

GIRT-Model: Automated Generation of Issue Report Templates

Nafiseh Nikeghbal[†]

Independent Researcher

Munich, Germany

lastname.firstname@gmail.com

Amir Hossein Kargaran

Ludwig Maximilian University

Munich, Germany

amir@cis.lmu.de

Abbas Heydarnoori

Bowling Green State University

Bowling Green, USA

aheydar@bgsu.edu

2024 IEEE/ACM 21st International Conference on Mining Software Repositories (MSR), Pages 407- 418

発表者：野口 優輝

背景

ソフトウェア開発では、多くのissueが報告される

- 問題点) 多くのIssue報告が整理されていない

→Issue管理が複雑になり、開発者の負担が増加[2][43]

GitHubなどは、課題報告テンプレート（IRT）を導入

- より早くissueが解決され、コメントのカバレッジも高い[25]
- 問題点：IRTを採用するリポジトリは5%未満[33]

原因) どのテンプレートや情報を載せるべきかが理解できていない

支援言語型モデルツールのGIRT-Modelを提案

[2] Nicolas Bettenburg, Sascha Just, Adrian Schr.ter, Cathrin Weiss, Rahul Premraj, and Thomas Zimmermann. 2008. What makes a good bug report?. In 16th ACM SIGSOFT International Symposium on Foundations of software engineering. 308–318.

[43] Mozhan Soltani, Felienne Hermans, and Thomas B.ck. 2020. The significance of bug report elements. Empirical Software Engineering 25, 6 (2020), 5255–5294.

[25] Hongyan Li, Meng Yan, Weifeng Sun, Xiao Liu, and Yunsong Wu. 2023. A firstlook at bug report templates on GitHub. Journal of Systems and Software 202(2023), 111709.

[33] Nafiseh Nikeghbal, Amir Hossein Kargaran, Abbas Heydarnoori, and Hinrich Schütze. 2023. GIRT-Data: Sampling GitHub Issue Report Templates. arXiv preprint arXiv:2303.09236 (2023).

目的とアプローチ

目的

- 開発者の指示に基づいたIRTを自動生成するGIRT-Modelを提案し、IRTの採用率を向上させること

アプローチ

- IRT生成のための指示データセット「GIRT-Instruct」を構築し、GIRT-Modelの性能をベースラインモデルと比較

GIRT-Instructの構築

収集したIRTのフィルタリング

Nikeghbalらの提供するGIRT-Data^[33]を使用

- 50,032のリポジトリがIRTをサポート
→MarkDown形式のIRTを採用(98,550件)

採用理由^[33]

- MarkDown形式のIRTの設定が容易
- 別のプラットフォームであるGitLabもMarkDown形式のみ採用している

前処理

- Nullのフィルタリング, 匿名化, 正規化, 重複削除, 非ラテン文字のフィルタリング
→51363件のIRTに絞る

GIRT-Instructの構築

インストラクションデータの生成

インストラクションデータ

- 特定の指示や命令（入力）と対応するIRT（出力）のペア
 - 各ペア：＜インストラクション（入力）＞＜IRT（出力）＞
 - インストラクション：(name, about, title, labels, assigness, **summary**)
- 本研究はsummaryの有無で2種類のインストラクションを生成
 - Zephyr-7B-Bata言語モデル^[46]を使って要約を生成
 - GIRT-Modelが開発者の意図を自然言語でカスタマイズする能力を調査

マスクインストラクション

- インストラクション内を＜MASK＞トークンで2つマスク
 - モデルに柔軟性を持たせるため

最終的なデータセット

$$51,363 \times 2 \times 2 = 205,452$$

GIRT-Modelの設定

アーキテクチャ

- 事前学習済みモデルのT5^[37]モデルを使用

ファインチューニング

- Sequence to Sequence (Seq2Seq) 学習アプローチを使用
- 指示データセット「GIRT-Instruct」を用いる

訓練設定

- エポック数：30
- バッチサイズ：8
- 学習率：5e-5

実験設定

データセットの分割

- 80 : 10 : 10 = 訓練データ : 検証データ : テストデータ

テストセット

テストセット	説明
META	GIRT-Modelのメタデータのみ
META + MASK	指示の一部がランダムにマスク
META + SUM	要約フィールドを含む
META + MASK + SUM	要約フィールドを含み、指示がマスク

ベースライン

- T5^[37]とFlan-T5^[7] (0.2B, 0.8B, 3B) で比較
- 評価：ゼロショット学習とワンショット学習

実験評価

自動評価

- ROUGE^[27], BLUE^[36], METEOR^[1]

人間評価

- IRTの概念に詳しく課題管理ツールなどを5年以上経験したSE
- Flan-T5-Largeのワンショットが生成したIRTを使用
- 評価項目：
 - 一貫性と整合性
 - 指示の遵守
 - フォーマット

}	0~5で評価
}	「はい」, 「部分的に」, 「いいえ」の3種類で評価

[27] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out. 74–81.

[36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In 40th Annual Meeting of the Association for Computational Linguistics. 311–318.

[1] Satyanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MTEvaluation with improved correlation with human judgments. In ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 65–72.

自動評価結果①

テストセットタイプ（一部抜粋）

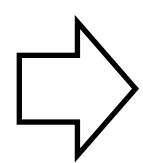
指示に詳細（要約）が含まれているか

		ROUGE-1
META	T5-3B	16.97
	Flan-T5-Large	13.30
	GIRT-Model	72.34
META + SUM	T5-3B	22.43
	Flan-T5-Large	17.10
	GIRT-Model	77.90

マスクされているか

		ROUGE-1
META	T5-3B	16.97
	Flan-T5-Large	13.30
	GIRT-Model	72.34
META + MASK	T5-3B	9.92
	Flan-T5-Large	10.35
	GIRT-Model	63.35

- マスクしていない方がマスクするよりも良いスコアを出した
 - モデルの柔軟性を高める一方で、出力の不一致を多く出した可能性



どの指標においても、ベースラインモデルより
比較的高いスコアを出した結果になった

自動評価結果②

Zero-shot vs One-shot（一部抜粋）

- 例) Flan-T5-Large, ROUGE

	Zero-shot	One-shot
META	13.30	27.95
META+SUM	17.70	42.06
META+MASK	10.35	18.23
META+SUM+MASK	13.48	39.47

- ワンショットはゼロショットよりも高いスコアを出した
 - ワンショットは文脈学習を可能にし，モデルが指示の目的に適合した
- ベースラインモデルはGIRT-Modelには及ばない
 - ユーザーはゼロショットでモデルとやりとりすることが多く，IRTの概念に詳しくない人にとっても，ゼロショットに対応したモデルが重要

人間評価結果

Test Set Type	Aspect	Score %	
		Flan-T5-Large (One-Shot)	GIRT-Model (Zero-Shot)
META	Consistency and Coherence	6.4	90.8
	Instruction Following	13.6	94.0
	Format (Yes, Partially)	0.0, 4.0	100.0, 0.0
META + MASK	Consistency and Coherence	8.4	84.4
	Instruction Following	11.2	88.4
	Format (Yes, Partially)	2.0, 6.0	100.0, 0.0
META + SUM	Consistency and Coherence	24.0	91.6
	Instruction Following	26.4	90.0
	Format (Yes, Partially)	40.0, 12.0	100.0, 0.0
META + SUM + MASK	Consistency and Coherence	22.4	85.2
	Instruction Following	24.4	88.8
	Format (Yes, Partially)	38.00, 2.00	100.0, 0.0

Consistency and Coherence

→一貫性と整合性

Instruction Following

→指示の遵守

Format (Yes, Partially)

→完全, 部分的に合っているか

- 指示に詳細（要約）が含まれるとGIRT-Modelはスコアが低い
 - 複雑な指示の場合, モデルが制約に従っているのかを人間が評価するのに難しさが生じるため
- マスクされていない方がスコアが高い
 - モデルに柔軟性が与えられると, 結果に一貫性のない部分が発生してしまうため

まとめ

目的

- 開発者の指示に基づいたIRTを自動生成するGIRT-Modelを提案し、ベースライン（T5, Flan-T5）と比較してどうか

結果

- GIRT-Modelはベースラインよりも比較的高いスコアを出した
- Zero-shotとOne-shotではOne-shotの方が性能が良かったが、いずれにせよ、GIRT-Modelには及ばなかった

今後の課題

- Markdown形式以外にも、YAML形式のIRTもサポート

補足：実験結果

ベースラインモデルとの比較

	Model	Size	Zero-Shot				One-Shot					Zero-Shot				One-Shot			
			ROUGE-1	ROUGE-L	BLEU	METEOR	ROUGE-1	ROUGE-L	BLEU	METEOR		ROUGE-1	ROUGE-L	BLEU	METEOR	ROUGE-1	ROUGE-L	BLEU	METEOR
META	T5-base	0.2B	0.05	0.05	0.00	0.02	16.83	14.60	2.23	9.55	META + MASK	0.07	0.06	0.00	0.03	9.80	8.91	0.88	6.03
	T5-Large	0.8B	1.39	1.35	0.00	0.66	10.56	8.41	1.51	7.03		0.83	0.78	0.00	0.30	10.74	8.27	1.77	7.42
	T5-3B	3B	16.97	15.44	0.44	9.28	21.29	18.02	3.60	15.03		9.92	8.95	0.33	5.98	18.03	<u>15.78</u>	2.73	13.38
	Flan-T5-base	0.2B	10.42	10.12	1.61	4.14	14.83	13.05	7.51	14.43		3.87	3.61	1.08	2.01	5.80	5.36	3.13	5.50
	Flan-T5-Large	0.8B	13.30	12.60	0.02	4.81	<u>27.95</u>	<u>21.88</u>	<u>14.11</u>	<u>25.36</u>		10.35	8.83	0.51	3.77	<u>18.23</u>	15.18	<u>7.52</u>	<u>16.61</u>
	Flan-T5-XL	3B	9.41	9.36	0.00	3.48	18.20	16.57	9.16	15.76		3.84	3.80	0.00	1.70	8.70	7.92	3.02	7.47
	GIRT-Model-base	0.2B	72.34	70.20	59.96	69.95	X	X	X	X		63.44	59.74	53.93	62.95	X	X	X	X
META + SUM	T5-base	0.2B	0.09	0.08	0.00	0.04	9.07	7.49	0.42	5.86	META + SUM + MASK	0.08	0.06	0.00	0.03	10.81	9.23	0.58	6.29
	T5-Large	0.8B	5.83	4.74	0.00	2.36	10.98	8.16	3.27	7.89		8.01	6.21	0.00	3.18	11.07	8.54	3.45	8.68
	T5-3B	3B	22.43	18.72	1.06	11.90	27.41	20.86	5.26	18.03		15.51	12.34	0.48	8.16	25.47	19.81	4.21	16.76
	Flan-T5-base	0.2B	16.93	14.66	0.23	6.24	27.60	23.43	14.19	28.63		15.23	12.53	1.02	6.03	26.51	22.18	13.27	28.19
	Flan-T5-Large	0.8B	17.10	13.56	0.01	6.13	<u>42.06</u>	<u>33.28</u>	<u>20.96</u>	<u>36.90</u>		13.48	10.50	0.00	4.58	<u>39.47</u>	<u>32.26</u>	<u>21.68</u>	<u>36.19</u>
	Flan-T5-XL	3B	17.97	15.16	0.01	7.03	18.17	15.19	6.36	15.71		16.15	12.93	0.08	6.96	20.75	17.74	9.44	17.61
	GIRT-Model-base	0.2B	77.90	75.33	62.94	74.38	X	X	X	X		73.92	70.32	57.12	69.85	X	X	X	X

補足：GIRT-Modelの例

```
name: Bug report
about: <|MASK|> ←モデルに判断を任せたい時
title: <|MASK|>
labels: <|MASK|>
assignees: <|EMPTY|> ←フィールドを明確的に空にしたい時
headlines_type: # Heading
headlines: ['Describe the bug', 'To Reproduce',
'Expected behavior', 'Screenshots
(if appropriate)', 'Environment', 'Additional
context']
summary: This issue template is designed to
help users report bugs they encounter while
using our software. Environment section should
asks for "operating system" such as Ubuntu.
```

```
---
name: Bug report
about: Create a report to help us improve
title: '[Bug]'
labels: bug
assignees: ''
---

## Describe the bug
A clear and concise description of what the bug
is.

## To Reproduce
Steps to reproduce the behavior:
1. Go to '...'
2. Click on '....'
3. Scroll down to '....'
4. See error

## Expected behavior
A clear and concise description of what you
expected to happen.

## Screenshots (if appropriate)
If applicable, add screenshots to help explain
your problem.

## Environment
- OS: [e.g. Ubuntu]
- Version [e.g. 22]

## Additional context
Add any other context about the problem here.
```

The screenshot displays the GIRT interface, which is used to generate GitHub issue templates. It is divided into two main sections: 'Settings' and 'Model Configs'.

Settings Section:

- Issue Template Inputs:** A dropdown menu showing 'bug report' (labeled 1).
- Name Metadata:** A text input field containing 'Bug report' (labeled 2).
- About Metadata:** A text input field containing 'e.g., File a bug report'.
- Title Metadata:** A text input field containing 'e.g., [Bug]:'.
- Labels Metadata:** A text input field containing 'e.g., feature, enhancement'.
- without title:** A checkbox.
- without labels:** A checkbox.
- without assignees:** A checked checkbox.
- How would you like to be your Headlines?:** A dropdown menu showing '# Heading' (labeled 3).
- Headlines:** A list of headlines: 'Describe the bug', 'To Reproduce', 'Expected behavior', 'Screenshots (if appropriate)', 'Environment', 'Additional context'.
- Summary:** A text area containing a summary of the issue template.

Model Configs Section:

- max_length:** A slider ranging from 30 to 512, currently set at 300 (labeled 4).
- min_length:** A slider ranging from 0 to 300, currently set at 30.
- top_p:** A slider ranging from 0.00 to 1.00, currently set at 0.92.
- top_k:** A slider ranging from 0 to 100, currently set at 0.

Generated Template Section:

The right side of the interface shows the generated GitHub issue template (labeled 5 and 6). It includes a header with the GIRT logo and a 'Submit' button. The template content is as follows:

```
name: Bug report
about: <|MASK|>
title: <|MASK|>
labels: <|MASK|>
assignees: <|EMPTY|>
headlines_type: # Heading
headlines: ['Describe the bug', 'To Reproduce', 'Expected behavior', 'Screenshots
(if appropriate)', 'Environment', 'Additional
context']
summary: This issue template is designed to help users report bugs they encounter
while using our software. Environment section should asks for "operating system"
such as Ubuntu.
```

The template also includes a 'Submit' button and a 'Duplicate Space' button.