

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

杉本駿^{†1} 南田幸紀^{†1} 原田山人^{†1}

概要: 変化するユーザニーズに対応するために、ソフトウェアを効率的に開発することが求められている。特に上流工程での開発文書の品質を確保することが、下流工程での誤りや手戻りを防ぐことにつながるため重要である。上流工程での開発文書の誤りや理解を妨げる表現の自動検出が可能となると、効率的な品質向上が期待できる。一般的に単語の意味を補足するために使われているかっこ表現だが、用法を誤ると文意が曖昧になり開発文書の品質低下につながる可能性がある。そこで本稿ではかっこ有無の混在やかっこ内の語のばらつきを自動検出するツールを提案する。開発文書から見直すべきかっこ表現のばらつきを抽出する実験を行い、ツールによる作業の効率性を評価した。実験にあたっては、人手による抽出タスクを均質化するために、見直すべきかっこ表現とそうでないものを仕分けするためのガイドラインを作成した。

キーワード: 文書校正, 仕様書, 要件定義書, かっこ表現, 日本語

Extracting Differences in Parenthetical Information from Software Documentations

SHUN SUGIMOTO^{†1} KOUKI MINAMIDA^{†1} YAMATO HARADA^{†1}

Keywords: Proofreading, Specification, Requirement Definition Document, Parenthesis, Japanese Language

1. はじめに

ソフトウェア開発では、変化するユーザニーズに対応するために、品質は維持し、より短期間の開発が求められている。高品質かつ効率的な開発のために、開発の各工程において自動化・省力化の技術が数多く研究されている[1,2]。上流工程の成果物である、要件定義書や基本設計書は、下流工程での成果物の基になるため、ソフトウェアの品質に大きく影響する。さらに効率面でも、要件定義や基本設計などの上流工程起因での不具合を下流工程で発見した場合には、ソフトウェアの仕様に関わるため、手戻りの原因となり、開発が遅延するリスクが大きくなる。そのため、要件定義書や設計書などの上流工程での開発文書を高品質に作成することが重要である。実際のソフトウェア開発プロジェクトでは、チェックリスト等を利用し、人の目によってレビューを行うことで開発文書の品質を確保している。しかし、チェックすべき観点には仕様の内容のみならず、表記ゆれや体裁等の表層的なものまで多岐にわたり、目視では見落としが起きてしまうのが実情である。

開発文書では、特に誰が読んでも一意に解釈できることが重要とされる。あいまいな表現があると、文書レビューが長時間化する可能性や、設計工程で読み手が理解を誤り不具合につながる可能性が高くなる。そのため、あいまいな表現を自動で抽出することが可能となれば、ソフトウェアの開発効率化、品質向上に有用である。

本稿では、開発文書におけるあいまいな表現の一つとし

て丸かっこの用法に着目する。同一の名詞に対しかっこによる補足の有無が混在していることや、かっこ内の語がばらついていることで文意があいまいとなる場合がある。そこで、かっこ表現のばらつきの自動抽出方法を提案する。提案手法の有効性を示すために、文書から見直すべきかっこ表現のばらつきを抽出するタスクで評価実験を行い、提案手法による作業の効率性を評価した。実験にあたっては、見直すべきばらつきと見直す必要のないばらつきとを仕分けするためのガイドラインを作成し、抽出タスクの均質化を図った。

2. 開発文書品質向上の課題

2.1 かっこ表現のばらつき

ソフトウェア開発において、誤りを修正するコストは、上流において修正するほど安価で、下流になるほど高価になる[3]。したがって、要件定義や設計工程などの上流工程での開発文書の品質が向上することで、ソフトウェアの品質向上や、開発遅延のリスク低減が期待できる。ここでいう開発文書とは、要件定義書や設計書などの各工程で作成される文書のことである。開発文書の品質は通常、人の目視レビューにより確保されている。

著者らは、あるソフトウェア開発プロジェクト2件の要件定義書225ページ分のレビューコメント(欠陥の指摘や内容の確認・質問)885件を調査した。コメントを、文や文書の表層的な部分に対するもの(a)、要件や仕様の内容に対するもの(b)、文書の修正を伴わない質問等(c)に分類したところ、(a)が10%、(b)が42%、(c)が48%であった。なお、文法的指摘、フォント、必須事項の欠落、文書間の項目の

^{†1} 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
NIPPON TELEGRAPH AND TELEPHONE CORPORATION
Network Service Systems Laboratories

対応の過不足の指摘などを表層的なコメントに分類した。文書から表層的なコメントを排除しておくことで、レビューワーが内容の吟味に専念でき、文書の品質向上に繋がると考えられる。そのためには、文書から表層的な問題を確実かつ効率的に抽出する技術が必要である。

著者らは、文のあいまいさに関わる、あるひとつのコメントに着目した(表1)。表1の修正前記述はレビュー対象文書に記載されていた文言である。「PPA」は通信プロトコル Diameter で定められた信号名の略称である。丸かっこ「Diameter 信号」という語によって補足された語と補足のない語が混在しており、補足のない「PPA」は Diameter 信号ではない別の略語を指しているのか、と無用な疑念を生み、文意の理解を妨げるため、補足有無を統一するようレビューワーが求めている。

開発文書では、読み手によらず一意に意味が読み取れることが重要とされる。複数の解釈の可能性がある、要件のあいまい性につながる。かっこ表現については、かっこの有無が混在している場合や、かっこ内の語にばらつきがある場合に、複数の解釈につながる可能性があると考えられる。またかっこ表現は経験的に開発文書内で多用されており、人の目ですべてを確認するには、大きな稼働が必要となるだけでなく、見落としも懸念される。著者らはかっこ表現のばらつきを自動抽出することが、開発文書品質の向上につながると考え、検討することとした。

表1 レビューコメントの一例

修正前記述	コメント	コメント回答
～SIP サーバからの PPA 受信により加入者データを確定する。～SIP サーバへ PPA(Diameter 信号)を送信し～	補足の記載レベルが異なっておりますので、合致をいたします。	記載レベルの統一を行います。

2.2 かっこ表現の用法

かっこ表現のばらつきが文意のあいまいさにつながると述べたが、ばらついていても理解を助ける適切な補足となっている場合や、概念の差異を適切に表している場合などは、文意の理解に問題がない。つまり、ばらつきには、理解を妨げるものと、そうでないものが存在する。しかし、ばらつきが理解を妨げるか否かの線引きは容易ではなく、人によって判断に差が出てくる。とはいえ、文書の品質を高めるには、どのようなかっこ表現あるいはかっこ表現のばらつきが理解を妨げるかが定められている必要がある。

3. 関連研究

品質の高い文書を作成するための自動校正の技術として、文の係り受け構造のあいまいさに着目した自動校正、自動校正支援技術が広く研究されている[4,5]。しかし、著者らが調べた範囲では、かっこ表現のあいまい性に着目し、自動校正を行う研究は見られない。

またかっこ表現に関する研究については、新聞記事にお

けるかっこ表現に着目し、用法の分類[6]、言い換え表現の抽出に活用する方法[7]が提案されている。新聞記事では、かっこ表現の利用例や用法が整理されているが、著者らの属する組織のソフトウェア開発プロジェクトでは、かっこの使い方に関する明文化されたルールはない。

4. かっこ表現ばらつき自動抽出手法

本節では、かっこ表現のばらつきを自動抽出する方法を提案する。

4.1 自動抽出対象の定義

本稿で自動抽出の対象とするかっこ表現は、丸かっこのみとし、鉤かっこなどは対象としない。また丸かっこの使い方も複数考えられるが、「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を自動抽出の対象とする。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。

かっこ表現のばらつきとして判断する基準は、下記の二点とする。

- (a) 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
- (b) 同一の被補足語に対して文書全体で異なる補足語が複数存在する場合

4.2 複合語の扱い

文書中では「開発文書」のように「開発」と「文書」という名詞が連なって構成される複合語も存在する。本稿では連続している名詞を結合し、一つの名詞として扱うこととする。単独の名詞と複合名詞を区別して扱うことにより、「文書」と「開発文書」のように、別の意味で使用している語の区別が可能となる。

4.3 同族語の扱い

2.1 で「PPA」と「PPA (Diameter 信号)」という語が混在している場合の理解の妨げについて説明した。Diameter 信号は他にもあり、UDR, MAR などが存在する、このように一つの上位概念にまとめられる個々の語を同族語と呼ぶこととする。仮に同じ文書内で「UDR (Diameter 信号)」と「PPA」が記載されているとする。その場合には、PPA は Diameter 信号以外の何かを表しているのではないかの疑念が生まれ、読み手の理解を妨げる可能性がある。そこで、同族語の関係にある被補足語は上位概念毎にばらつきを調べる。語 X', X"が上位概念 \bar{X} に属し、かっこ表現「X' (Y')」と「X" (Y")」があるとき、上位概念 \bar{X} を被補足語とみなし、被補足語 \bar{X} に対する補足語 Y' と Y"があり、ばらついているとみなす。

4.4 ばらつきの自動抽出手法

4.1, 4.2 および 4.3 での基準をもとに、かっこ表現のばらつきを次のように抽出する。

- (1) 対象文書の文章を一文ずつ抽出し、形態素解析と品詞付与を行う。

- (2) 付与された品詞情報から名詞を選別する。(図 1-a)
- (3) 文中で名詞が連続している場合には、複合語として結合する。(図 1-b)
- (4) 名詞の中で、同族語辞書に含まれている語を、上位概念毎にグルーピングする。なおグルーピングに用いる辞書は事前に人手で作成しておく。(図 1-c)
- (5) 被補足語各々に対して直後に開き丸かっこがあるか調べ、有の場合には開き丸かっこから閉じ丸かっこまでを補足語とする。
- (6) 被補足語各々に対して、4.2(a)または(b)の基準に当てはまるかどうか判定し、当てはまる場合には、ばらつきのある被補足語として出力する。

5. かっこ表現のガイドライン

2.2 で提起したかっこ表現によって理解を妨げるか否かの判断基準の課題に対して、かっこ表現の使い方のガイドラインの策定を試みた。策定したガイドラインの一部を表 2 に示す。策定にあたり、あるソフトウェア開発プロジェクトの要件定義書 69 ページからかっこ表現をすべて抽出し、用法を分類した。本文中には 187 組のかっこ表現があり、同様の使い方をしている表現をグループにまとめ、15 項目に分類した。次に各分類に対して、理解の妨げとなりうるか吟味し、利用可否を区分けした。具体的には、かっこの有無、補足語のばらつきがあった場合に複数の解釈の可能性があるかどうか、補足語を読み飛ばした場合に解釈が大きく変わらないかどうか、の二つの観点で区分けを行った。

利用可否が○のものは、ばらつきがあっても許容される表現、×はばらつきがない場合にもかっこを使用することを避けた方がよい表現、△はばらつきの出現の仕方によっては許容されることを表している。△としたものの利用可

図 1 ばらつきの自動抽出手法

表 2 ガイドラインの例

分類名	使い方	利用可否
参照先	文書内の図表の参照先を記載 例：状態遷移 (図 5.2-2)	○
具体例	被補足語の具体例を記載 例：工事 (ファイル更新, プラグ投入)	○
所属	同名だが異なる概念を、それが属する別の概念をカッコ内に記載することにより区別する 例：契約者番号 (オーダ投入システム) 契約者番号 (呼制御システム)	△
限定	被補足語の意味をより限定する目的で記載 例：DB サーバ (スレーブ)	△
言い換え	同様の意味の別の語を記載 例：監視 (ヘルスチェック)	△
並列	被補足語と並列で並べたい語を記載 例：レコードを追加 (削除) するときは	×
重要情報	注意事項等の読み飛ばせない内容を記載 例：救済 (スレッド障害の場合は非救済)	×

否の判断の仕方は、具体的には「所属」と「限定」は、ひとつの表、章、節などのある範囲の中で全ての被補足語に補足語が付いていれば補足語が異なっても許容、「言い換え」は、章、節などのある範囲の中で、最初に出現する被補足語にのみ補足語が付き、その後補足が付かずに被補足語が現れるなら許容する。

6. 評価実験

提案手法を利用し、文書の高品質化作業を行う場合の稼働削減効果を評価するために、評価実験を行った。

6.1 ばらつき自動抽出の実装

かっこ表現のばらつきの自動抽出を行うツールを実装した。実装したツールの概要を説明する。まず 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する。次に MeCab により形態素解析および品詞付与を行う。そして、複合語の生成、同族語のグルーピング、補足語の抽出、ばらつき有無の判定を行う。判定の方法は 4.1 節に示した単純なものである。ばらつきがあると判定された被補足語を、その補足語らとともに csv ファイルで図 1-d のように出力する。

6.2 実験方法

文書のかっこ表現のばらつきを排除する推敲作業を想定し、文書から見直すべきかっこ表現を指摘するタスクによって評価することとした。抽出元の文書は、あるソフトウェア開発プロジェクトの MS Word で書かれた要件定義書 50 ページ (27000 文字) とした。比較のため、下記 A, B の 2 通りの方法で、タスクを実行した。C を比較用の参

名詞	複合語
ネットワーク	要求条件
サーバ	警報監視
接続	輻輳制御
PPA	エラー応答
情報	トラヒック制御
...	...

(a)抽出した名詞の例

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

(b)複合語の例

グループ名	被補足語	補足語
なし	疎通	なし
		応答
Diameter信号略称	PPA,UDR,MAR ...	なし
		Diameter信号

(c)同族語辞書の例

グループ名	被補足語	補足語
なし	疎通	なし
		応答
Diameter信号略称	PPA,UDR,MAR ...	なし
		Diameter信号

(d)自動抽出の出力例

考として設ける．評価の流れを図 2 に示す．

- A) 人が対象文書のみを参照し，目視でかっこ表現のばらつき抽出と見直し要否の判断を行う．
- B) ツールでかっこ表現のばらつきを自動抽出した結果を，人が対象文書を参照しながら見直し要否の判断を行う．
- C) ツールでかっこ表現のばらつきを自動抽出した結果をそのまま見直しを要する表現とする．

実験 A, B のどちらも，タスクを均質化するために，被験者に 5 章で述べたガイドラインを参照してもらい，補足語毎のかっこ表現の分類付与と被補足語毎の見直し要否の判断をしてもらった．なお，今回のタスクは見直し後の文書の作成や文書の修正作業は含まない．

見直すべきかっこ表現の正解集合は，実験 A, B の抽出結果を参考に複数人で見直し要否を検討した上で定めた．

実験 A, B のタスクをそれぞれ 2 人，計 4 人の被験者に実施してもらった．

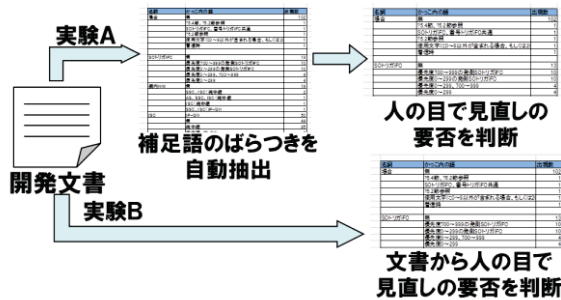


図 2 評価実験の流れ

7. 実験結果

7.1 指摘要否判断の結果

実験 A, B, C の結果を表 3 に示す．また図 3 では実験結果のイメージを弁図として表現している．正解集合は 24 件だった．作業時間は実験 A (目視のみ) と実験 B (自動抽出と目視) を比較すると 82 分，つまり 37%短縮されている．今回は 50 ページの文書での作業だったので，1 ページあたり 1.6 分程度のレビュー作業時間の短縮が見込める．

再現率，適合率については実験 A (目視のみ) が実験 B (自動抽出と目視) を上回る結果となった．実験 B の再現率が実験 A のそれを下回った一因は，ツールが正解を十分に抽出していないことによる．実験 C (自動抽出のみ) で指摘漏れとなった表現 (図 3 のア, イの領域) が 6 件あった．指摘されなかった原因は，

- 図や表中のテキストで，Apache Tika によって抽出されなかったことによるもの
- かっこ表現のばらつきはないが，ガイドラインにおいて使用を避けた方がよいとされていたもの (例：重要情報)．

があった．ツールにおいて上記二点の原因を解消できるような改善が必要である．また，自動抽出後の人手による見

表 3 実験結果

実施方法	平均指摘数	平均作業時間 (分)	再現率	適合率	F 値
A 目視のみ	19	222	65% (15.5/24)	82% (15.5/19)	0.72
B 自動抽出+ 目視	21.5	140	38% (9/24)	41% (9/21.5)	0.46
C 自動抽出のみ	63	5	75% (18/24)	29% (22/63)	0.41

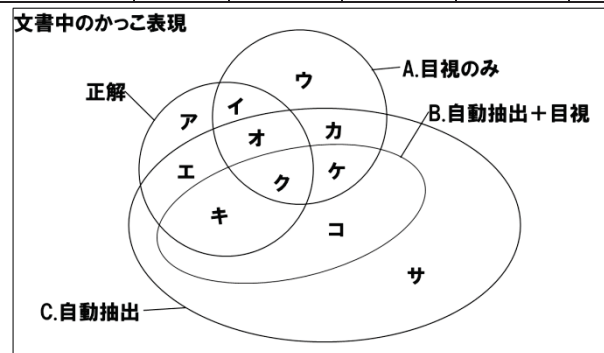


図 3 実験結果のイメージ図

直し要否判断において，見直すべき表現を見直し不要と誤って判断しているもの (図 3 のエ, オの領域) があつたことも再現率が低かつた要因である．これは，被験者のガイドラインの理解が不足していたことによると考えられる．ガイドラインに関する考察は次節で述べる．実験 B の適合率が実験 A のそれを下回ったのは，ツールが，見直し不要なかっこ表現まで機械的に出力すること (図 3 のカ, ケ, コ, サの領域) による．

次に実験 A (目視のみ) と実験 C (自動抽出のみ) の結果を比較すると，実験 C の再現率が 10 ポイント高い．これは，目視では指摘できていない表現をツールによって指摘できていることを表している (図 3 のエ, キの領域)．ツールでのみ指摘できていた件数は平均 7 件あつた (実験 A の被験者 2 名の平均)．その内訳は，かっこ表現を抽出する際に見落とししたものが 1 件，見直すべき表現を見直し不要と誤って判断したものが 6 件であつた．ツールでは人が見落としした表現を抽出できており，ツールが見落としを防ぐために有効であるといえる．人が見直し不要と判断してしまつた 6 件はガイドライン改善による均質化を検討しており，考察は次節で述べる．

7.2 かっこ表現の分類結果

本実験で高い再現率/適合率を得るためには，ガイドラインに基づきかっこ表現を正しく分類する必要がある．4 人の被験者のかっこ表現の分類の正答率を表 4 に示す．実験 B の被験者の分類の再現率が A の被験者と比べ低い値と

なっている。これは実験 A の被験者はガイドラインの内容を熟知しており、実験 B の被験者は実験 A の被験者と比べると理解度が低い状況だったため、このような差が現れたと思われる。そのため、実験 B の被験者の分類の正答率が低く、その結果見直し要否判断の正答率が低くなったと考えられる。ガイドラインの理解度の差が作業品質に顕著に影響してしまった理由には次のことが考えられる。

- (1) ガイドラインの記載が不十分
- (2) ガイドラインのカッコ表現の分類の独立性が不十分でひとつのカッコ表現が複数の分類に該当する場合があります

(1)は、各分類の例を多く示すことで改善できると考える。(2)の具体例として、「限定」と「重要情報」の類似性がある。たとえば、「特番（緊急呼を除く）」という表現は、「緊急呼」や「緊急呼でないもの」を含む「特番」という広い概念に対し、「緊急呼を除く」ものに限定しているとみなせるし、「(緊急呼を除く)」というカッコ表現を省くと仕様としては誤ったものになってしまうから「重要情報」ともみなせる。(2)の問題は、分類を見直すことや、優先させる分類を細かく定義することによって改善できると考える。

表 4 補足語分類の結果

実施者	実験 A		実験 B	
	実施者 1	実施者 2	実施者 3	実施者 4
分類の再現率	72%	86%	44%	31%

8. まとめ

開発文書の品質向上のための推敲作業の自動化・省力化の一助となるカッコ表現のばらつき自動抽出方法を提案した。また提案手法による作業効率化の評価実験を行った。見直しを要するカッコ表現を抽出するタスクにおいて 37% の作業時間の短縮が確認できた。また本手法を活用することにより、目視では見落としていたカッコ表現の抽出もれを減らせる可能性が示唆された。

また、カッコ表現の見直し要否判断を均質化するために、カッコ表現の用法に関するガイドラインを策定した。ガイドラインは、評価実験のタスクを統一する効果があった。またガイドラインがあることで、カッコ表現の良さあしについて議論や分析が可能となった。しかしながら、用法分類の独立性や定義の説明が不十分であり課題が残る。

今後は以下の課題に取り組んでいく。

- 本実験における実装では、カッコ表現のばらつきを一覧のみを出力したが、元の文書中のカッコ表現の位置とともに出力する機能を追加し、更なる作業効率化を図る。
- カッコ表現ガイドラインの見直しを行い、誰もがかつ

こ表現の良さあしを同じように判断できるようにする。判断基準を改善するために、過去の開発のレビューコメント等を参考にする。

- かつこ表現ガイドライン中の利用を避けた方がよいとした分類を、自動的に識別する方法を検討する。
- かつこ表現に限らず、他の開発文書のチェック観点の自動化を検討し、ソフトウェア開発の効率化に寄与する技術確立を行っていく。

参考文献

- [1] 小形真平, 松浦佐江子, “UML 要求分析モデルからの段階的な Web UI プロトタイプ自動生成手法,” コンピュータソフトウェア, Vol.27, No.2, pp.14-32(2010).
- [2] 下村翔, 古畑慶次, 松尾谷徹, “大量の状態とイベントを持つプログラムの自動解析とテスト手法の提案～想定外のイベントを自動テストする～,” ソフトウェア・シンポジウム 2016 論文集, pp.59-67(2016)
- [3] 独立行政法人情報処理推進機構, “形式手法活用ガイド 導入の手引き,” 独立行政法人情報処理推進機構(2012)
- [4] 林良彦, “技術文章向けの日本文推敲支援システムの実現と評価,” 信学論 D, Vol. J77-D2, No.6, pp.1124-1134, (1994).
- [5] 今枝恒治, 河合敦夫, 石川裕司, 永田亮, 榊井文人, “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).
- [6] 中山悟, 森田和宏, 泓田正雄, 青江順一, “括弧表現の抽出・分類に関する研究,” 言語処理学会年次大会発表論文集, Vol. 16, pp. 379-382, (2010).
- [7] 久光徹, 丹羽芳樹, “統計量とルールを組み合わせる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).