Q3-9 一般性を考慮した言語処理モデルのShortcut Reasoningの自動検出

原口 大地 (JAIST), 白井 清昭(JAIST), 井之上 直也 (JAIST/理研)

NLP 2023 OKINAWA

背景

Shortcut Reasoning：
学習データ内の入力-ラベル間の疑似相関にモデルが依存することで，推論プロセスにおいて発生する非合理的な推論
頑健性：
学習データと同じ分布を持つデータ(IID) と，異なる分布を持つデータ(OOD) に対するモデルの解析性能の差の小ささの程度

Shortcut Reasoningは頑健性を低下させる

Training step
“*Titanic is great.*” → positive

Inference step
“*The sinking of Titanic was tragedy.*” → positive

先行研究の課題

1. Shortcut Reasoning の形態を事前に定義している – Han+(2020), etc.

事前に想定したShortcut Reasoning に対してのみ検証をするため，モデルに潜在する未知の形態のものを明らかにできない可能性

2. 人手による判定が必要となる – Pezeshkpour+(2022), etc.

単純にコストがかかるのに加え，一見Shortcut Reasoning に見えないような事例を見逃す可能性がある

3. モデルの内部情報を直接考慮していない – Ribeiro+(2020), etc.

特定の編集や特徴を加えた入力に対する出力のエラー分析を行った手法が多いが，そこから我々が得られる情報には限度がある

これら3つの課題をクリアした手法を提案

提案手法

Shortcut Reasoning検出の手順
1. モデルから推論パターンの候補を抽出
2. 推論パターンの候補の一般性を計算し，高いものを採用
3. IIDで有効かつOODで効かない推論パターンShortcut Reasoningとして検出

推論パターン：
ある入力に対するモデル(f)の推論プロセスにおいて，何らかのトリガー(t)が特定のラベル(l)の予測をもたらす規則

$$p \overset{\text{def}}{=} t \overset{f}{\rightarrow} l$$

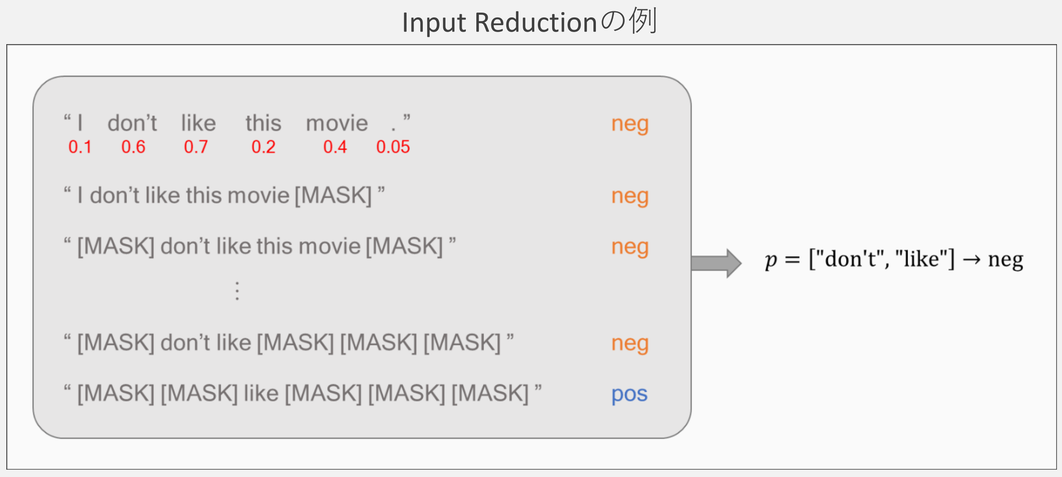
推論パターンの抽出 (Input Reduction)：
予測に必要な最低限の単語の系列($\boldsymbol{w}$)をトリガーt，その予測をラベルlと仮定
1. Integrated Gradient(IG)を用いて各単語の予測への寄与度を計算
2. スコアの低い順にMASKをかけ，予測が変わる直前に残った単語とそのラベルを出力

推論パターンの一般性 (g)：
推論パターンが，モデルの予測にどの程度影響を与えているかを示す
トリガーを含む事例の集合($E(\boldsymbol{w})$)に対して，モデルがラベルと同じ予測をする割合

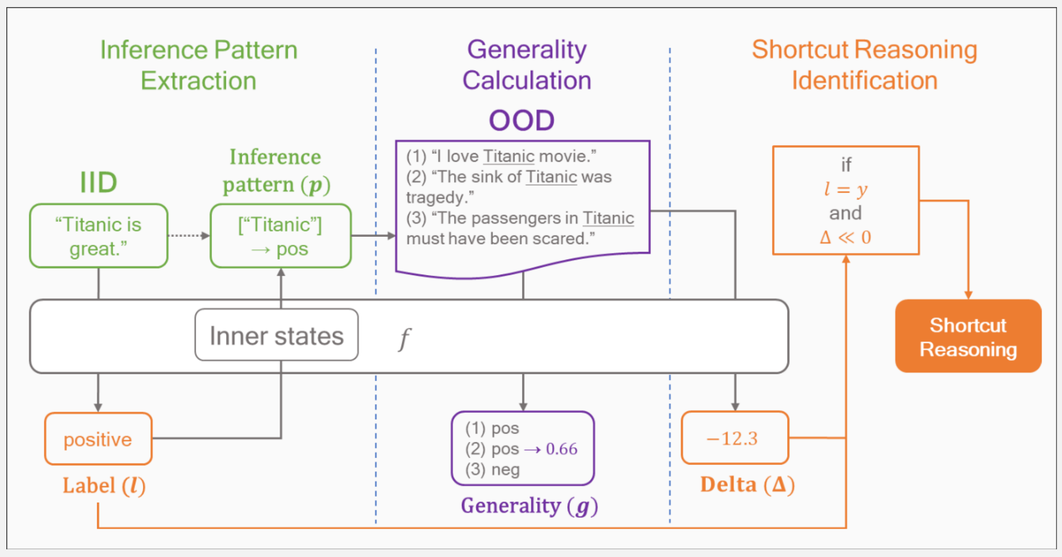
Shortcut Reasoningの判定：
i. ラベルが正解ラベルと一致 (IIDで有効)
ii. OODのデータセットと，そこから抽出された事例の集合$E(\boldsymbol{w})$のF1スコアの差(Δ)が一定以下

$$\Delta \overset{\text{def}}{=} \mathrm{F1}(E(\boldsymbol{w}), f) - \mathrm{F1}(\mathcal{D}_{\mathrm{OOD}}, f)$$

Input Reductionの例



提案手法の概観



結果と考察

適用タスク：自然言語推論(NLI)・感情分析(SA) (ともに3値分類)
モデル：huggingfaceに公開されているfine-tune済みRoBERTa

NLI：
- 先行研究でも報告されている，**否定表現・Hypothesisに依存**したShortcut Reasoningを検出
- 加えて，[“is”, “popular”]，[“as”, “well”]といったShortcut Reasoningも新たに検出された

SA：
- 感情語が主にトリガーとして検出・一見適切に見えるが，Δが小さい
- 多くがpositiveとnegativeが混ざったneutralの文を誤って解答，**文中の片方の感情語に依存している可能性**
- “Titanic”のように全くラベルに関連性のない単語は実はあまりShortcut Reasoningになっていない？

今後の課題：頑健性の向上・他タスクへの適用・結果の評価方法の模索

p	g	Δ	$\lvert E\rvert$
[“/s”, “is”, “popular”]→ neutral	85.3	-9.4	291
[“/s”, “never”]→ contradiction	80.9	-7.7	1515
[“/s”, “as”, “well”]→ neutral	60.9	-12.0	151
[“/s”, “not”]→ contradiction	54.5	-22.8	8708

p	g	Δ	$\lvert E\rvert$
[“worst”]→ negative	97.5	-24.0	158
[“Excellent”]→ positive	96.2	-11.9	184
[“Perfect”]→ positive	96.0	-13.4	324
[“Poor”]→ negative	95.3	-8.8	169