

Semantically Equivalent Adversarial Rules for Debugging NLP Models

Marco Tulio Ribeiro¹, Sameer Singh², Carlos Guestrin¹

¹University of Washington, ²UC Irvine

読み手: 三輪誠 (豊田工業大学)

※一部の図表は著者の論文・スライドより引用

概要

- モデルのOversensitivity
 - 複雑な機械学習モデルは意味がほとんど同じ入力でも違う結果
- Oversensitivityを見つける例・ルールの生成・利用方法を提案
 - Semantically Equivalent Adversaries (SEAs)
 - Semantically Equivalent Adversarial rules (SEARs)

The biggest city on **the river Rhine** is Cologne, Germany with a population of more than 1,050,000 people. It is **the second-longest river** in Central and Western Europe (after the Danube), at **about 1,230 km (760 mi)**

How long is the Rhine?

1230km

How long is the Rhine??

More than 1,050,000



やりたいこと

- 意味を保ちつつ、モデルが間違える例を作りたい (SEAs)
- 間違える例をたくさん作るルールを作りたい (SEARs)



What color is the sky?

Blue

Which color is the sky?

Gray

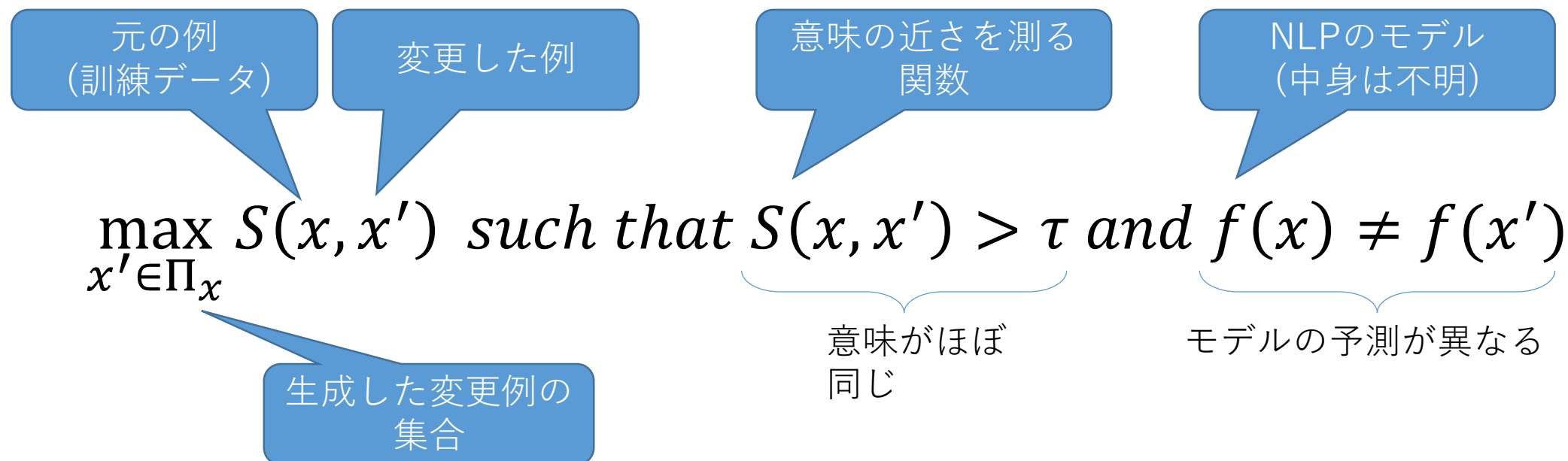


Rule $r(\textit{What NOUN} \rightarrow \textit{Which NOUN})$

ACL2018読み会@名大

Semantically Equivalent Adversaries (SEAs)

- 変更例を自動生成 (➡次ページ)
- モデルが間違え, かつ, 最も意味に近い例を (あれば) 選択



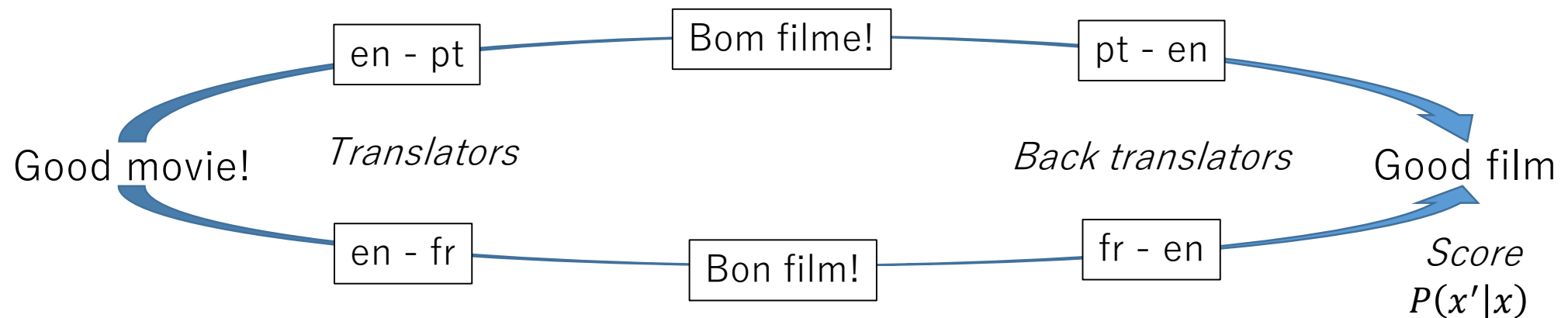
変更例の生成

- 複数ピボット言語を介した翻訳・逆翻訳による Paraphrasing [Mallinson et al., 2017]

$$S(x, x') = \min \left(1, \frac{P(x'|x)}{P(x|x)} \right)$$

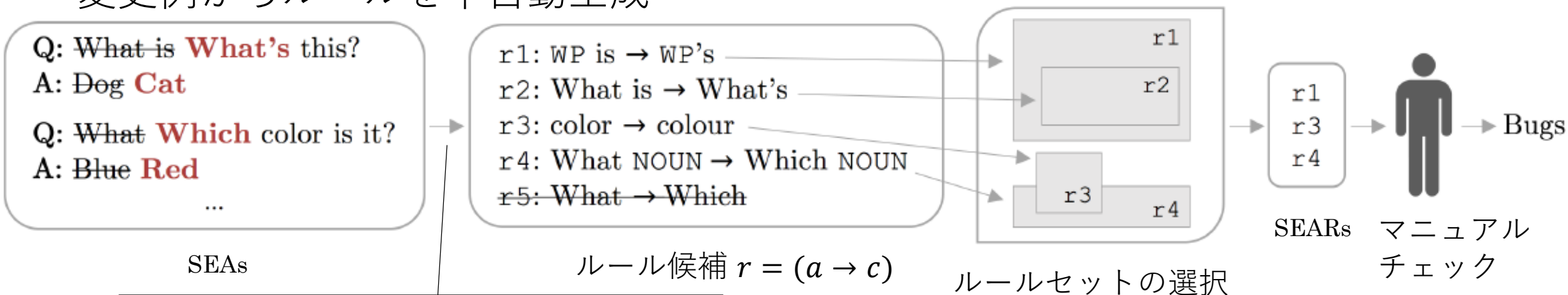
変換した文の生成確率

元文の生成確率



Semantically Equivalent Adversarial Rules (SEARs)

変更例からルールを半自動生成

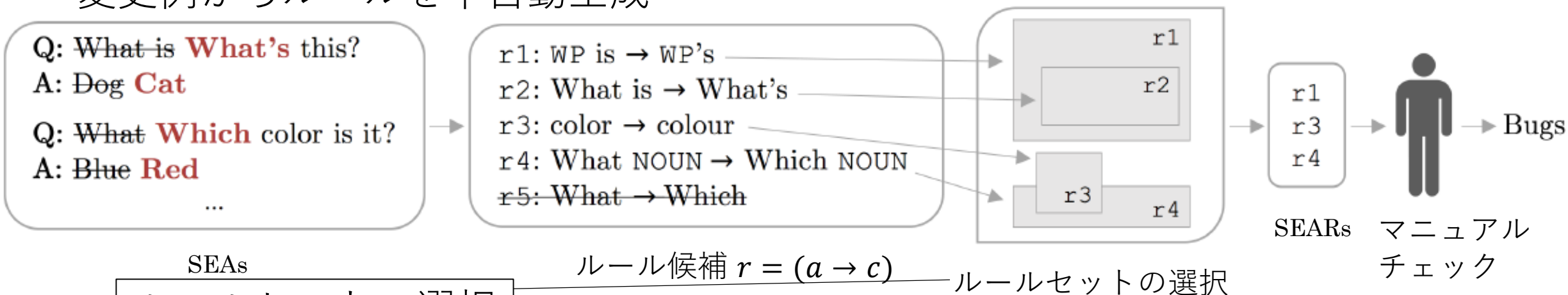


ルール候補($r = (a \rightarrow c)$)の作成

- 元の文と変更した文 (SEAs) の**最小の連続した変更部分**を抽出
- 変更部分の**前後1単語**まで含めて, **ルール候補**として抽出
 - 前後1単語を品詞にして**汎化**したものも追加

Semantically Equivalent Adversarial Rules (SEARs)

変更例からルールを半自動生成



ルールセットの選択

- 意味の異なる不自然な例を作るルールを削除
 - $E[\mathbb{I}[S(x, r(x)) > \tau]] \geq 1 - \delta$ ($\mathbb{I}$ はindicator function)
- 多くの例を被覆する・重複しないルールセット R を選択
 - $\max_{R, |R| < B} \sum_{x \in X} \max_{r \in R} S(x, r(x))$ such that $S(x, r(x)) > \tau$ and $f(x) \neq f(x')$
 - この最大化問題は解けないので, Greedyに一つずつ選択

Semantically Equivalent Adversarial Rules (SEARs)

変更例からルールを半自動生成

Q: What ~~is~~ **What's** this?

A: Dog **Cat**

Q: What **Which** color is it?

A: Blue **Red**

...

r1: WP is → WP's

r2: What is → What's

r3: color → colour

r4: What NOUN → Which NOUN

r5: What → Which

r1

r2

r3

r4

r1

r3

r4

SEARs



Bugs

マニュアル
チェック

SEAs

ルール候補 $r = (a \rightarrow c)$

ルールセットの選択

マニュアル・チェック

- Paraphrasingモデルは完璧ではない（文法的な誤り等も含む）
 - 選んだルールはデータ依存で汎用でない可能性がある
- 人手でチェック
- ※実験ではルールを選ぶごとにチェックして、次のルールを選択

評価設定

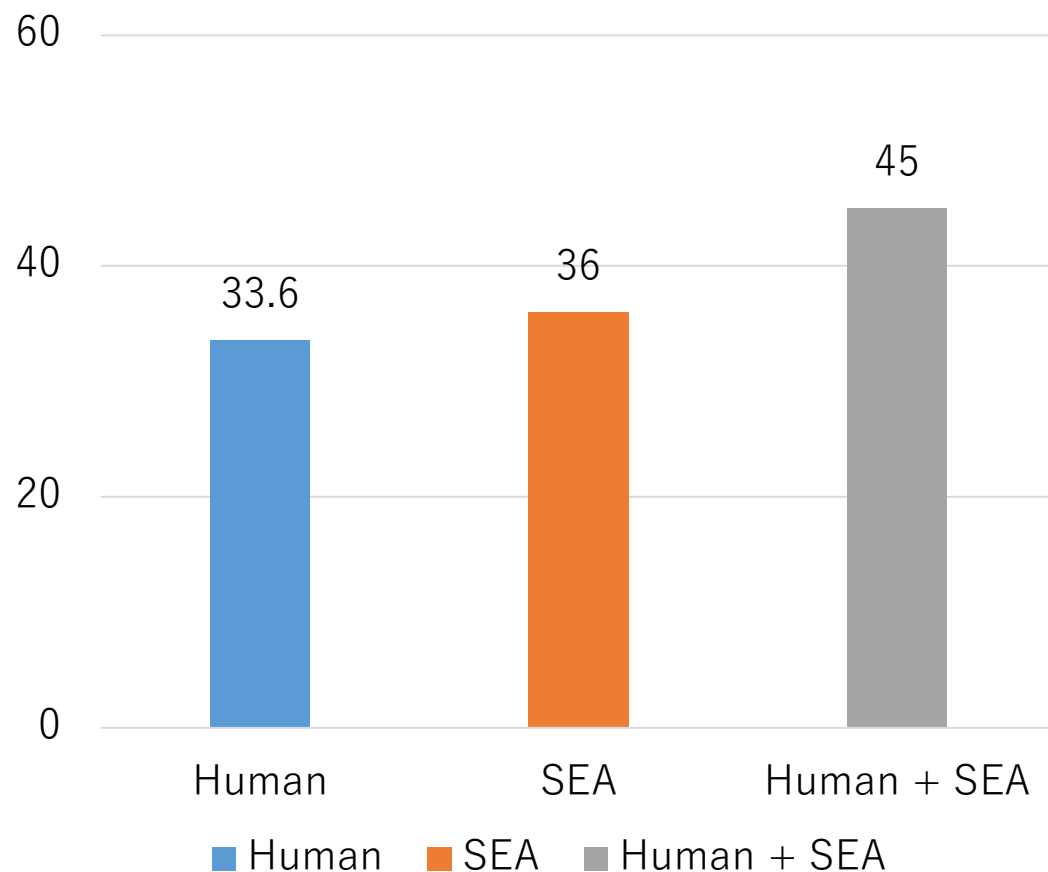
- 3つのデータセットで評価
 - Machine Comprehension (SQuAD) BiDAFモデル [Seo et al., 2017]
 - Visual QA (Visual7W-telling) Attentionモデル [Zhu et al., 2016]
 - Sentiment Analysis (Movie Review) fastTextモデル [Joulin et al., 2016]
- このプレゼンではVisual QAタスクの結果のみ紹介
 - 他の実験に関しては論文を参照

作成したルール (SEARs) の例

正解の変化した
割合

SEAR	Questions / SEAs	f(x)	Flips
WP VBZ→ WP's	What has What's been cut? Who is Who's holding the baby	Cake Pizza Woman Man	3.3%
What NOUN→ Which NOUN	What Which kind of floor is it? What Which color is the jet?	Wood Marble Gray White	3.9%
color →colour	What color colour is the tray? What color colour is the jet?	Pink Green Gray Blue	2.2%
ADV is→ ADV's	Where is Where's the jet? How is How's the desk?	Sky Airport Messy Empty	2.1%

モデルが間違える意味の近い変更例 (SEAs) を作成できたか？



- 100個の例それぞれについて
 - **Human:** 10個の変更例を人手 (3人のTurker) で作成し, 意味の近いものを選択
 - モデルで変更例を評価
 - **SEA:** 提案手法
 - **HUMAN+SEA:** 提案手法で5個の変更例を作り, 人手で元の文に意味の近いものを選択
- 生成した例を他の10人以上のTurkerが意味の近さで評価

自動生成した例と手動生成した例の違い

- 自動生成

Original	SEA
Where are the men?	Where are the males?
What kind of meat is on the boy's plate?	What sort of meat is on the boy's plate?

- 手動生成

Original	Human-generated SEA
How many suitcases?	How many suitcases are sitting on the shelf?
Where is the blue van?	What is the blue van's location?

- 編集した平均文字数

- 自動生成 6.2
- 手動生成 9.0

- 人間は小さな変更は思いつかないが、文構造を変えたり文脈を考えたりできる

手動ルール (Expert) と半自動ルール (SEARs) の比較

- 手動生成したルール - モデルの出力を参考に最大10個
- 半自動生成したルール - 1つずつ候補を20個まで生成し, 人手で最大10個まで選択 (候補を除外したら次の候補は再計算)

The screenshot displays the SEARs interface, which is used for evaluating and refining rules for question-answer pairs. The interface is divided into several sections:

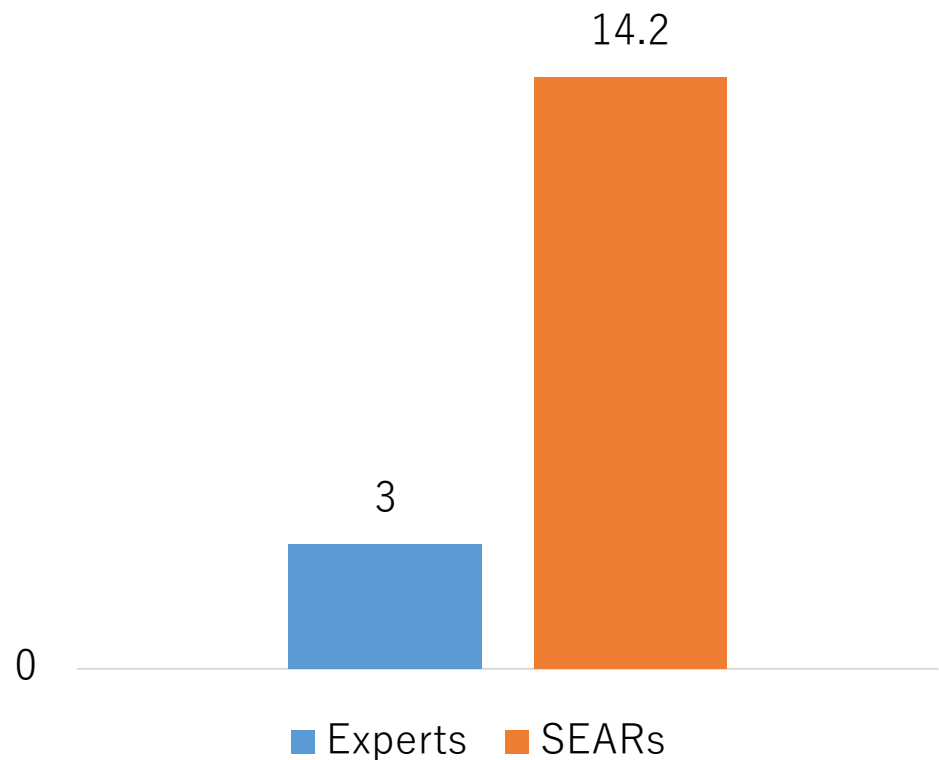
- Individual predictions**: A section for trying different rules. It includes a "List of POS tags" button, a text input for "Replace first instance of:" (currently "What NOUN"), a text input for "With:" (currently "Which NOUN"), and a "Submit" button.
- Rules to evaluate**: A section for evaluating a specific rule. It shows the "List of POS tags" button, the current rule "replace(What NOUN, Which NOUN)", and a "Does the current rule induce a bug?" section with "No" and "Yes" buttons.
- Progress**: A circular progress indicator showing "1 of 20" rules evaluated.
- Results**: A section showing the results of the rule evaluation. It includes a "replace(What NOUN, Which NOUN)" button, a "Mistake examples" section with a "Compact" checkbox, and a table comparing the "Original" and "After rule" results for three examples.

The "Results" section shows the following examples:

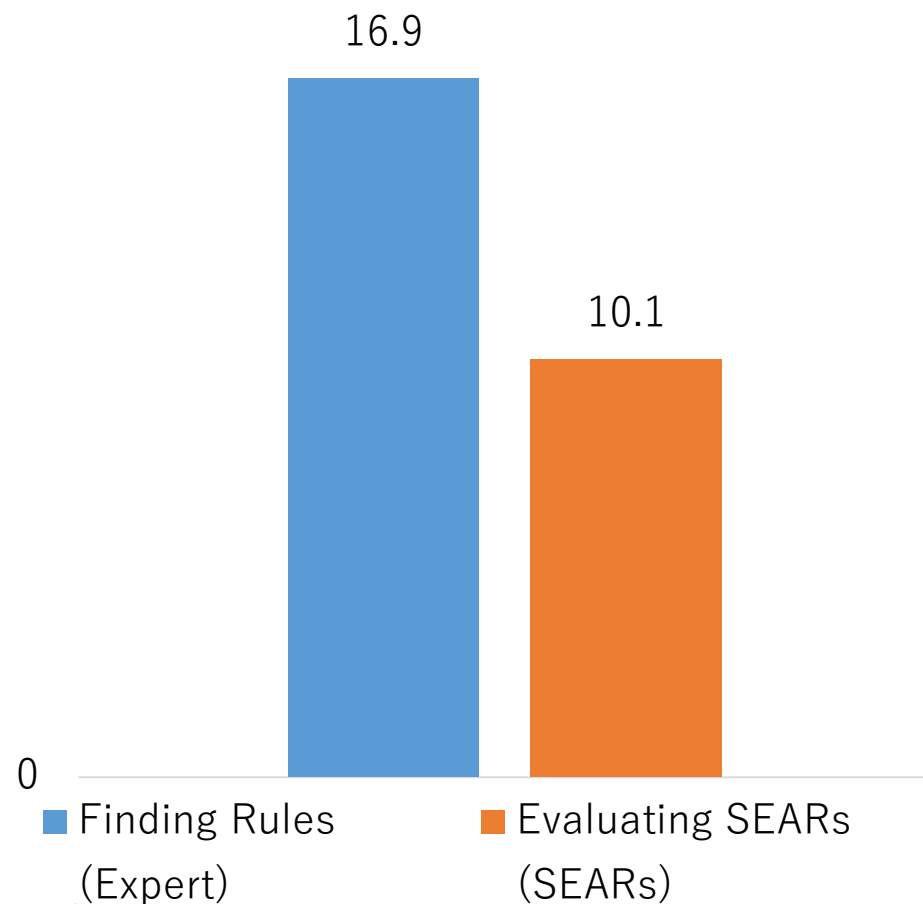
Image	Original	After rule
	Q: What color are the pots ? Answer: a) Silver. b) Black. c) White. d) Gold.	Q: Which color are the pots ? Answer: a) Silver. b) Black. c) White. d) Gold.
	Q: What color is the lampshade ? Answer: a) A light yellow. b) A bright red. c) A subtle green. d) A vivid orange.	Q: Which color is the lampshade ? Answer: a) A light yellow. b) A bright red. c) A subtle green. d) A vivid orange.
	Q: What animal is running in the background ? Answer: a) A dog. b) A horse.	Q: Which animal is running in the background ? Answer: a) A dog. b) A horse.

手動ルール (Expert) と半自動ルール (SEARs) の比較

20 予測の変化した例の割合

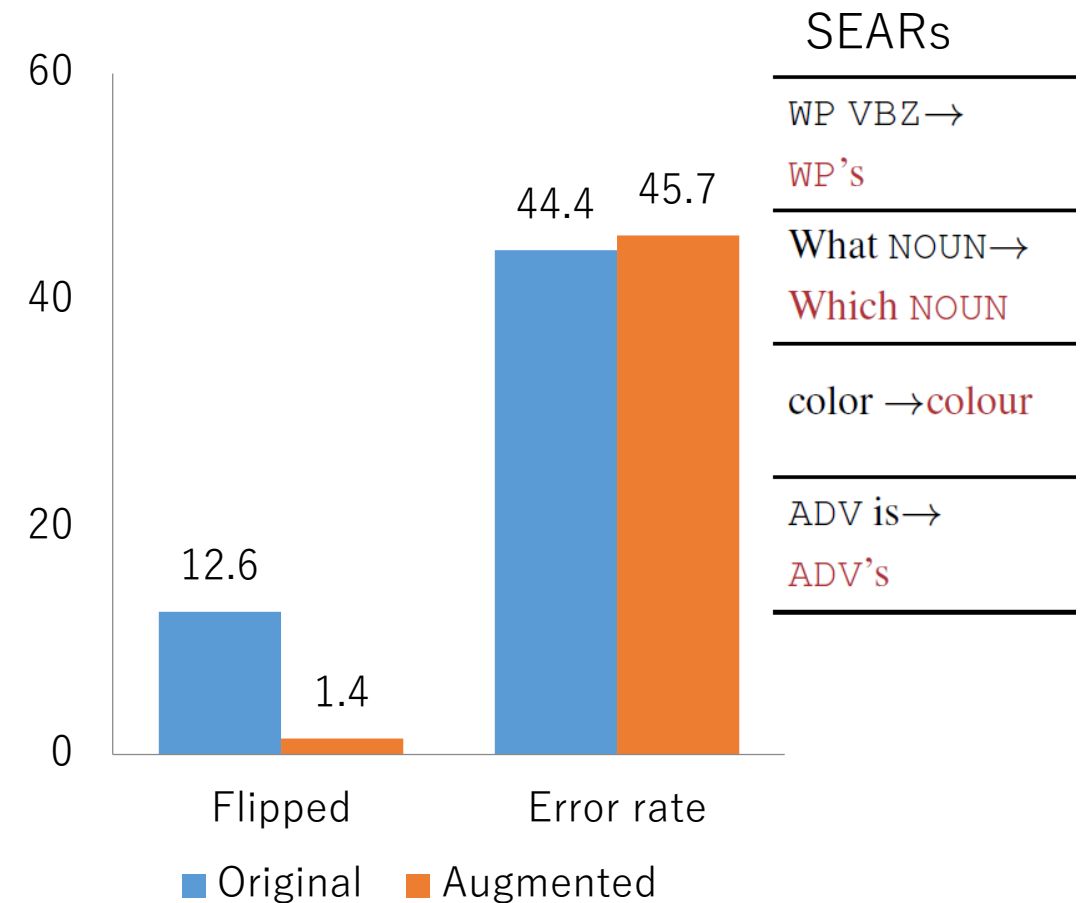


20 ルール作成にかかった時間



ルールで生成した例を用いた再学習

- 手順
 - 訓練データにSEARsを適用したデータを追加して, 再学習
 - 評価データにSEARsを適用したデータで評価
- モデルの精度を有意に下げることなく, 「バグ」を排除



まとめ

- 学習を用いたNLPモデルのデバッグ手法を提案
 - モデルを修正する例・ルールを（半）自動作成