

ソースコード自動要約における生成文の有用性の予測

山村慎太郎

背景 | ソースコード要約

| ソースコードの内容や動作を自然言語で記述したもの
→コード理解, メンテナンス, など様々なタスクを支援

| 要約文の作成は人手で行うため, 時間・労力がかかる
→要約文生成の自動化に関する研究が行われる

背景 | ソースコード自動要約

機械学習モデルなどを用いて、要約文を自動生成



要約品質を向上させる手法の提案が研究の中心

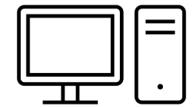
背景 | 自動要約手法の評価

生成文の品質を評価することで、要約手法を評価

主な評価方法

①自動評価

自動評価メトリクスを用いて、自動的に品質を評価



②人間評価

人間の評価者による主観的な評価を手動で行う



背景 | ①自動評価

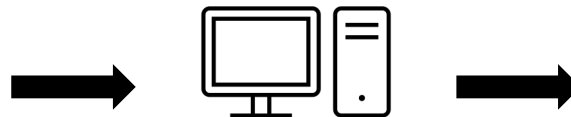
生成文と参照文のN-gramの重複度を測定することで評価



生成文



参照文



0~1の数値で
重複度を評価

- 一般的に、自然言語処理分野で用いられる評価指標(ツール)が
用いられる

—例：BLEU, METEOR, ROUGEなど

労力やコスト，一貫性の問題などから，研究では①が
主に利用される

背景 | ②人間評価

経験豊富な開発者が、あらかじめ定義された指標に従って生成文を1つずつ評価していく評価方法



生成文



各指標について
5段階評価

- ・ 指標：自然さ，表現力，適切性，簡潔性，有用性，理解性
- ・ 高品質で総合的な評価を行うことができる

デメリット

- ・ 労力・コストがかかる
- ・ 主観的な判断に依存する（評価に一貫性がない）

既存研究 | 自動評価と人間評価

自動評価は、人間評価との相関が高くない※

(例) 自動評価では高評価だが、人間評価では低評価な生成文

Source Code:

```
boolean goFromRetiredToDead(){  
    return compareAndSet(Status.RETIRED, Status.DEAD);  
}
```

Reference: attempts to transition the entry from retired to dead when releasing the handle.

Generated comments by *Re²Com*: attempts to transition the entry from idle to customise when evicting from the idle cache.

重複するN-gram

"attempts to transition the entry from" が自動評価スコアにプラスの影響を与えたため、自動評価で高い評価を得たと考えられる

自動評価では高評価でも、人間にとって有用とは限らない^[1]

原因：自動評価は、文脈や同義語、コード内容を考慮しないため

[1] Correlating Automated and Human Evaluation of Code Documentation Generation Quality

有用性の高い要約

| 自動評価では、人間にとって有用な要約文がどんなものかわからない

自動評価と人間評価は必ずしも一致しないため

| 要約文のこういった要素が有用性に影響するかを分析

有用性とその他人間評価指標との関係性の分析

得た特徴から要約文の有用性の予測

頓挫理由

| 人間評価と自動評価の相関， いうほど低くないことが
発覚

| 特に，最も重視していた「有用性」はMETEORとピア
ソン相関0.74と最大値に

→ 「人間評価『有用性』と自動評価にギャップがある」
という前提が崩れ， 研究方針が破綻

反省点

| 先行研究の論文内容全体をよく読めていなかった

論文が長いこともあり、とりあえず必要な部分だけよく読み込んでいた

→後々になって前スライドの内容が論文に載ってることに気づく

反省：先行研究の論文は全体をよく読みましょう

| 仮定を前提にしてしまう思考のクセ

誤った解釈を前提にして研究を進めてしまうことで、間違いに気づけなかった

反省：論文で一つ一つ事実の裏付けをしましょう

今後の方針

| 研究テーマをもらう

田口先輩の卒研の派生

分類器にNaïve bayesを用いたら？をやることになりそう