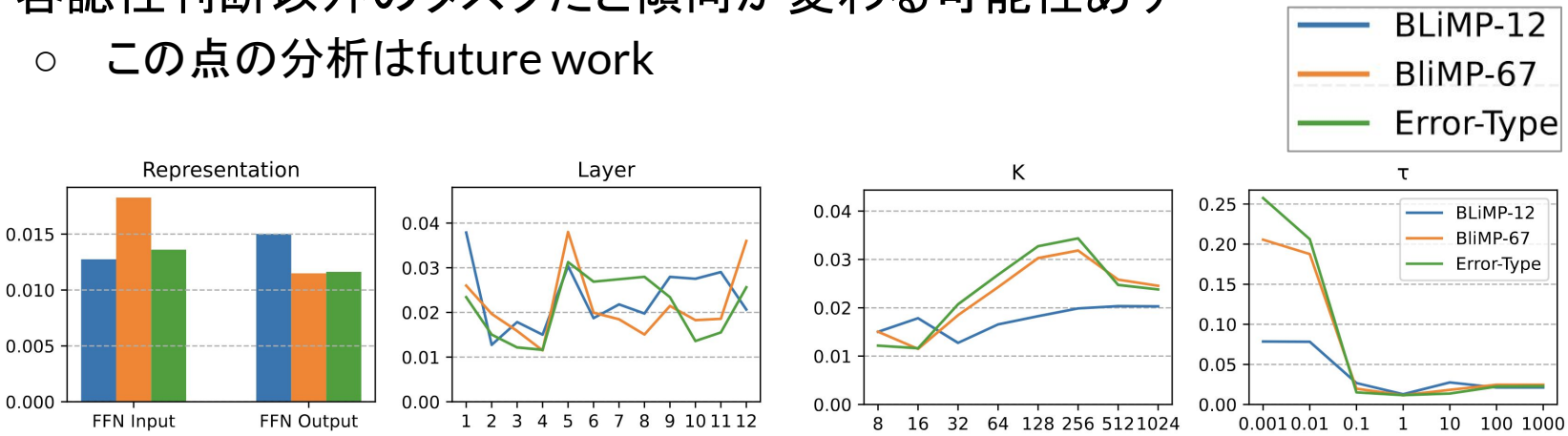
- 解釈結果から，特に冠詞と名詞の一致を根拠に非文法的と推論する余地を残している

クエリ: Jane sees some mirror that shocks Katherine.	
BLiMP-12	確率（%）
ACCEPTABLE	54.96
determiner_noun_arg.	23.89
filler_gap	17.74
binding	3.41
BLiMP-67	確率（%）
ACCEPTABLE	54.96
determiner_noun_agreement_with_adj_irregular_2	15.04
wh_questions_subject_gap	10.57
wh_questions_object_gap	7.17
determiner_noun_agreement_2	5.25
誤りタイプ	確率（%）
CORRECT	54.96
DET	20.96
DET,PRON	10.57
OTHER	6.71

分析：検索設定による解釈性能の変化

- 検索設定に応じたECEの変化を分析
  - 色の違いはラベルセットの違い
  - 表現以外は，ラベルセット間で概ね同じ傾向を示す
    - あるラベルセットで適切な表現を決定すれば，
      他のラベルセットにも使い回せる

- 容認性判断以外のタスクだと傾向が変わる可能性あり
  - この点の分析はfuture work



まとめ

概要

- **動機**：ニューラルモデルの解釈において，従来手法が高コストである点を指摘
- **手法**：解釈ラベルセットを用いて，それに属するラベルの確率分布を提示
- **実験**：容認性判断タスクにおいてECEによる定量評価，ケーススタディによる定性評価を実施

今後の課題

- 他タスクでの実験
- モデルの大局的な解釈への拡張
  - 事例ごとに提示された確率分布をうまく統合したい
- LLMへの応用
  - プロンプト入力時点の表現を用いて生成結果を解釈するなど