README における見出しの同定処理に向けた内容分析

石岡 直樹 † 伊原 彰紀 ‡ 南 雄太 §
和歌山大学 † 和歌山大学 ‡ 和歌山大学 §

1 はじめに

多くのオープンソースソフトウェア (OSS) プロジェクトは，ソフトウェアの使用方法，開発への貢献方法，ソフトウェアのライセンスなどの情報を開発者や利用者に発信するために，ソフトウェアと共に README を作成し，公開する．特に，ソフトウェア公開用プラットフォームとしても利用される GitHub を使用するプロジェクトでは，リポジトリのトップページに README を提示することが推奨されている．また，ソフトウェア開発において README といったドキュメントファイルは，ソフトウェアの開発環境やシステムの保守において重要な役割を果たしていると示唆される [1].

README は，ソフトウェアの情報を簡潔に説明するために見出し（例: インストール方法）とその説明文（例: インストールするためのコマンド）を記載する．したがって README の内容はソフトウェアによって異なり, README に記述すべき内容の明確なガイドラインは確立されていない．従来研究では，README のガイドラインの作成を目的に GitHub に保存される JavaScript ライブラリにおいて共通して記述される見出しを調査し，表記揺れする見出しの統合を目視で行っている [2]．見出しの表記揺れは，開発者が勘や経験に基づき見出しを決定していることが原因と示唆される．

本論文では，README において表記揺れする見出しの自動同定処理に向けて，見出しと見出しの説明文の一貫性を機械学習モデルを用いて明らかにする．具体的には，見出しの説明文中に含まれる単語を説明変数として，見出しを予測するモデルを構築する．高い精度で予測できる見出しは，見出しと見出しの説明文に一貫性があると判断する．

2 分析

2.1 データセット

本論文では，Libraries.io [*1]において公開される JavaScript ライブラリを対象に，README を作成して 3 年以上が経過し，英語で記述されるリポジトリ，ランダムに選択した 1,000 リポジトリを分析する．対象を 1,000 リポジトリのみに限定する理由は，表記揺れする見出しを目視によって同定するためである．対象とする 1,000 リポジトリから最新の README を取得し，従来研究 [2] において最も記述頻度が多い「使用方法」に関する見出しと，その内容の一貫性を分析する．使用方法が記述された見出しと，その見出しが記述されたリポジトリ数は，usag (452 件)，exampl (119 件)，use (98 件) であり，そのほかは，1% 未満のリポジトリにのみ記述されるため対象外とする．対象とする見出しは，ステミング処理をしているため，usag には Usage，usages なども同一の見出し usag とする．

English title
† Naoki Ishioka, Wakayama University
‡ Akinori Ihara, Wakayama University
§ Yuta Minami, Wakayama University

[*1] Libraries.io:https://libraries.io/

表 1: 予測結果

	モデル 1	モデル 2	モデル 3
Precision	0.45	0.50	0.55
Recall	0.70	0.87	0.81
F1	0.55	0.63	0.65

2.2 分析手法

使用方法を示す見出し（usag，exampl，use）のそれぞれの説明文の一貫性を分析するために，説明文に含まれる単語をデータ整形する．

1. 正規表現を用いて，プログラム，URL，記号を削除
2. NLTK を用いて，説明文の分かち書き
3. NLTK を用いて，ストップワードの削除
4. NLTK を用いて，ステミング処理
5. Bag-of-Words により，単語のベクトル化

ベクトル化した説明文を説明変数とし，使用方法を示す見出し（usag，exampl，use）を予測するモデルを構築する．本論文では，自然言語解析において広く利用される教師あり学習モデル SVM（Support Vector Machine）を用いる．SVM は，2 クラスのパターン識別器であるため，本論文では，モデル 1（usag と その他），モデル 2（exampl と その他），モデル 3（use と その他）を構築する．予測モデルの評価には，本論文の対象データが膨大ではないため Leave-One-Out 交差検証を用いる．モデルの評価には適合率・再現率・F1 値を用いる．

2.3 結果

表 1 は，モデル 1〜モデル 3 の予測結果を示す．モデル間の予測精度の差は小さく，適合率は 0.45〜0.55 であったため，ランダムに予測する場合の予測精度と同じと考えられる．言い換えると，開発者は usag，example, use の見出し語を区別せずに使用していると考えられる．ただし，3 つのモデルの再現率は 0.70〜0.87 であり，ランダムで行うよりも高い精度であることから，判別する閾値を調整することにより，再現率を維持したまま高い適合率を得る可能性もあり，モデルの改善を試みる．

3 おわりに

本論文は，README における見出し，特に利用頻度の高い「使用方法」の見出しと見出しの説明文の一貫性を SVM モデルを用いて分析した. 分析の結果，開発者は usag，example, use の見出しを区別せずに使用していると考えられる. 今後は，README の説明文に対応する統一した見出しを分析し，README を作成するためのガイドラインの確立を目指す.

謝辞

本研究は JSPS 科研費 18H03222 の助成を受けたものです.

参考文献

[1] Kipyegen, N. J. and Korir, W. P. K.: Importance of Software Documentation, *IJCSI International Journal of Computer Science Issues*, Vol. 10, No. 1 (2013).

[2] Ikeda, S., Ihara, A., Kula, R. G. and Matsumoto, K.: An Empirical Study on README contents for JavaScript Packages, *IEICE Transactions on Information and Systems*, Vol. E102.D, No. 2, pp. 280–288 (2019).