

Good things come in three: Generating SO Post Titles with Pre- Trained Models, Self Improvement and Post Ranking

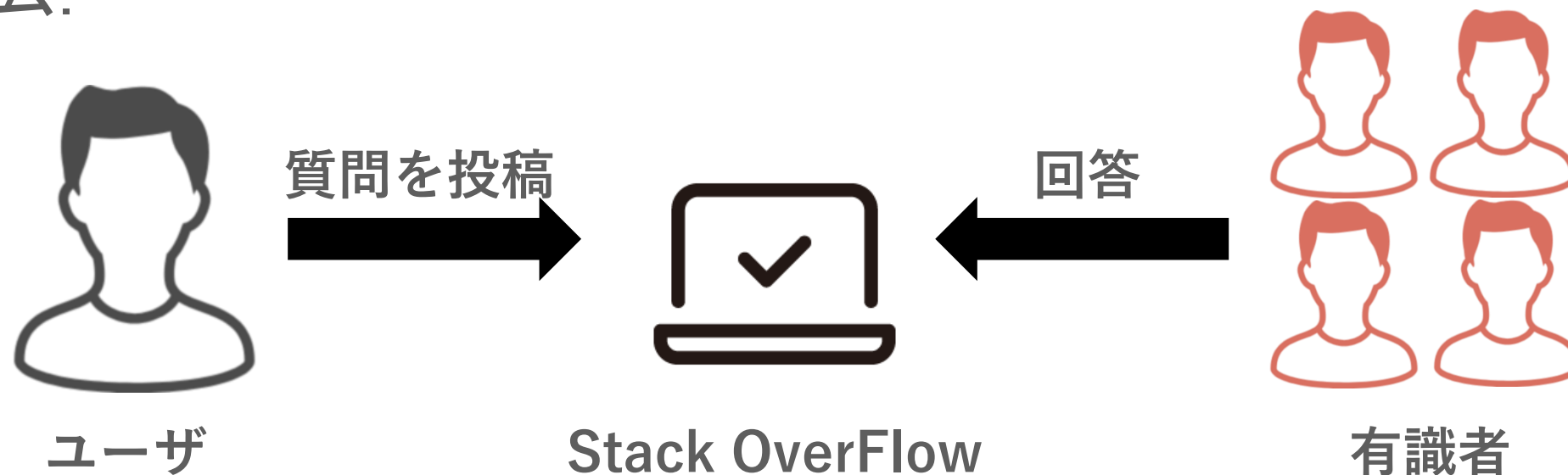
Duc Anh Le, Anh M. T. Bui, Phuong T. Nguyen,
Davide Di Ruscio, 21 Jun 2024.

ESEIW 2024 ESEM Technical Papers

Stack Overflow (SO) とは

ソフトウェア開発のためのQ&Aフォーラム

開発者がプログラミング関連の質問や解決策を探すためのプラットフォーム。



課題：質問に対して回答が少ない。

タイトルに問題が十分に記載されていない場合が多いため、内容に適したタイトルを質問につけることが効果的であると考えられる。

Stack Overflow (SO) とは

既存研究では

事前学習済みモデル(CodeT5)を使用して, コードや問題記述からタイトルを自動生成していた.

▶データの質や生成モデル固有の制約により, 質の高いタイトルを得ることは依然として困難.

そこで...

ファインチューニングされた言語モデルに自己改善と投稿ランキング機能を組み合わせた**FILLER**を提案する.

- ・ 自己改善

ファインチューニングされたモデル自身の予測を訓練データに組み込むこと.

- ・ 投稿ランキング

推論中に複数のタイトルを生成し, 最も関連性の高いものを選択する.

SOとCodeT5

現状のCodeT5利用における課題

①露出バイアス

モデルの訓練時と推論時の動作の違いから生じる課題.

訓練時：正しい単語が次の単語を予測する入力

推論時：モデルの予測が入力 ➤ 初期の誤差が連鎖的に増幅される

②生成品質のばらつき

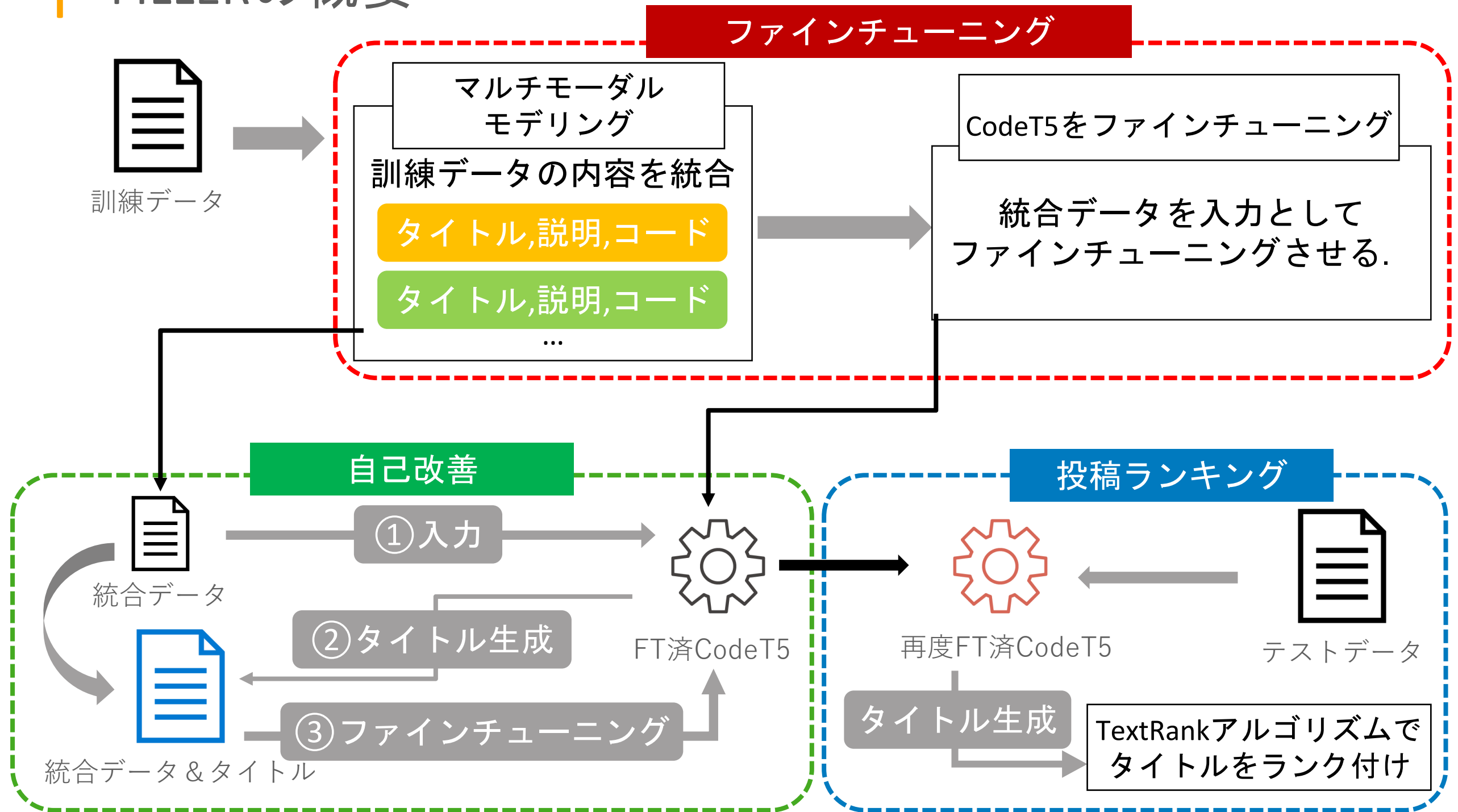
生成モデルのランダム性により品質に大きなばらつきが生じる問題.

Ex. CodeT5^[1]が生成した30個のタイトルは内容の関連度が40~99%であった.

[1] Yue Wang, Weishi Wang, Shafiq Joty, and Steven CH Hoi. 2021. CodeT5: Identifier-aware Unified Pre-trained Encoder-Decoder Models for Code Understanding and Generation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 8696–8708.

FILLER

FILLERの概要



FILLER

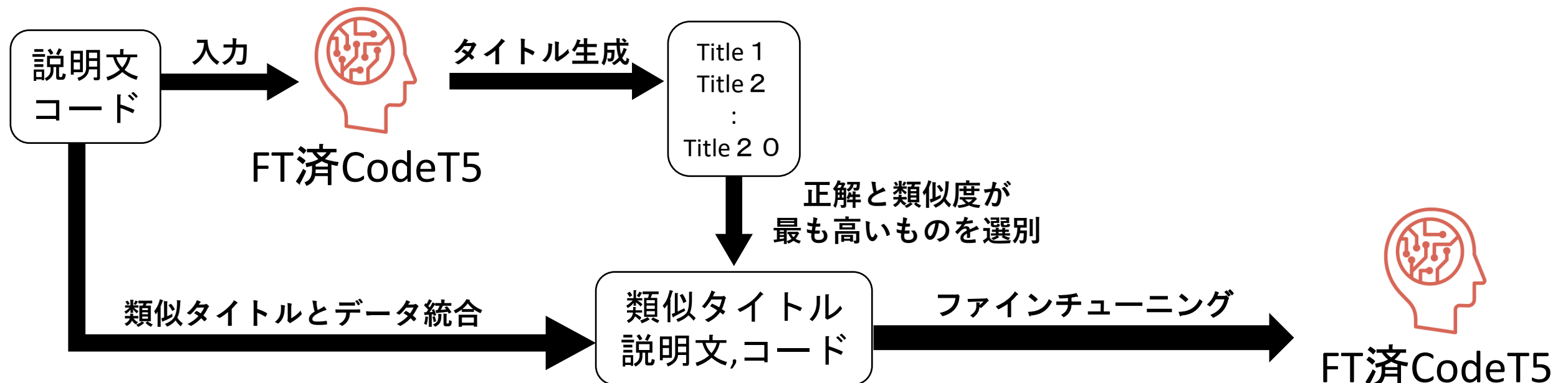
ファインチューニングプロセス

実際の投稿データでCodeT5をファインチューニングする。



自己改善プロセス

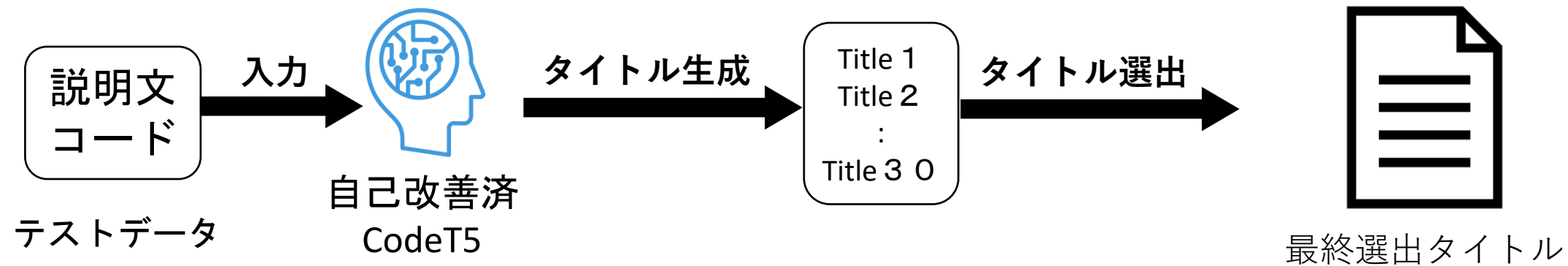
モデル自身の予測を入力データに統合し，再度ファインチューニングを行う。



FILLER

投稿ランキングプロセス

テストデータから複数のタイトル候補を生成し，その中から最も関連性の高いタイトルを選択する．



タイトルの選出手法は以下の通り

- ①各タイトルをノードとするグラフを生成.
- ②全タイトル間に重みを付け，類似度を計算.
- ③各タイトルに関連性スコアを割り当てる.
- ④最も高いスコアのタイトルを選出.

評価設定とRQ

データセット

既存研究^[2]のベンチマークデータセットを使用.

284298件の投稿データが含まれる（訓練，検証，テスト用）

対象言語はJava, C#, Python, JavaScript.

リサーチクエスチョン

RQ1: 事前学習済みモデルはSOのタイトル生成に影響するか？

RQ2: FILLERは最先端ツールと比較してどの程度効果があるか？

RQ3: FILLERのどの要素がその有効性に貢献しているか？

RQ4: FILLERの生成したタイトルは開発者にどう評価されるか？

[2] Ke Liu, Guang Yang, Xiang Chen, and Chi Yu. 2022. SOTitle: A Transformer-based Post Title Generation Approach for Stack Overflow. In 2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER). IEEE Computer Society, 577–588.

RQ1：事前学習モデルの比較

T5, BART, CodeBERT, UnixCoder, CodeT5を比較

生成されたタイトルと正解のタイトル間の重なりを比較, 計算.

単一の単語(R-1), 2つの単語(R-2), 最も長い文章(R-L)の3つの評価に分けて比較を行う.

結論

C#, Python, JavaScriptにおいてCodeT5は他のモデルを全ての評価指標で上回った.

Model	Java			C#			Python			JavaScript			Average		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
T5 ₂₅₆₊₂₅₆	26.49	9.80	24.60	26.94	10.76	25.17	29.00	10.79	26.70	28.53	10.77	26.47	27.74	10.53	25.74
T5 ₅₁₂	28.72	11.14	26.59	29.03	12.07	27.21	31.39	12.37	28.87	30.96	12.65	28.80	30.03	12.06	27.87
BART	28.40	10.96	26.35	28.48	11.80	26.72	31.25	12.29	28.79	30.62	12.16	28.51	29.69	11.80	27.59
CodeBERT	25.22	9.31	23.62	25.17	10.28	23.85	28.16	10.75	26.36	27.49	10.67	25.85	26.51	10.25	24.92
UnixCoder	25.28	9.34	23.58	25.54	10.49	24.17	27.49	10.00	26.10	27.84	10.66	26.08	19.73	7.65	18.52
CodeT5	28.64	11.10	26.53	29.24	12.38	27.31	31.94	12.71	29.40	31.23	12.74	29.16	30.26	12.23	28.10

RQ2：既存研究ツールの比較

既存研究ツールとFILLERを比較

Stack Overflowの投稿タイトルを生成する提案アプローチの有効性を評価するため、FILLERを4つのツールと比較した。

結論

どの種類のコードにおいても一貫してFILLERが上回った性能を見せた。

Model	Java			C#			Python			JavaScript			Average		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Code2Que [9]	25.25	9.17	23.57	26.55	10.57	25.01	27.93	10.36	25.87	27.50	10.40	25.70	26.81	10.13	25.04
CCBERT [47]	26.25	9.65	24.57	26.00	10.76	24.72	28.55	10.69	26.54	28.34	10.83	26.64	27.29	10.48	25.62
SOTitle [21]	26.49	9.80	24.60	26.94	10.76	25.17	29.00	10.79	26.70	28.53	10.77	26.47	27.74	10.53	25.74
M3NSCT5 [46]	27.90	10.22	25.52	28.77	11.55	26.53	31.35	11.80	28.44	30.39	11.65	27.96	29.60	11.31	27.11
GPT3.5-turbo	25.68	7.62	22.52	25.21	7.80	22.69	28.48	8.92	25.06	27.15	8.19	24.11	26.63	8.13	23.60
FILLER	30.48	11.52	27.94	30.89	12.81	28.60	33.64	13.00	30.63	33.14	13.24	30.55	32.04	12.64	29.43

RQ2：既存研究ツールの比較

生成例の比較

元のタイトル：Equivalent of Java's SortedMap.tailMap in C# SortedDictionary

SOTitle：How to copy / mimic tailMap in C#

M3NSCT5：How to copy a Dictionary <IComparable, value> to SortedDictionary <k, v> in C#

FILLER：C# equivalent of Java's SortedMap.tailMap

►FILLERのみ、**Javaの機能との対比**という重要なニュアンスを捉えている。

Model	Java			C#			Python			JavaScript			Average		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Code2Que [9]	25.25	9.17	23.57	26.55	10.57	25.01	27.93	10.36	25.87	27.50	10.40	25.70	26.81	10.13	25.04
CCBERT [47]	26.25	9.65	24.57	26.00	10.76	24.72	28.55	10.69	26.54	28.34	10.83	26.64	27.29	10.48	25.62
SOTitle [21]	26.49	9.80	24.60	26.94	10.76	25.17	29.00	10.79	26.70	28.53	10.77	26.47	27.74	10.53	25.74
M3NSCT5 [46]	27.90	10.22	25.52	28.77	11.55	26.53	31.35	11.80	28.44	30.39	11.65	27.96	29.60	11.31	27.11
GPT3.5-turbo	25.68	7.62	22.52	25.21	7.80	22.69	28.48	8.92	25.06	27.15	8.19	24.11	26.63	8.13	23.60
FILLER	30.48	11.52	27.94	30.89	12.81	28.60	33.64	13.00	30.63	33.14	13.24	30.55	32.04	12.64	29.43

RQ3：FILLERの要素の有効性

自己改善，投稿ランキングに有意性はあるのかを調査

自己改善，投稿ランキングをしない場合のFILLERと比較を行い，パフォーマンスの向上に貢献している点を分析する．

結論

どちらの手順を抜いても精度が低下したため，全ての要素においてタイトルの精度向上に貢献していることが分かった．

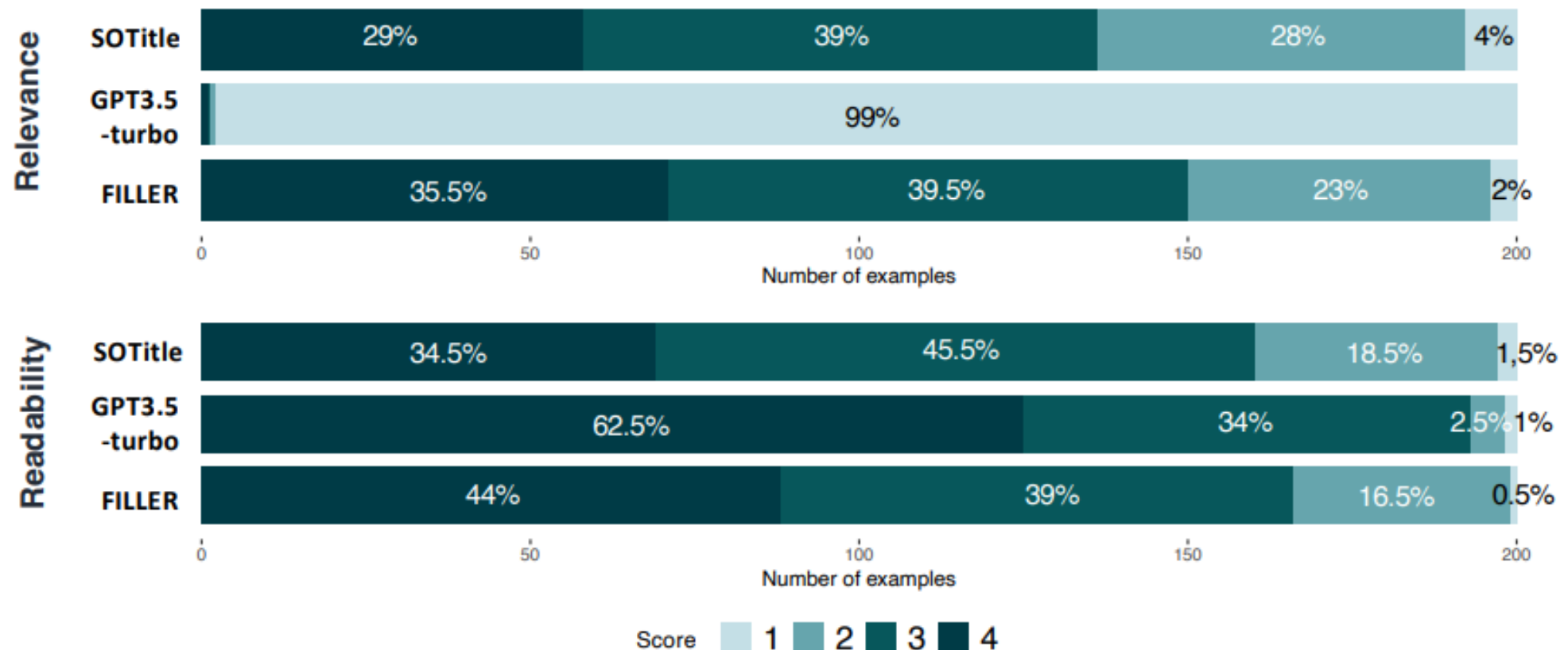
Model	Java			C#			Python			JavaScript			Average		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
FILLER_{w/o SI+PR}	28.64	11.10	26.53	29.24	12.38	27.31	31.94	12.71	29.40	31.23	12.74	29.16	30.26	12.23	28.10
FILLER_{w/o PR}	28.93	11.26	26.89	29.27	12.46	27.45	32.17	12.78	29.64	31.77	12.89	29.59	30.54	12.35	28.39
FILLER_{w/o SI}	29.71	11.29	27.35	29.86	12.46	27.78	33.05	12.97	30.20	32.34	12.90	29.95	31.24	12.41	28.82
FILLER	30.48	11.52	27.94	30.89	12.81	28.60	33.64	13.00	30.63	33.14	13.24	30.55	32.04	12.64	29.43

RQ4：開発者による評価

FILLERと既存ツール2つでユーザ調査を実施
関連性と読みやすさについて4段階評価を行った。

結論

FILLERがどの評価においても最高の評価を得た。



まとめ, 展望

Stack Overflowの質問タイトルを自動生成するツール

ファインチューニングに加えて, 自己改善, 投稿ランキングの手順を加えることで従来よりも高性能なツールが作成できた.

今後の展望

データ拡張

他のプログラミング言語やGitHubなどのデータソースでも検証してみる.

手法の改良

より高度なランキングメカニズムを組み込み推薦性能を向上させる.