

A Transformer-based Approach for Augmenting Software Engineering Chatbots Datasets

Ahmad Abdellatif
University of Calgary
Calgary, Canada

ahmad.abdellatif@ucalgary
.ca

Diego Elias Costa
Concordia University
Montreal, Canada

diego.costa@concordia.
ca

Khaled Badran
Concordia University
Montreal, Canada

k_badran@encs.conc
ordia.ca

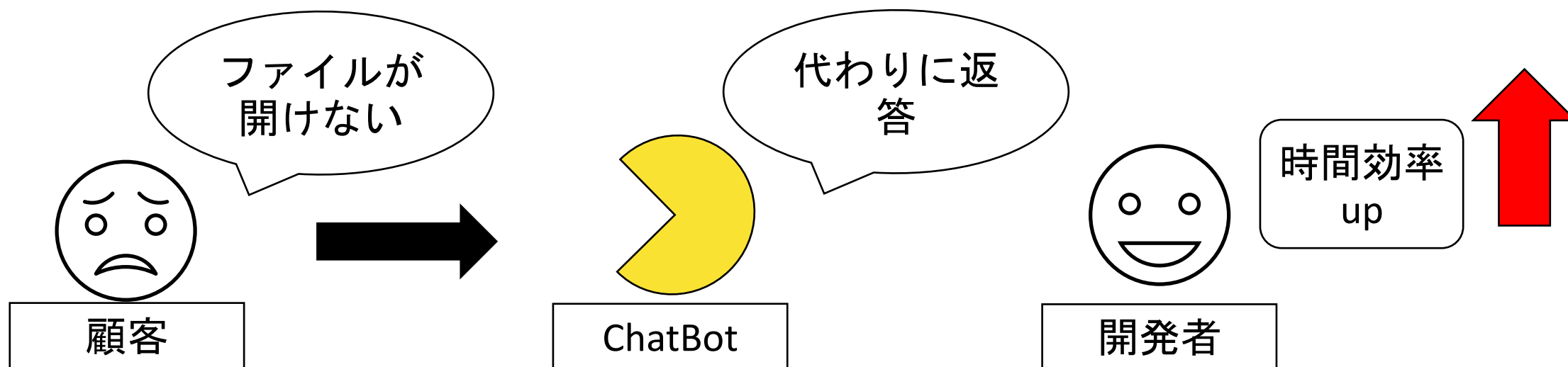
Emad Shihab
Concordia University
Montreal, Canada
emad.shihab@concor
dia.ca

発表者: 福井 遥陽

背景

近年、ChatBotは**時間とコスト**を節約できるという面から、SEタスクで開発者を支援することを目的に開発・研究されている。

例) マイクロサービスプロジェクトの構築や管理, カスタマーサポート

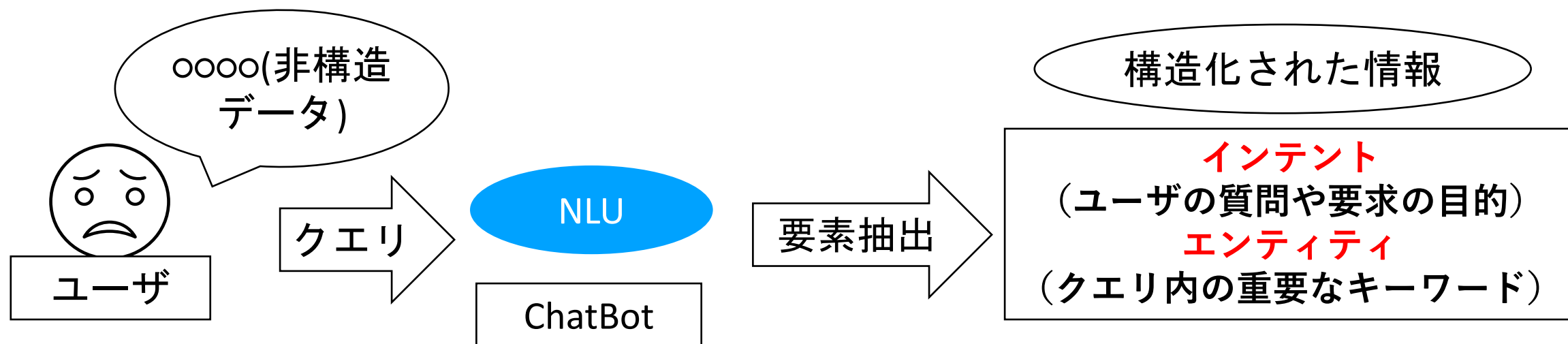


背景

ChatBot

ユーザと自然言語で対話を行うソフトウェアボット。主にユーザとサービスをつなぐ手段として活用。

→ **自然言語理解(NLU)**コンポーネントにより、ユーザからのクエリを理解



背景

チャットボットが正確に動作するには、NLUがクエリ情報を正確に抽出する必要がある。

→精度が低いとユーザの意図を誤認し、不適切な返答を行う

NLUは分類したインテントに対し、「信頼度スコア」を返す。

構造化された情報

インテント

(ユーザの質問や要求の目的)

エンティティ

(クエリ内の重要なキーワード)

インテント

信頼度スコア

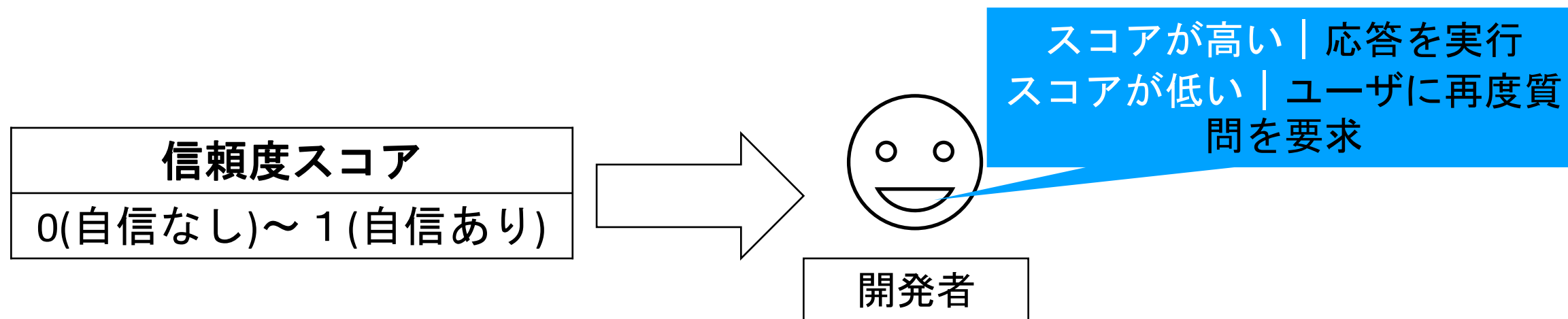
0(自信なし)～1(自信あり)

背景

チャットボットが正確に動作するには、NLUがクエリ情報を正確に抽出する必要がある。

→精度が低いとユーザの意図を誤認し、不適切な返答を行う

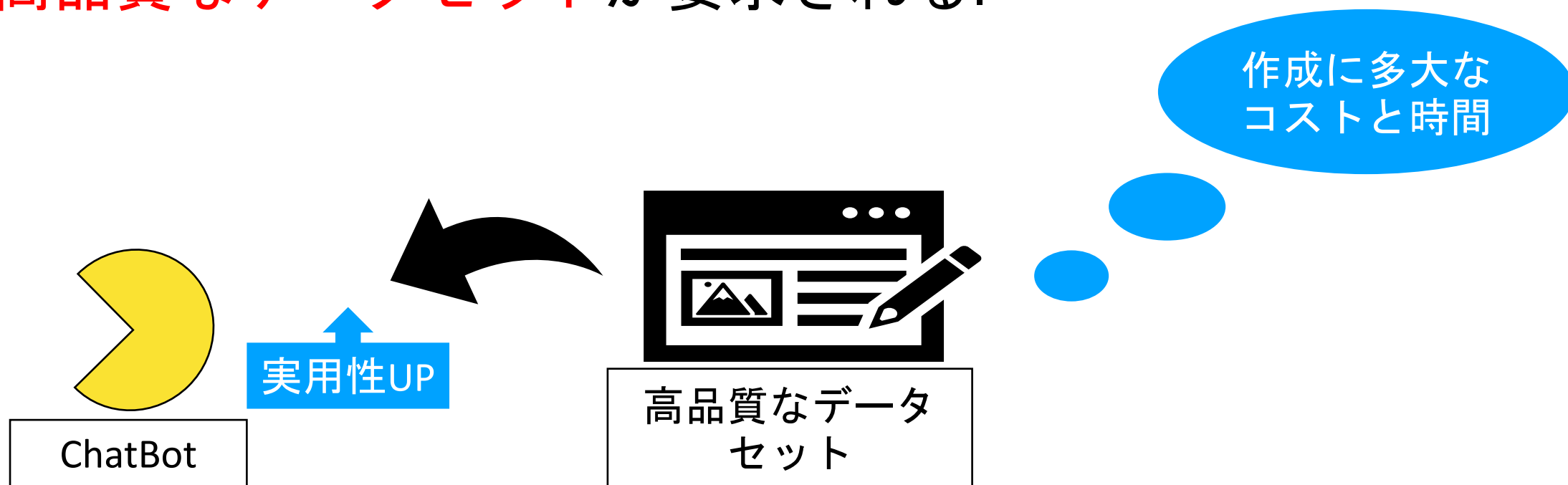
NLUは分類したインテントに対し、「**信頼度スコア**」を返す。



背景

実用性のあるチャットボットの活躍には高精度のNLUが必要不可欠

→ 高品質なデータセットが要求される。



目的

ソフトウェア工学チャットボットのデータセットを自動的に拡張するトランスフォーマーベースのアプローチを提案.

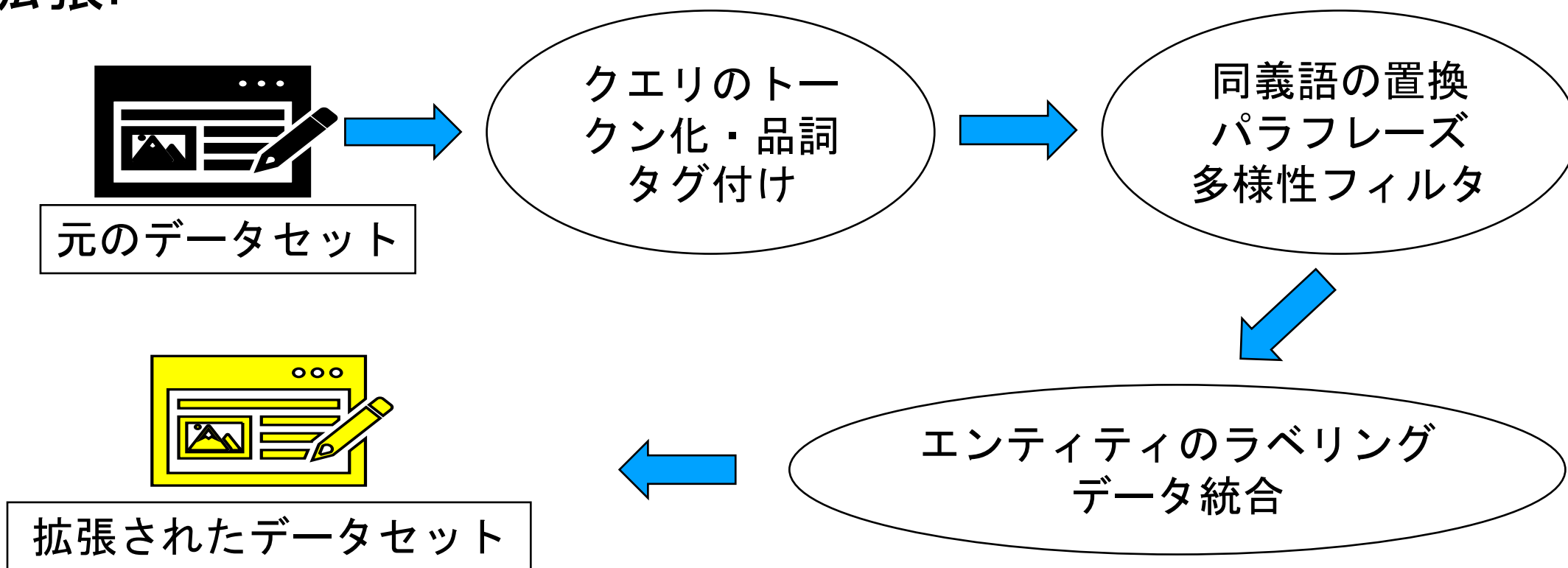
つまり、、、

チャットボット向けの高品質なデータセットを多大な時間とコストをかけずに作成することを目的とする.

本研究では質ではなく量の改良に焦点を当てる.

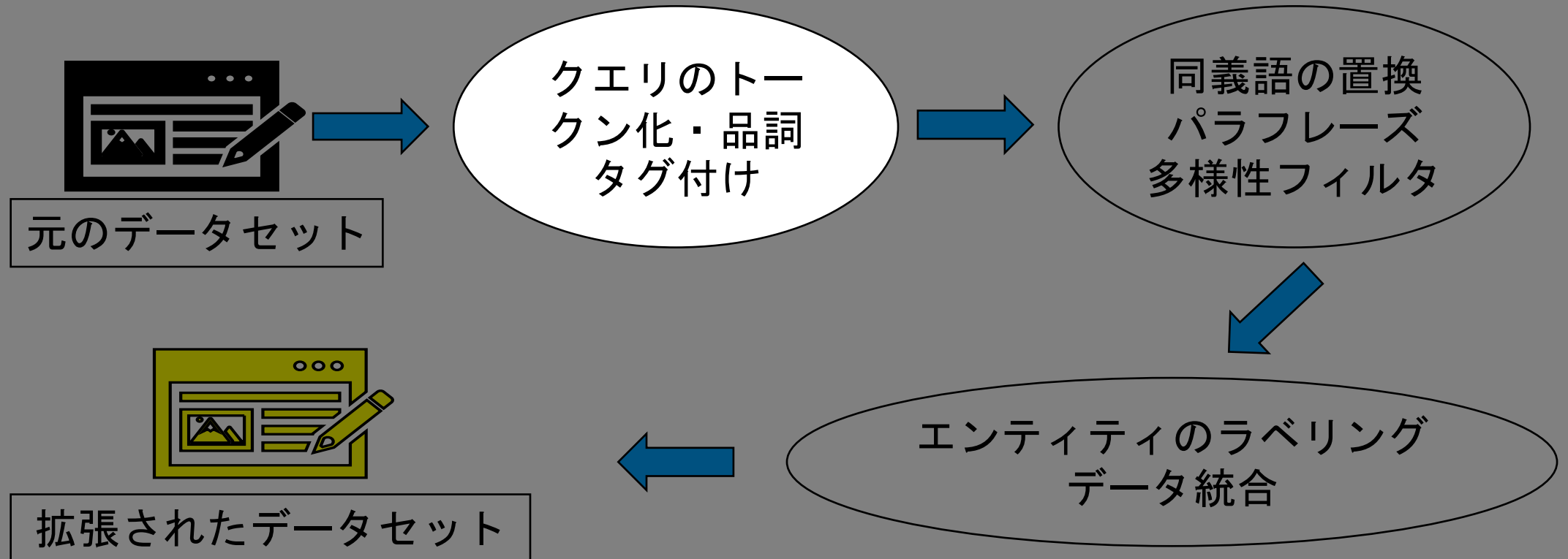
提案手法

元のトレーニングセットを入力として、意味を保持しつつ用語や文構造が異なる新しいクエリを生成し、データセットを拡張.



提案手法

データの前処理



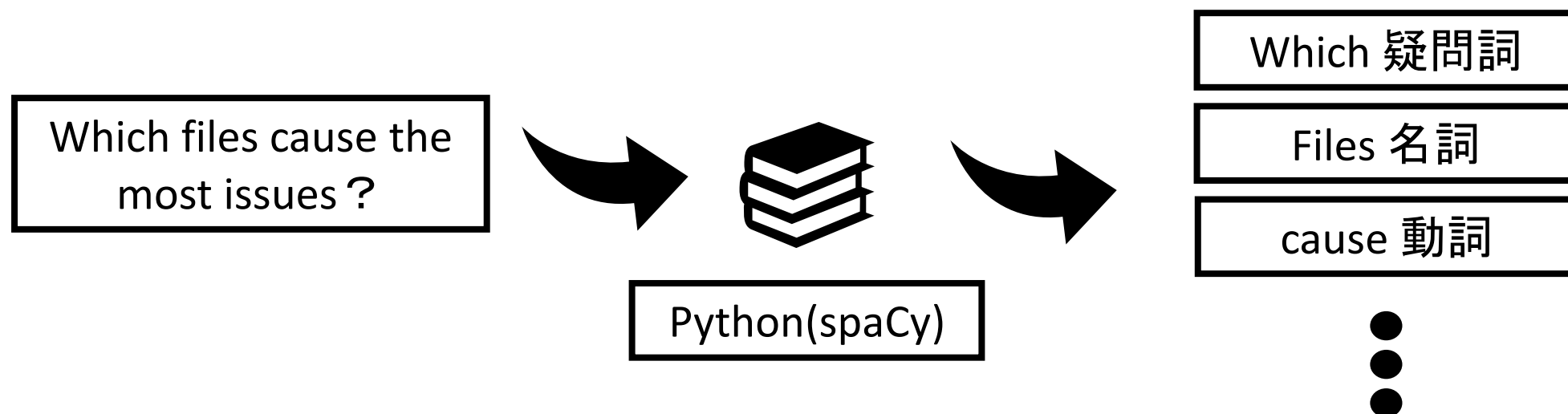
提案手法

トークン化

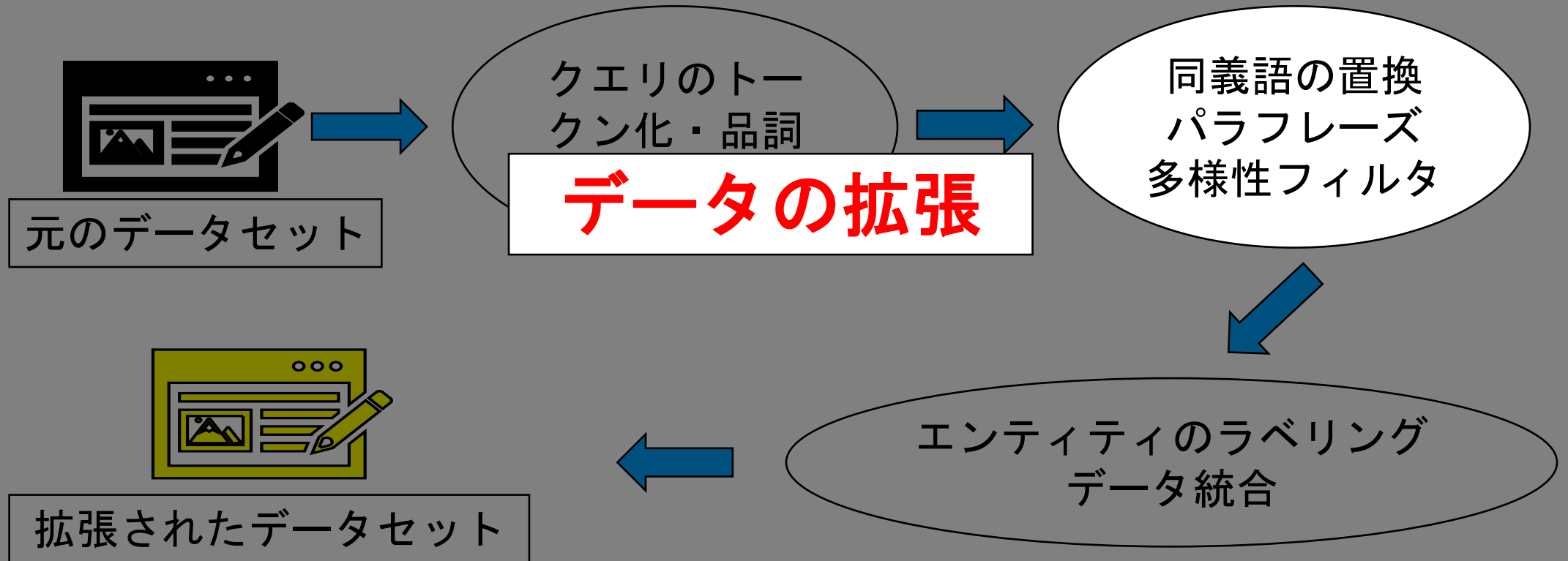
| 入力データを単語単位に分割

品詞タグ付け

| 分割したトークン（単語）に対して品詞を特定



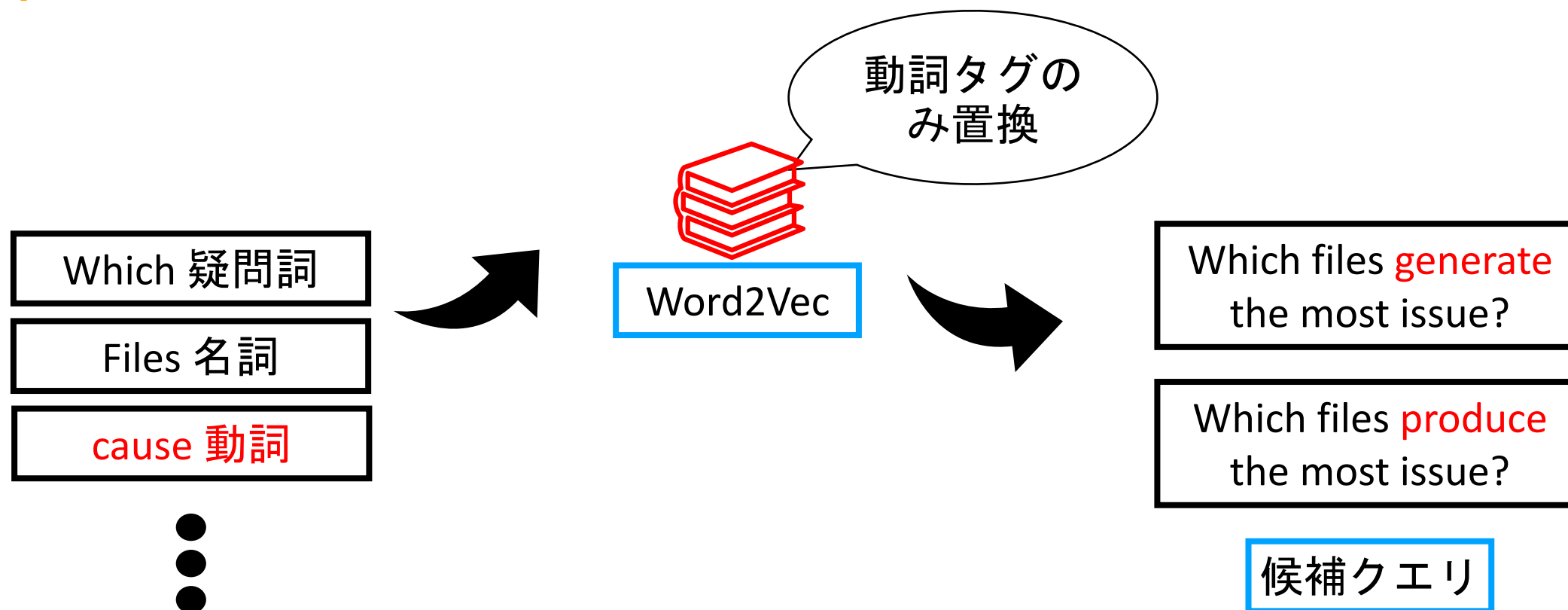
提案手法



提案手法

同義語の置換

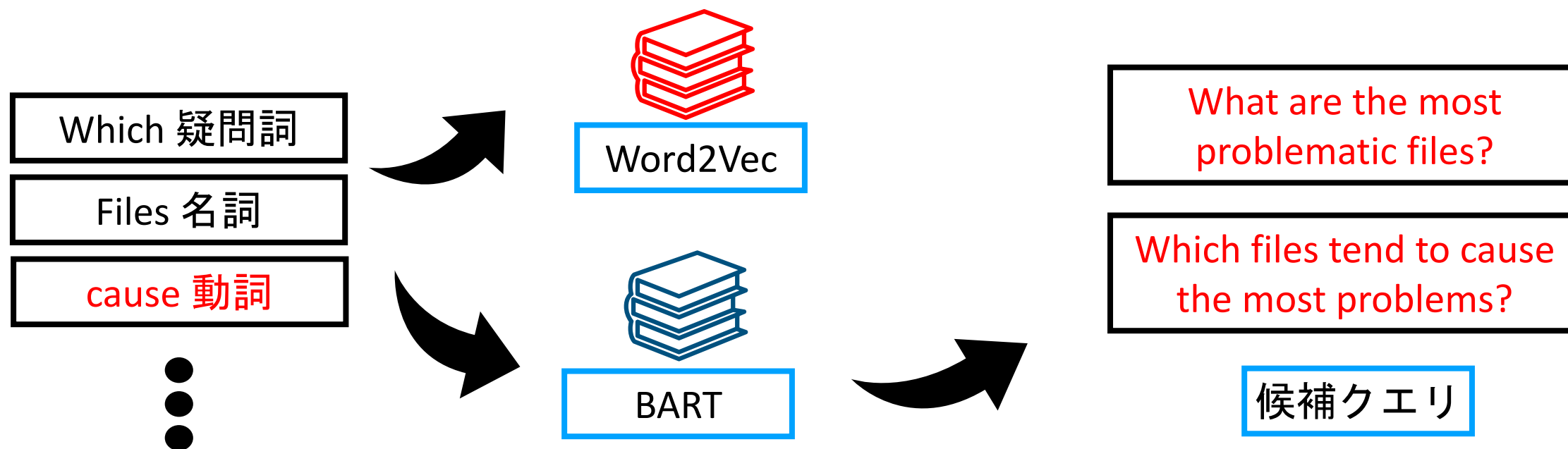
同義語の置換により，新しく「候補クエリ」を生成



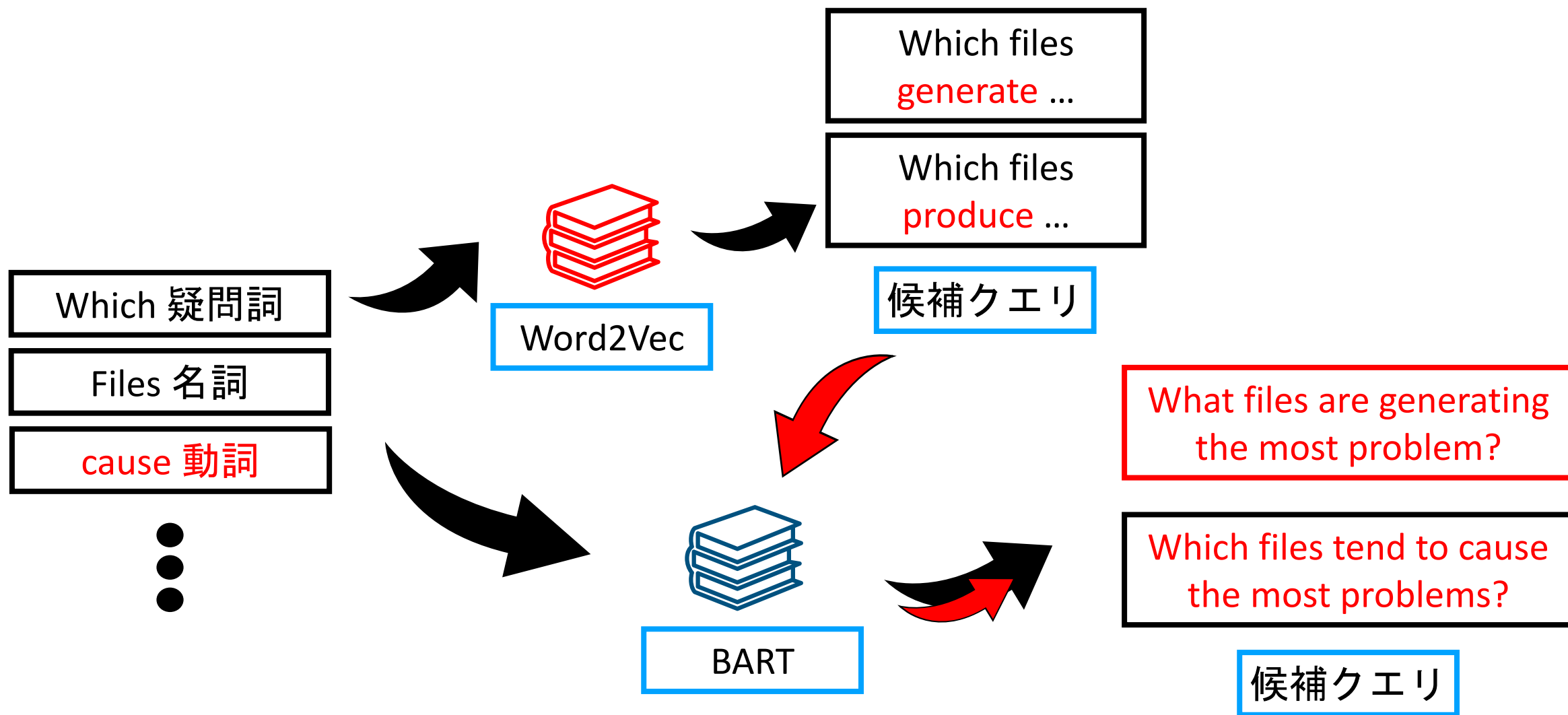
提案手法

パラフレーズ

文構造そのものを変更した「候補クエリ」を生成



提案手法

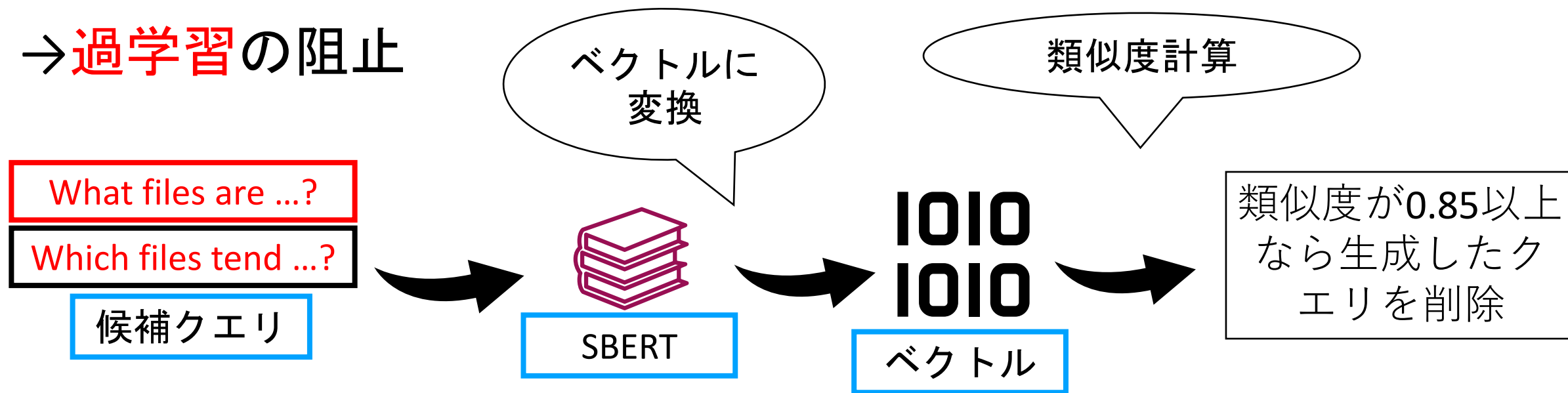


提案手法

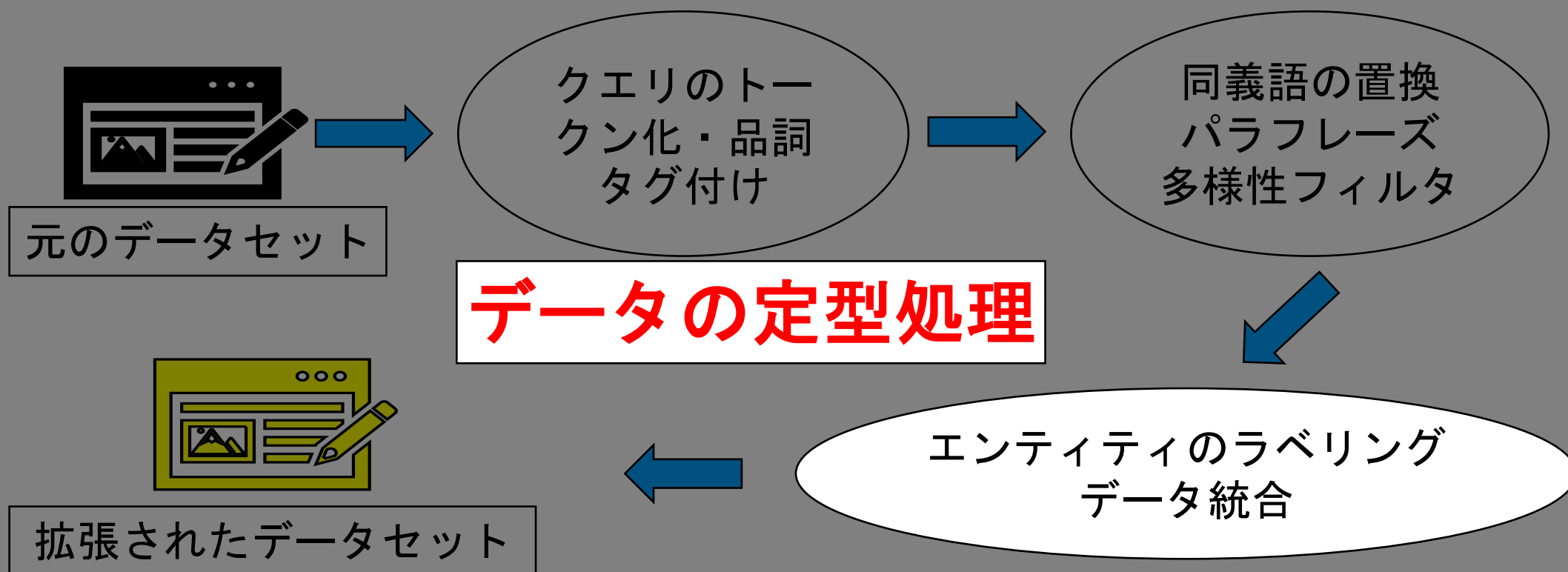
多様性フィルタ

生成した候補クエリから，元のクエリと意味が近すぎるクエリを除去．

→ **過学習**の阻止



提案手法



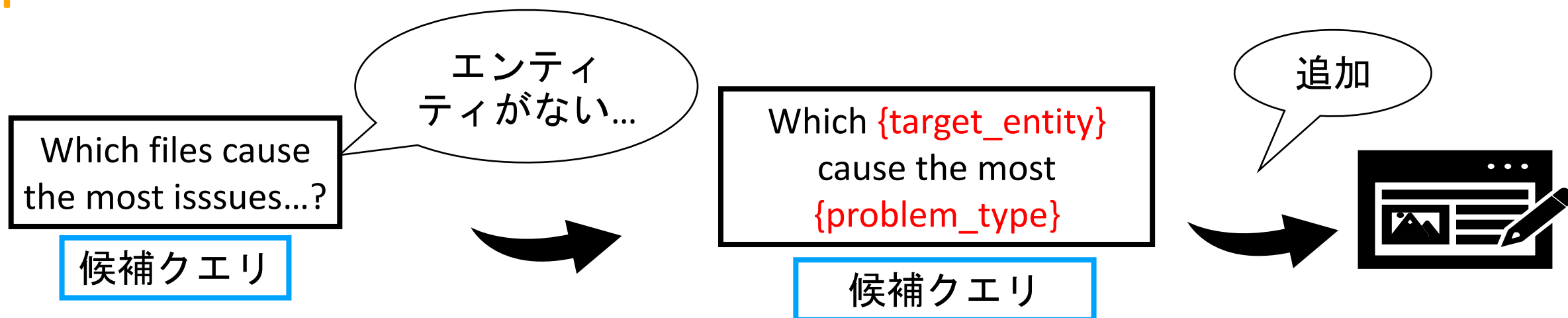
提案手法

エンティティのラベリング

生成クエリにエンティティをルールベースで自動ラベリングを行う.

データ統合

拡張されたクエリを元のトレーニングセットに追加



評価設定

用いるデータセット

- Repository....10種類のインテント, 398クエリ

- Ask Ubuntu....4種類のインテント, 154クエリ

- Stack Overflow....5種類のインテント, 215クエリ

モデルの訓練, 評価

- Rasa(ver2.5)

評価設定

パラフレーズ生成に用いるBARTのファインチューニング

- Quora Question Pairs

- Microsoft Research Paraphrase Corpus

- Paraphrase Adversaries from Word Scrambling

使用モデル

- BART-large

評価設定

評価シナリオ

- 1. 1クエリ/インテント
- 2. 3クエリ/インテント
- 3. 5クエリ/インテント

性能評価指標

- 適合率 $\text{正しく分類された数} \div \text{分類された全数}$
- 再現率 $\text{正しく分類された数} \div \text{実際の正解数}$
- F1スコア $(2 \times \text{適合率} \times \text{再現率}) \div (\text{適合率} + \text{再現率})$

結果・RQ1

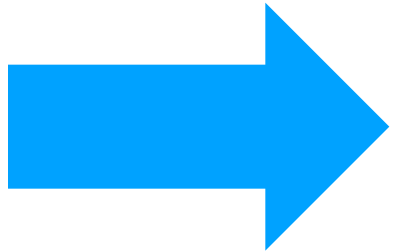
拡張手法はNLUの性能を向上させるか？

- 1.元のトレーニングデータのみ使用
- 2.今回の手法で拡張したデータを使用
- 3.人間が手動で拡張したデータを使用

Dataset	No. of Queries/Intent	Baseline F1-score	Augmentation Approach			Human	
			F1-score	% Improvement	% Optimal	F1-score	% Improvement
Repository	One	43.7	44.7*	2.3	6.9	58.1*	33.0
	Three	62.8	64.3*	2.4	34.1	67.2*	7.0
	Five	66.4	68.5*	3.2	60.0	69.9*	5.3
Ask Ubuntu	One	71.6	73.5*	2.7	17.3	82.6*	15.4
	Three	84.1	83.6*	-0.6	-	90.4*	7.5
	Five	87.1	84.2*	-3.3	-	90.2*	3.6
Stack Overflow	One	32	32.2	0.6	2.5	40*	25
	Three	36.9	37.9*	2.7	12.3	45*	22.0

結果・RQ1

- | 拡張手法は通常時に比べて， NLUの性能を2%向上させた.
- | 基本的に人手による拡張のほうが常に高い性能を出すが，実験設定によっては60%の最適性能を見せた.

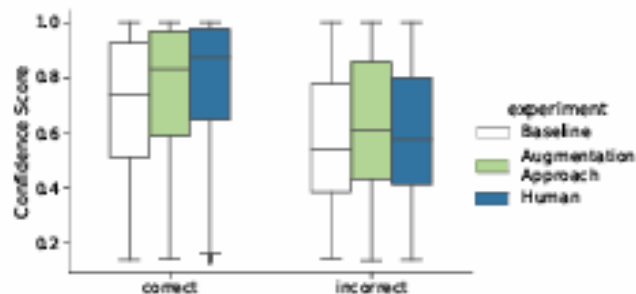


小さな改善だが， NLUの性能を
自動で向上させることが可能

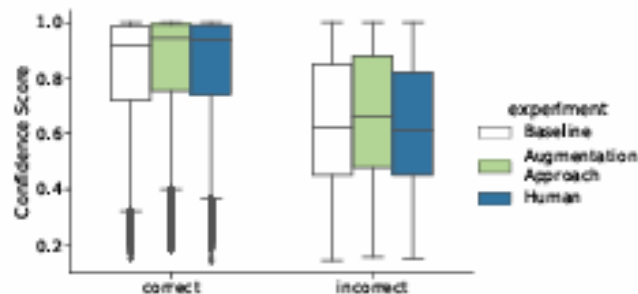
結果・RQ2

拡張手法はNLUの信頼度を高めるか？

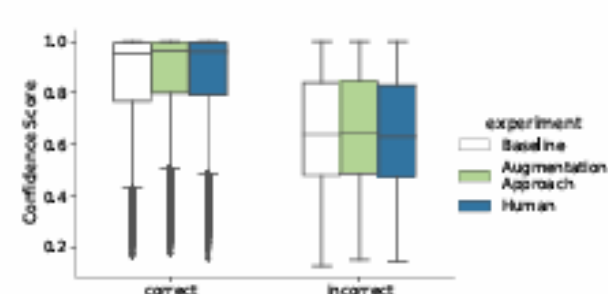
前述した信頼度スコアに対して，拡張手法が与える影響を確認．つまり，誤分類されたインテントと正しく分類されたインテントを比較



(a) 1 query/intent



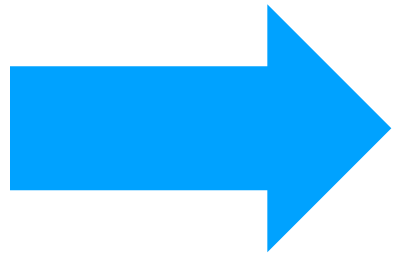
(b) 3 queries/intent



(c) 5 queries/intent

結果・RQ2

- 拡張手法を使うことで、正しく分類されるインテントの信頼度スコアが向上
- 一方で、誤分類されたインテントに対しても信頼度が上がってしまった。



どちらの親鸞緯度スコアも上がる
ことから、閾値調整が重要になる

考察

結果から、小数のトレーニングクエリからでも、ある程度の性能を発揮.

→データ量ではなく、データの質が重要

今回使用したパラフレーズ作成用のデータセットはSE特化ではなかったため、専用のパラフレーズデータセットが必要.

今後の展望

| SE特化のパラフレーズデータセットを構築

| 誤分類の信頼度上昇を抑制する改良

| 拡張データによるインテント間の干渉を緩和する仕組みの導入

所感

本論文ではデータセットそのものを改善することに焦点を当てたものであり、モデルの改良するなどchatBotの仕組みを改良するものではなかった。

→自分の研究分野ではデータセットの改良を目的とした論文がほとんどないため、新たな着想を得ることが出来た。