1B-01

開発文書中のかっこ表現ばらつき抽出技術の評価

杉本駿† 南田幸紀† 原田山人†
日本電信電話株式会社 ネットワークサービスシステム研究所†

1. はじめに

ソフトウェア開発では，上流工程での要件定義書などの開発文書の品質が，ソフトウェアの品質に大きく影響することが知られている．そのため，要件の矛盾点や記述のあいまい性など，欠陥の原因となりうる記述を取り除くためにレビューを行っている．レビューでは，ほかにも誤字や表記揺れなどの要件の内容とは関係の少ない軽微な指摘も行われる．軽微な指摘を制限することで，より効率的な指摘が実施できるという報告もされている[1]．つまり，レビュー前に，軽微な修正を効率的に取り除くことができれば，レビューの品質が向上すると考えられる．著者らは，軽微な修正を効率的に抽出する手法の一つとしてかっこ表現のばらつき抽出手法を提案している[2].

本稿では，著者らが提案したかっこ表現のばらつき自動抽出手法を，実際の開発文書に適用し，作業効率化の評価を行なった結果について報告する．

2. 開発文書の課題

ソフトウェア開発において，欠陥を修正するコストは下流の工程になるほど，高価になると言われている．そのため，より上流の工程で欠陥を発見することで，ソフトウェアの品質向上や，開発期間延伸のリスク低減につながると考えられる．一般的には上流工程の文書に対してレビューを行い，品質を確保しているが，レビュー稼働は限られており，さらなる品質向上のためには，レビューの効率化が必要である．

著者らは，レビューの効率化に向けて，実際のソフトウェア開発のレビューコメントの中に機械的に抽出できるコメントがないか，調査を行った．調査の結果，丸かっこ内の記載のばらつきに関するコメントに着目した．以降，かっこ表現とは丸かっこについて言及していることとする．かっこ表現に着目した理由は二点ある．一点目は，かっこ表現は文書中で多用される表現であり，目視での網羅的な確認は困難なためである．二点目は，固有表現の抽出を目的として，かっこ表現の分類に関する研究[3]もされていることから，かっこ表現に注目することで，固有表現の表記揺れなどが効率的に抽出できると考えたためである．著者らの先行研究において，かっこ表現を効率的に確認することを目的として，かっこ表現のばらつきの抽出手法の検討を行った．

3. 関連研究

文書の自動校正技術は，広く研究されており，例えば，文の係り受け構造のあいまいさに着目した自動校正技術がある[4]． しかし，文書校正の観点でかっこ表現のばらつきやあいまい性に着目した研究は見られない．

名詞
ネットワーク
サーバ
接続
要求条件
警報監視
・・・

(a) 抽出した名詞の例

グループ名	Diameter 信号略称	SIPメソッド
名詞	UDR	INVITE
	UDA	ACK
	PPR	BYE
	PPA	CANCEL
	・・・	・・・

(b)同族語辞書の例

被補足語	補足語
疎通	なし
	応答
リクエスト	Ini-INVITE
	Initial-INVITE
・・・	・・・

(c)自動抽出結果の出力例

図 1 かっこ表現のばらつき自動抽出手法

4. かっこ表現のばらつき自動抽出手法

本手法で自動抽出対象とするかっこ表現は「疎通(応答)」のように名詞 X と語句 Y が「X（Y)」の形で記載されたもののみとする．以降の説明では，X にあたる語句を「被補足語」，Y にあたる語句を「補足語」と呼ぶ．また，INVITE や BYE のように，SIP メソッドという上位概念にまとめられる名詞を，「同族語」と呼ぶこととする．同族語を定義することで，INVITE（SIP メソッド)，BYE（SIP メソッド）というように，被補足語が異なるが，補足語が同じになる語句の表記揺れ等を検出することが可能となる．

かっこ表現のばらつきの判断基準は，同一の被補足語に対して，文書全体で補足語の有無が混在している場合と，補足語が複数種類ある場合とする．上記の基準をもとにかっこ表現のばらつきを次のように抽出する．

(1) 文章を一文ずつ抽出し，形態素解析と品詞付与を行い，名詞だけを抽出する．名詞が連続している場合には 1 つの名詞として扱う．（図 1-a）
(2) 名詞の中で，同族語辞書に含まれている語を，グループ名毎にグルーピングする．なお同族語辞書は事前に人手で作成しておく．（図 1-b）
(3) 各名詞に対してかっこ表現の有無を確認する．

An Evaluation of the Technique to Extract Differences in Parenthetical Information from Software Documentations
† SHUN SUGIMOTO, KOKI MINAMIDA, YAMATO HARADA,
NIPPON TELEGRAPH AND TELEPHONE CORPORATION
Network Service Systems Laboratories

(4) 被補足語各々に対して，かっこ表現のばらつきがあるか判定し，被補足語と補足語の対応関係を出力する．（図 1-c）

5. 評価実験

提案手法による稼働削減効果を評価するために，実際の開発文書を利用し，評価実験を行った．

5.1 かっこ表現のばらつき自動抽出手法の実装

提案手法を実装したツールの概要について説明する．MS-Word で記載された文書から Apache Tika によってテキストを抽出し，MeCab によって形態素解析，品詞分析を行う．そして提案手法で示した判定方法を用い，かっこ表現のばらつきを抽出する．その結果を MS-Word 上に一覧としてユーザへ提示する．

5.2 実験方法

実際の要件定義書に対して提案手法を利用した修正箇所の抽出作業を行い，実際のレビュー結果と比較し作業の効率化を評価した．対象文書は，あるソフトウェア開発プロジェクトのレビュー前の要件定義書 230 ページとした．評価の流れを図 2 に示す．

A) 対象文書に対して，提案手法を適用し自動抽出した結果をもとに，人が目視でかっこ表現のばらつきに関する修正要否の判断を行い，コメントを発出する．

B) 人が目視のみで，レビューを行う．かっこ表現に限らず，仕様書として修正が必要と判断した部分に対してコメントを発出する．

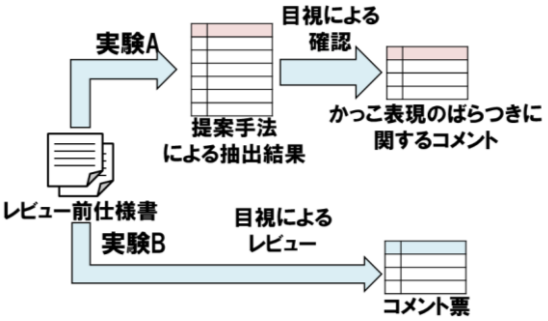

図 2 評価実験の流れ

6. 実験結果

実験 A，B の結果を表 1 に示す．実験 A では，提案手法によって文書中のかっこ表現のばらつきがある被補足語が 282 件抽出された．そこから人の確認によって修正すべきコメントが 28 件抽出された．人の確認によって抽出されたコメントを正解集合とすると，提案手法の適合率は 9.9%と計算される．現時点での抽出アルゴリズムでは，かっこ表現のばらつきがあったものを全て抽出しており，そのうち表記揺れなどの実際に修正が必要なものが，今回は 10%程度だったと考えることができる．つまり，かっこ表現のばらつきを確認することで，修正すべき事項を抽出でき，レビュー時の確認の観点として有効であると考えられる．その一方，ばらつきはあるが修正不要なものが多く含まれている．具体的には，文書全体ではばらつきがあるが，ある章内では統一されており，文章の理解には影響がないもの，略語の正式名称を初回のみ記載し，次回以降省略しているもの等があり，これらの記載の特徴を，自動で判別し，自動抽出対象から除外することができれば，更なる適合率の向上が期待できる．

次に二つの実験での作業効率を比較する．コメント数と作業時間から，コメント 1 件あたりの平均作業時間は，実験 A で 9.1 分，実験 B で 15.6 分であった．また実験 A/B で共通のコメントは 2 件，つまり実験 A のみで抽出されたコメントは 26 件となる．比較のため，両実験の総コメント 58 件の抽出にかかる想定の時間を実験 A/B それぞれの条件で概算する．概算には，コメント 1 件あたりの平均作業時間を利用する．実験 A では，かっこ表現に関するコメント 28 件は提案手法，残り 30 件は実験 B と同条件とすると，724 分要する．実験 B では，コメント 58 件抽出するには，905 分必要となる．このことから総コメントの抽出，つまり文書品質を統一するために必要な時間は提案手法を用いたことにより，約 20%短縮できることが分かる．

表 1 実験結果

タスク内容	提案手法による抽出数	コメント数	作業時間	コメント1件あたりの作業時間	適合率
実験A	282件	28件	256分	9.1分	9.9%
実験B	ー	32件	500分	15.6分	ー

7. まとめ

かっこ表現のばらつき自動抽出手法を実開発の文書へ適用し，作業効率化の評価を行なった．かっこ表現のばらつきに着目し確認することで，目視だけでは抽出できなくかつ，文書品質向上に有効な指摘を行えることを確認した．またかっこ表現に関するコメントを抽出する作業では，1 件あたりの作業時間を約 39%短縮できることを確認した．

現時点では，かっこ表現のばらつきがあるものを全て提示しているが，ばらつきの走査範囲の限定や，かっこ表現の使われ方を自動分類する技術を確立させ，適合率の向上を目指していく．

参考文献

[1]三浦一輝，森崎修司，”軽微な指摘を抑制したソフトウェアレビューにおける指摘欠陥の分析，” 研究報告ソフトウェア工学（SE）2013-SE-179(7), 1-7,2013-03-04.

[2]杉本駿，南田幸紀，原田山人，“ソフトウェア開発文書におけるかっこ表現ばらつき抽出，” ソフトウェアエンジニアリングシンポジウム 2017 論文集, 2017,798-203(2017-08-23).

[3]久光徹，丹羽芳樹，“統計量とルールを組み合わせて有用な括弧表現を抽出する手法，” 情処研報 NL, Vol. 1997, No. 122, pp.113-118, (1997).

[4]今枝恒治，河合敦夫，石川裕司，永田,亮，桝井文人，“日本語学習者の作文における格助詞の誤り検出と訂正，” 情処研報, CE, Vol. 2003, No. 13, pp. 39-46, (2003).