

Stay Professional and Efficient: Automatically Generate Titles for Your Bug Reports

**プロフェッショナルかつ効率的であり続ける：
バグレポートのタイトル自動生成**

Xiaoyuan Xie, Songqiang Chen, Bangguo Yin, Yuanxiang Ji,
(School of Computer Science, Wuhan University)

Baowen Xu, Lin Chen

(State Key Laboratory for Novel, Software Technology, Nanjing University)

2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE)

紹介者：堀口 日向

GitHubのIssueは、タイトルや更新日時、投稿者などの情報がリストで1行ずつ表示される
⇒ Issueタイトルを正確かつ簡潔に作成することは、問題を適切に把握し、
効率的に対処するため重要

高品質なタイトルの利点

本文を詳細に読まなくても、

- 問題の理解
- 適切なラベル、優先順位の割り当て
- 類似した問題の発見

多くのIssueは、正確かつ簡潔な高品質のタイトルが付けられていない（次スライドに例）

Case A

Title: Cuda 3.0 not working

Body:

When installing from pip tensorflow GPU version (Linux), minimum cuda version is 3.5.

Install instructions have been changed to:

TensorFlow GPU support requires having a GPU with NVidia Compute Capability ≥ 3.0

However, gpu_device.cc still requires cuda 3.5:

```
std::vector supported_cuda_compute_capabilities = { CudaVersion("3.5"), CudaVersion("5.2")}; (omit trivial code)
```

Collected from <https://github.com/tensorflow/tensorflow/issues/1440>

- タイトル

CUDA 3.0が動かない

- 本文の要約

TensorFlow GPUのインストール説明書に、
CUDA 3.0以降が必要と書いてあるが、
実際には3.5以降が必要

⇒ このタイトルは情報不足

Case B

Title: Combined/Complete Guide of A Certificate (without the challenge part) for people with slow or restricted internet connection

Body:

I wish to request a feature which will benefit people with a slow, limited or restricted internet connection, that providing a downloadable, combined/complete guide of a certificate, in one HTML, pdf, epub, etc. So that people can study the complete guide whenever and wherever possible, and then when connected, return to the normal procedure of completing each test and challenges.

This also has the potential benefit for experienced ... *(omit other trivial description)*

Collected from <https://github.com/freeCodeCamp/freeCodeCamp/issues/17895>

- タイトル

インターネット接続が遅いまたは制限されている人のための、組み合わせた/完全な証明書ガイド（チャレンジ部分なし）

- タイトルの改善案

インターネット接続の悪いユーザーにダウンロード可能な証明書ガイドを提供する

⇒ このタイトルは冗長

目的，アプローチ

目的

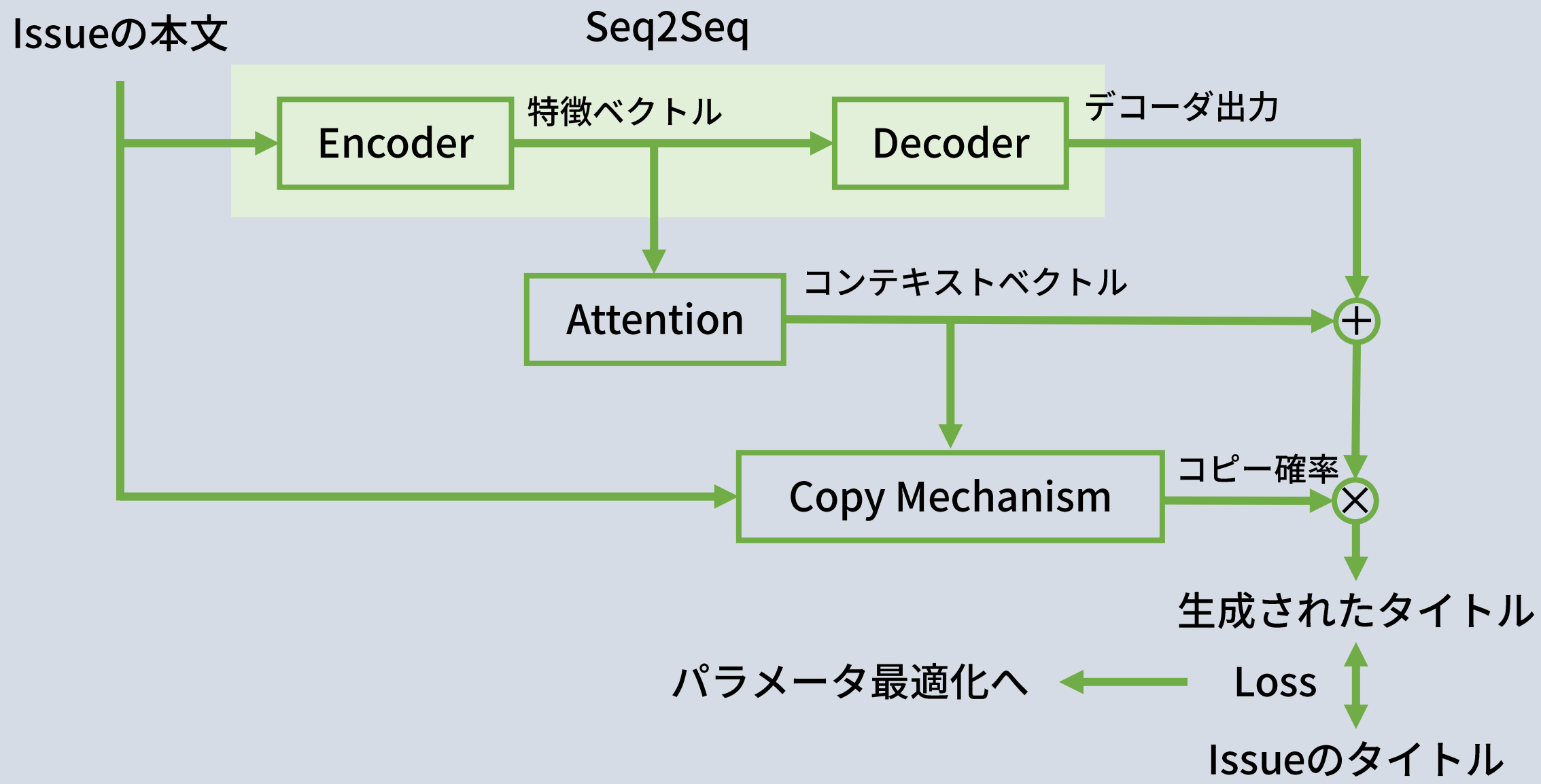
Issueタイトルの全体的な品質を改善し，メンテナのIssue管理や，開発者がIssueを利用する際の効率を高める

アプローチ

Issueタイトルを自動的に生成する機械学習モデル「iTAPE」を構築

- ・ タイトル生成をIssue本文を要約するタスクとして定式化
- ・ Seq2Seq + Attentionモデルをベースとした深層学習モデル

モデル概要



モデル構築時のドメイン固有の課題

課題1

モデルの学習に利用可能な、高品質のタイトルが付いたバグレポートのデータセットが存在しない

課題2

バグレポート特有の固有表現（変数名や関数名といった識別子や、バージョン番号など）は、出現頻度が低いためRNNでの学習が困難

課題1：高品質タイトルのデータセット作成

GitHubのスター数上位200件のリポジトリから約90万件のIssueを収集
以下の条件でフィルタリング（閾値などは，予備調査を基に決定）

- ・ **タイトルの長さが異常に短い/長い**（5~15単語以外）
- ・ **タイトルが本文の要約になっていない**
 - ・ タイトルの大部分（7割）の単語が本文に含まれていない
 - ・ タイトルが本文の一部をコピーしたもの

⇒ 約30万件のIssueを学習に利用

課題2：出現頻度が低い単語の学習 | Copy Mechanism

Copy Mechanismを利用[1]

Seq2Seq + Attentionのデコード時に，入力単語を確率でコピーする

$$p_{copy}^i = \sigma(\omega_c^T c_i + \omega_h^T h'_i + \omega_x^T x_i^{emb} + b_{copy})$$

Attentionの出力

Encoderの出力

入力単語のベクトル

これらを重みづけする ω を学習時に調整

文章要約の場合，入力には出力すべき単語が含まれている可能性が高い

⇒ 学習時に出現頻度が低くても，入力に含まれている単語を重みづけして

出力すれば，要約の精度を上げられる

[1]: Jiatao Gu, Zhengdong Lu, Hang Li, & Victor O. K. Li. (2016). Incorporating Copying Mechanism in Sequence-to-Sequence Learning.

課題2：出現頻度が低い単語の学習 | Tagging

Tagging

固有表現をタグ付けすることで，出現頻度が少ない単語に意味情報を付加

⇒ タグ自体は高頻度で出現するのでRNNが文脈を学習する助けになる

この論文では，識別子とバージョン番号にタグ付け

Issue body AFTER tagging:

When I add the <I> ad_mob <I\$> (<I> firebase_admob <I\$> : ^ <V>
0.6.1 <V\$> +1) Android emulator crashing before any code executes.
As soon as I take it out, it works. I tried to start <I> LogCat <I\$>, but
no logs are generated (I am using VS Code)
I have the following plugins working successfully with firebase
<I> flutter_crashlytics <I\$> : ^ <V> 0.1.1 <V\$>
<I> firebase_analytics <I\$> : ^ <V> 1.0.4 <V\$>
But <I> ad_mob <I\$> just won't work.

* This issue is collected from <https://github.com/flutter/flutter/issues/23659>.

提案モデルのパフォーマンス評価方法（リサーチクエスチョン）

提案モデルのパフォーマンスを評価するための3つのRQ

RQ1：人間が書いた高品質なタイトルに近いタイトルを生成できるか？

RQ2：前述の2つの課題への対処方法は、どの程度役に立つか？

RQ2.1：データセット作成時のフィルタリングはどの程度効果的か？

RQ2.2：Copy Mechanismや固有表現のTaggingはどの程度効果的か？

RQ3：実務者から見て、生成されたタイトルはどの程度の品質か？

RQ1：高品質なタイトルを生成できるか？

Issueタイトル生成の既存研究が存在しないため、**既存の文章要約モデルの性能と比較**

- ・電子メールの件名生成モデル×1
- ・ニュースの見出し生成モデル×2

これらは、最新の要約生成タスクにおいて満足できる性能であることが確認されている

評価指標

- ・ **Precision**：生成されたタイトルのうち、一致する単語の割合
⇒ 人間のタイトルを、いかに少ない単語生成で正確に当てられるか？
- ・ **Recall**：人間のタイトルのうち、一致する単語の割合
⇒ 生成したタイトルに、いかに多くの人間のタイトルの単語を含められるか？

RQ1：高品質なタイトルを生成できるか？

評価指標（つづき）

Precision, Recall計算時の「一致する単語数」のカウント方法

- ROUGEスコア
 - ROUGE-N：N-gram単位で一致する単語数をカウント
 - ROUGE-L：単語が連続で一致する最大の長さをカウント

結果

Method	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1
iTAPE	31.36	13.12	27.79
<i>Zhang et al. [45]</i>	26.11	11.43	25.64
<i>Xu et al. [42]</i>	27.92	10.27	24.03
<i>Song et al. [35]</i>	34.23	15.35	29.48

すべての評価指標において、
既存の文章要約モデルと同等の性能

RQ2.1：データセット作成時のフィルタの効果

3つのフィルタをそれぞれ単体で適用した結果と，すべて同時に適用した結果を比較
フィルタの内容

- ・ R1：長さが長い/短いタイトルの削除
- ・ R2：タイトルの大部分の単語が本文に含まれていないタイトルの削除
- ・ R3：本文の一部をそのままコピーしたタイトルの削除

Refinement	ROUGE-1			ROUGE-2			ROUGE-L		
Strategies	F1	P	R	F1	P	R	F1	P	R
∅	29.89	30.60	30.96	12.46	12.89	12.96	26.33	28.24	28.56
R1	30.05	30.57	31.27	12.55	12.92	13.11	26.47	28.20	28.83
R2	31.12	32.12	31.91	13.06	13.64	13.44	27.41	29.60	29.41
R3	30.64	31.92	30.96	12.76	13.42	12.95	27.14	29.48	28.61
R1+R2+R3	31.36	32.61	31.72	13.12	13.77	13.34	27.79	30.10	29.32

ルール未適用と比べて，
ほぼすべてのルールで
わずかにスコアが上がる

RQ2.2：出現頻度が低い単語の学習

Copy MechanismとTaggingの適用/未適用時のRecallを比較

Resolving Approach	Identifier	Version No.
∅	38.99	26.62
Copy	55.50	41.46
Copy + Substitute	58.46	41.99
Copy + Tagging	59.20	43.33

- **Copy Mechanism**を適用することで、スコアが大きく向上
 - Copy Mechanismに加えて、Substituteや**Tagging**を加えるとさらにスコアが向上
 - Substitute：識別子やバージョン番号を特定の単語に置換する方法
- SubstituteよりもTaggingのほうが若干スコアが高い

RQ3：実務者からのフィードバック

提案モデルによって生成されたタイトルが，実務者視点でどの程度の品質か，
以下の2つの観点で確かめる

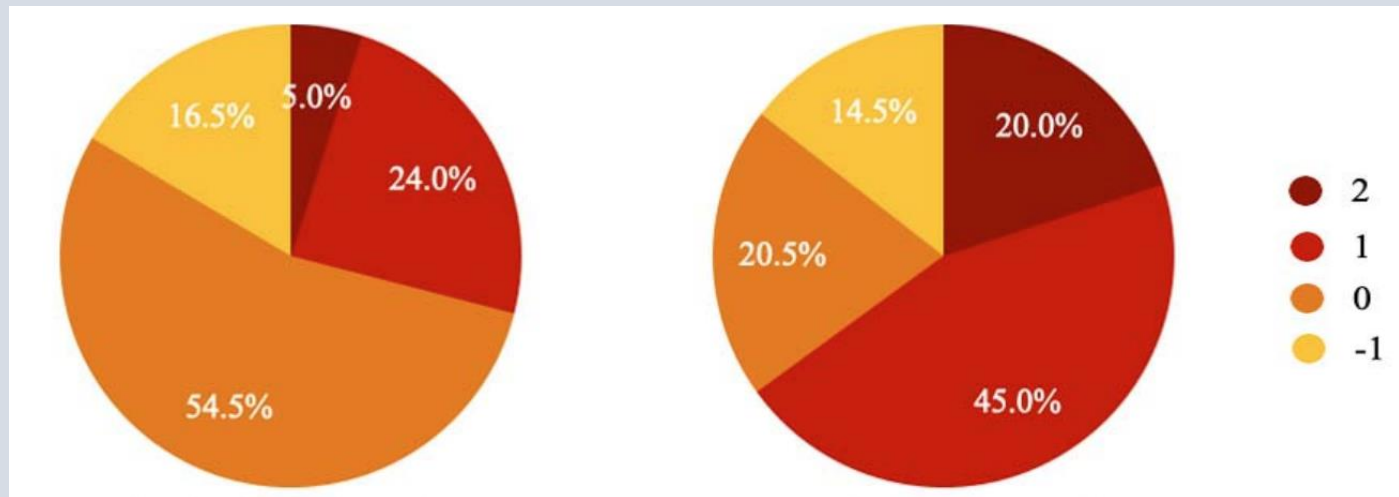
- ・人間の低品質なタイトルと比べ，iTapeがより高品質なタイトルを生成できるか
- ・人間の高品質なタイトルと比べ，iTapeが同等に高品質なタイトルを生成できるか

調査方法

- ・2人のボランティアが人間のタイトルが高/低品質のIssueを100件ずつサンプリング
これらのIssueに対して提案モデルを適用してタイトル生成
- ・別の6人のボランティア（大学院生）に対して，Issueの本文と，
人間のタイトル，生成されたタイトルを見てもらい，
両方のタイトルを4段階で評価

RQ3：実務者からのフィードバック

人間のタイトルが低品質のIssue



人間のタイトル

生成されたタイトル

2：完璧

1：わずかに情報が不足，
または文法的におかしい

0：本文のトピックは捉えているが，
情報が足りていない

-1：本文のトピックとは異なっている

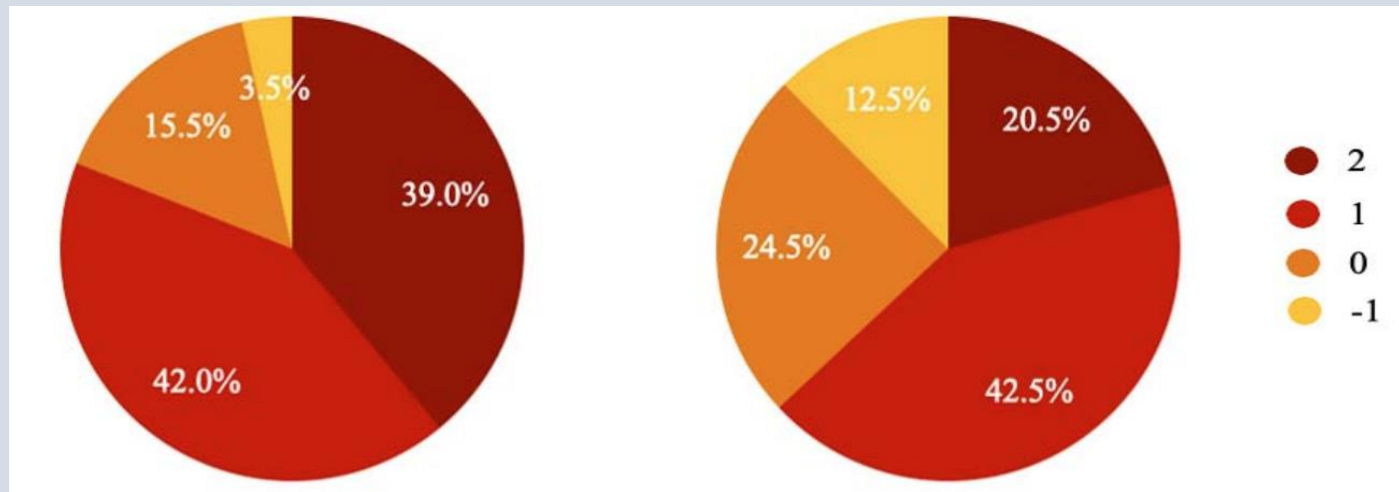
・ 人間のタイトル ： スコア1以上：29%，スコア0以下：71%

・ 生成されたタイトル：スコア1以上：65%，スコア0以下：35%

⇒ 提案手法を使うことで，人間の低品質なタイトルを改善することが可能

RQ3：実務者からのフィードバック

人間のタイトルが高品質のIssue



人間のタイトル

生成されたタイトル

2：完璧

1：わずかに情報が不足，
または文法的におかしい

0：本文のトピックは捉えているが，
情報が足りていない

-1：本文のトピックとは異なっている

・ 人間のタイトル：スコア2：39.0%，スコア1：42.0%，スコア0以下：19%

・ 生成されたタイトル：スコア2：20.5%，スコア1：42.5%，スコア0以下：37%

⇒ 完璧なタイトルは人間のほうが書ける（教師データが人間のタイトルであるための限界）

RQ3：実務者からのフィードバック

実際に生成されたタイトル

- ・人間のタイトルの品質：低品質
- ・生成されたタイトルの品質：高品質 と判断されたケース

Case i

Original Title: problems in firefox

Generated Title: file chooser dialog doesn't work in firefox

- ・人間のタイトル：Firefoxの問題
- ・生成されたタイトル：Firefoxのファイル選択ダイアログが動かない

RQ3：実務者からのフィードバック

実際に生成されたタイトル

- ・人間のタイトルの品質：高品質
- ・生成されたタイトルの品質：高品質 と判断されたケース

Case ii

Original Title: dist build only show a black screen

Generated Title: building dist on windows only shows a black window

- ・人間のタイトル：distビルドすると黒い画面のみ表示される
- ・生成されたタイトル：Windows上でdistビルドすると黒い画面のみ表示される

RQ3：実務者からのフィードバック

実際に生成されたタイトル

- ・人間のタイトルの品質：高品質
- ・生成されたタイトルの品質：低品質 と判断されたケース

Case iii

Original Title: improve connection to datasource lost indicator

Generated Title: lost connection in datasource when connection is lost

- ・人間のタイトル：datasourceとの切断を示すインジケータの改善
- ・生成されたタイトル：datasourceとの切断時の切断

データセットにこのような内容のIssueが少ないので、
より一般的な問題であるコネクションロストの問題と誤認した？

- 目的 Issueタイトルの全体的な品質を改善し，メンテナのIssue管理や，開発者がIssueを利用する際の効率を高め
- アプローチ **Issueタイトルを自動生成**する機械学習モデル「iTAPE」を構築
- 技術的な工夫
 - ・ 学習のためのデータセット構築
高品質なタイトルのデータを収集するためのフィルタリング
 - ・ Copy MechanismとTaggingを利用
出現頻度が低い単語でもタイトルに反映
- 結果 iTAPEの性能を定量的（ROUGEスコア），
定性的（実務者インタビュー）に確認

タイトルが低品質であることで実際に起きそうな弊害はどんなものか？

人気プロジェクトだとIssueが多すぎて、本文が見てもらいにくい？

タイトルが低品質だと、

- ・ 誰にも見てもらえず、自分の抱えている問題に気付いてもらえない？
- ・ 中身は非常にインパクトの高いIssueでも見逃される？

自分の研究に活かせることが多そう

モデル自体は要約生成タスクのものだったが、Issueの本文を入力とした自然言語処理モデルという意味では自分の研究（文章分類タスク）と共通していて、前処理テクニックやモデルそのもの、論文の書き方含め学ぶことが多かった