

不具合報告とレビューコメントの照合

Automatically Matching Bug Reports With Related App Reviews

Marlo Haering, Christoph Stanik , Walid Maalej.

Accepted for publication to the 43rd
International Conference on Software Engineering (ICSE21)

背景

- アプリ市場は競争が激しく，不具合報告も日々大量にある
 - 変化するニーズに対応するためにも迅速な対応が必要
- 利用者のニーズや問題の特定を開発者が理解するのは困難
 - レビューコメントを全て見るのは時間的コストが大きい
 - 賛否の声等の機能に関係ないレビューが多い
- フィードバックをフィルタリングするアプローチが研究されている

先行研究との違い

○ 先行研究

- バグレポート[44]や機能要求[16]など、関連するユーザのフィードバックの自動識別
- フィードバックをクラスタリングして、多くのユーザが抱えているトピックを分類[49]

○ 先行研究が抱える課題

- 課題管理システムのissueとレビューコメントのreviewのギャップが埋められていない
ex(開発者は、アプリのレビューで報告された問題がすでに課題追跡システムにバグレポートとして提出されているかどうかを簡単に追跡できない

目的

課題管理システムとレビューコメントを簡単に紐付ける

→ IssueとReviewを照合するシステムの提案

○ 貢献

- 課題追跡の効率化
- 繰り返し生じる問題の特定
- 開発者と利用者の語彙の違いを理解する

提案手法 | DeepMatcherの概要

Review:非公式, 非技術的, スペルミスが多い
Issue:公式, 技術的, フォーマルな形式

BERTの軽量版であるDistilBERT [41]を使用
周囲の単語の文脈を考慮することができる
→ スペルミス等に対応できる

Review

Problem Report

When I type into the search box's it type's random words on it's own even when I delete the random words it adds words in, it's not my prediction keyboard that's messing up it only happens on Firefox.

Bug Report Summary

I am able to type words in Firefox search bar but unable to type anything in the websites

Embedding

DistilBERT
Tokenizer

DistilBERT

Contextualized
Embeddings

Token
Mapping

SpaCy
Tokenizer

POS-Tagger

Nouns

Text
Embedding

Problem
Report
Embedding

Bug Report
Embedding

Cosine
Similarity
0.91

thumbs up
relevant
match

Issueの要約のみを
特徴量とするので十分
[22,25,48]

ストップワード除去
Spacyを用いて品詞を特定し,
名詞のみに限定

コサイン類似度の
スコアが大きい
Issueを選択
(1~3個で評価)

RQ

○ RQ1

DeepMatcherは、アプリのReviewとissueをどの程度正確に照合できるか？

- 定量的分析

○ RQ2

DeepMatcherの関連する提案，無関係な提案，見逃された提案はどのような違いがあるのか？

- 定性的分析

評価データセット

○ 4つのアプリのReviewとissueを収集

- Firefox (ブラウザ)
- VLC (メディアプレイヤー)
- Signal (メッセージャー)
- NextCloud (クラウドストレージ)



issue

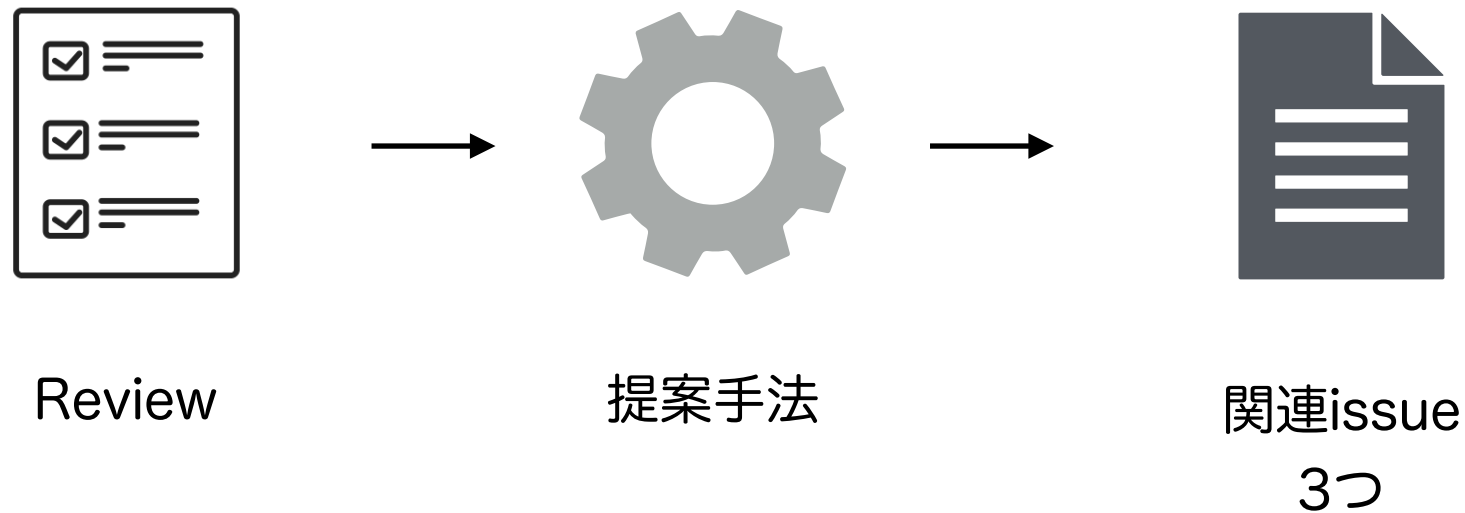
計40,724件
Bugzilla, Trac, GitHub



Review

計21,506件(2~4年)
ノイズ除去^[44] → 9,132件
↓
各アプリからランダムで
50件ずつ選択し目視
計200件になるように

評価方法



- 入力のReviewと出力のIssueに関連性があるかを著者らがラベル付け
 - ReviewとIssueが同じアプリの機能と動作を記述している場合は関連あり

- 全てのReview P に対する平均精度： $MAP = \frac{\sum_{p=1}^P AveP(p)}{P}$ [52]

- 指定する数の関連issueが存在することを前提としたスコア
- p : 各review
- $AveP$: 提案の平均精度 (P に関連する Issueを提案できているか)

RQ1 | Reviewとissueの照合

App	出力がIssue1つ			出力がIssue2つ			出力がIssue3つ				
	1 Suggestion			2 Suggestions			3 Suggestions				
	#	MAP	Hit Ratio	#	MAP	Hit Ratio	#	MAP	Hit Ratio	# Relevant Matches	Coder Agreement
Firefox	50	0.50	0.50	100	0.54	0.58	150	0.58	0.74	38	0.93
VLC	50	0.32	0.32	100	0.38	0.44	150	0.40	0.51	26	0.91
Signal	50	0.38	0.38	100	0.47	0.57	150	0.50	0.68	45	0.89
Nextcloud	50	0.62	0.62	100	0.73	0.84	150	0.73	0.89	58	0.88
Total	200	∅ 0.45	∅ 0.46	400	∅ 0.53	∅ 0.61	600	∅ 0.55	∅ 0.71	167	∅ 0.90

出力のIssueのうち、3個全て関連する場合、1
2個なら2/3、1個なら1/3、0個なら0の平均スコア

600件のIssueのうち
167件の関連Issueを特定
Review109件/200件をカバー

少なくとも1個の関連Issueを見つけたものの割合

- 関連マッチしないReviewが 91件/200件 存在
 - 関連するissueが存在しない or 提案手法の見逃し
 - キーワード検索で手動分析
- 47件存在せず(issue作成に活用できる), 44件見逃し

RQ2 | 関連するマッチと無関係なマッチの違い

- 文脈埋め込み(DistilBERT)を用いることで高いテキスト類似度に達している
 - ‘save pdf’, ‘download pdf’
 - ‘automatic synchronization’, ‘auto upload’
- 言語ギャップにより関連性のあるマッチが少なくなる
 - 目視の際にVLCのissueには専門用語が多く含まれていた → 類似度低
 - Nextcloudの開発者はReviewに返信したりissueの提出を求めた→類似度高

提案手法の活用方法とまとめ

○ 活用方法

- バグの早期発見
- ユーザーからのフィードバックによるバグレポートの改善
- 提案手法の拡張による重複, 反復, または類似したバグレポートの特定

○ まとめ

- 利用者が記述するReviewから問題部分を抽出し,
一致する不具合管理システムのissueを特定する手法を提案
- 200件のReviewから167件のissueとの関連性を特定
- 関連性を特定できなかった91件のReviewのうち47件が
まだissueになっておらず, 不具合早期検出に有効

所感

- 目視大変そう
- アプリ系の不具合報告には要約されたものが存在しているのか？
ソース等のリンクはなかったけど。
- 評価軸の説明等があんまりなかったので表の解釈が難しかった

追加資料

単語埋め込み技術の選択

○ フィルタリング

- 賛否の声や有益な情報を含んでいるものが多い[37][42]
- [44]を用いてissueをバグレポート, 機能要求, 無関係に分類 → 9,132件
- 信頼性確保のために調査対象4つのアプリの分類済みデータを手動分析→各50件計200件

○ それぞれでトークン化した単語を

- 全文の埋め込みを計算 = 各単語ベクトルの平均

○ 入力

- issueレポート

○ 出力

- 関連するバグ報告の候補をランク付け
- 距離関数を用いて類似度を計算→コサイン類似度を選択

単語埋め込み技術の選択

○ どのメトリクスを用いるか

- ユーザと開発者で使っている語彙が違う
 - ユーザのフィードバックはスペルミスが多い顔文字あり
- 構造的違い
 - レビューは非公式・口語・非技術的なものが大半
 - バグ報告はテンプレがあり技術的な内容が記載されているフォーマルな形

○ 埋め込み技術 → BERT [6]の軽量版である DistilBERT [41]

- バグ報告の要約分のみをベクトル化する

○ 前処理

- ストップワード除去
- spacyを用いて各単語トークンを768次元で表現 → スペルミスに対応
- 名詞のみ(動詞等があると結果に偏りが生じたため)

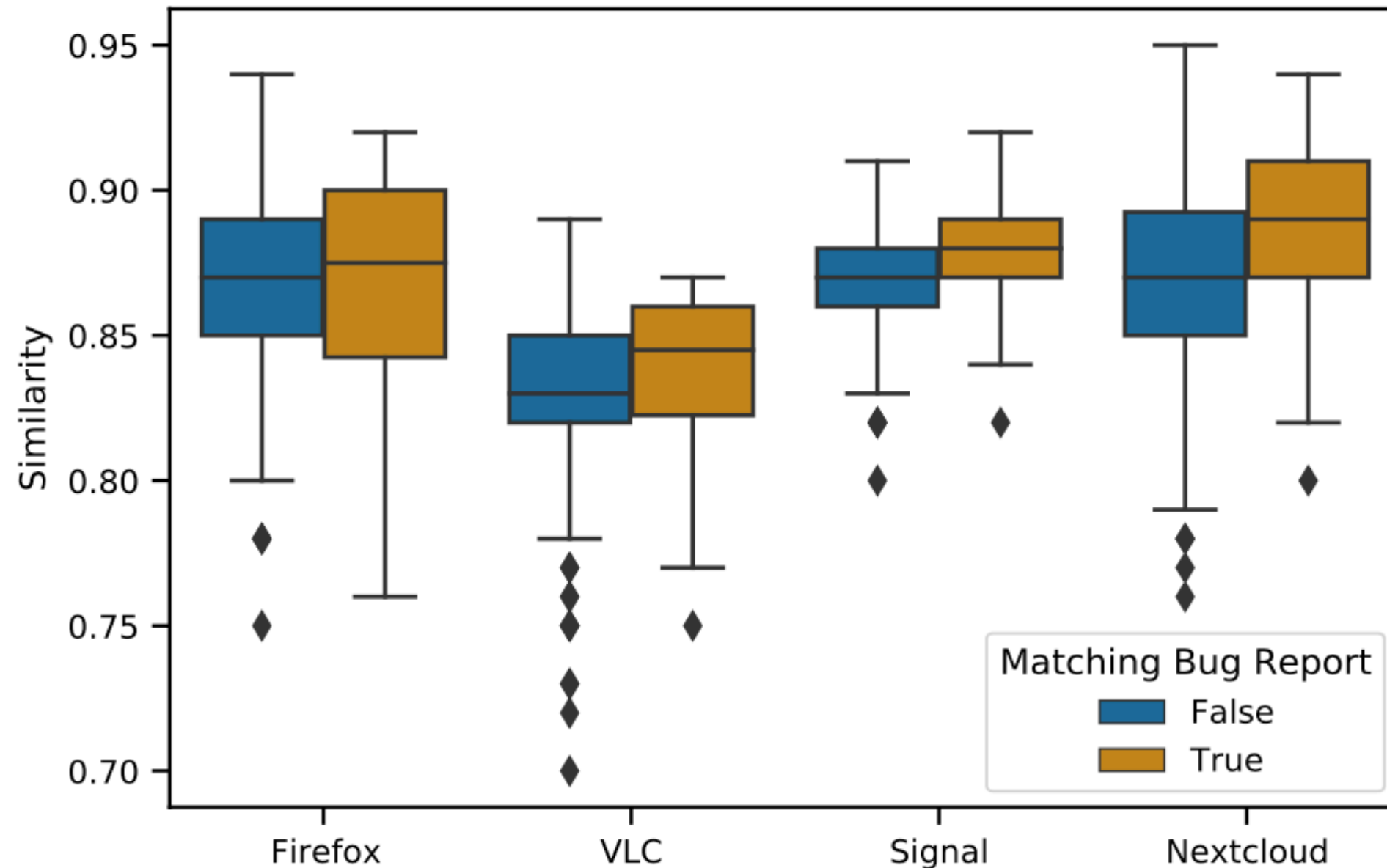
Spacy

	SPACY	SYNTAXNET	NLTK	CORENLP
Programming language	Python	C++	Python	Java
Neural network models	✓	✓	✗	✓
Integrated word vectors	✓	✗	✗	✗
Multi-language support	✓	✓	✓	✓
Tokenization	✓	✓	✓	✓
Part-of-speech tagging	✓	✓	✓	✓
Sentence segmentation	✓	✓	✓	✓
Dependency parsing	✓	✓	✗	✓
Entity recognition	✓	✗	✓	✓
Coreference resolution	✗	✗	✗	✓

評価データ

App name	Bug reports		App reviews		
	Time period	Count	Time period	Count	Problem reports [44]
Firefox Browser	01/2011 08/2019	29,941	09/2018 07/2020	5,706	3,314
VLC Media Player	05/2012 07/2020	553	09/2018 07/2020	5,026	2,988
Signal Messenger	12/2011 08/2020	7,768	09/2018 08/2020	10,000	2,583
Nextcloud	06/2016 08/2020	2,462	06/2016 08/2020	774	247
Total		40,724		21,506	9,132

テキストの類似度



- 目視の際にVLCのissueには専門用語が多く含まれていた → 類似度低め
- Nextcloudの開発者はReviewに返信したりissueの提出を求めることがあった