

キーワードと文法解析を用いた
要求文書からの非機能要求の抽出手法とツールの発展方針
A Retrieval Method and Tool Development Plan of Software Requirements
from Requirement Document with Grammatical Analysis and Keywords

小川 貴史[†] 原田 史子[†] 島川 博光[†]
Takafumi Ogawa Fumiko Harada Hiromitsu Shimakawa

1. はじめに

要求仕様書には、一般に開発体制に対する指示、会議の予定、導入による効果など、ソフトウェア品質要求以外の内容も記述される。また、文書のフォーマットは記述者に依存する。これらのことから、要求仕様書には、一読するだけでは、ソフトウェア品質要求のみを素早く、かつ正確に把握するのは難しいという課題がある。

そこで、この課題を解決するために、要求仕様書から要求が記述された文章だけを抽出し、JIS 規格に基づいた 12 種類のソフトウェア品質特性別に分類するツールを提案した[1]。[1]では、外部とのインタフェース要求、性能要求、論理データベース要求、設計制約、時間効率性、資源効率性、互換性、使用性、信頼性、セキュリティ、保守性、移植性の 12 種類のソフトウェア品質特性に対する分類精度を評価し、検証した範囲で高い再現率が確認できたことを報告した。

本論文では、[1]で誤検出文を排除しきれなかったことを受け手法を改善し、抽出精度を一部向上させることができたため報告する。また、[1]にて提案したツールについて発展方針を議論する。

2. 関連研究

木下ら[2]は、キーワードを用いた要求抽出手法を提案している。性能要求、互換性、使用性、信頼性、セキュリティの 5 要素に関して、それぞれキーワードを設定し抽出する手法をとっている。抽出漏れが発生しないようにキーワードを設定すると、必然的に多くの用語を設定しなければならず、結果として誤検出が増えてしまうことが課題として挙げられている。

北川ら[3]は、Word2Vec と深層学習を用いた要求仕様の自動分類手法を提案している。Word2Vec による単語の分散表現を生成し、ラベル付けた文章を入力とした教師あり学習を行うことで、自動的に分類するモデルを作成している。ただし、この研究では結果の再現率が 43%~69%と低いことから抽出漏れが懸念される。また、古川ら[4]も、単語のベクトルを用いた機械学習による要求の自動分類手法を提案しているが、こちらも精度は 60%程度となっており、決して高い精度が実現できているとは言えない。また、本研究は日本語の要求仕様書を対象としているが、[5][6]では英語の要求仕様書について機械学習を用いて分類を行っている。しかし、いずれも精度が高いとは言えない。

須堯ら[7]は、「ような」や「ように」といった直喩の表現に特定のキーワードがみられるというヒューリスティクスに基づき、キーワードと係り受け関係を持つ語を取

得することで喩表現データベースを構築する手法を提案している。画像から喩表現とその特徴を含む文章の作成を行う実験を通して、提案システムの有用性を示している。

石原ら[8]は、認知症患者の文法誤りや複雑な構文の利用頻度が減るなどの症状をヒューリスティクスとして、係り受け解析を用いた認知症傾向の検出を試みている。平均最大係り受け距離を利用した識別の正解率が 90%以上であることを報告している。

3. ヒューリスティクスに基づいた係り受け解析

第 2 章で挙げた関連研究から、単語ベクトルを入力とした機械学習技術を用いた分類では高い精度を得ることが困難であることと、ヒューリスティクスに基づく係り受け解析の利用に対する有効性を確認できた。これを受けて、本研究ではヒューリスティクスに基づく係り受け解析を利用した要求抽出手法を提案する。

3.1 キーワードと文法解析を用いた要求抽出

本研究では、まずキーワードを用いた抽出を行い、次に抽出した文章群を文法解析によって取捨選択する。提案する手法概要図を図 1 に示す。

キーワードを用いた抽出の目的は、再現率が 100%になるようなキーワード群を設定することである。文法解析の目的は、キーワードを用いて抽出された文章群から誤検出分を除外して要求文のみを抽出し、適合率を高めることである。この 2 段階をもって再現率、適合率ともに高い値となる抽出を目指した。

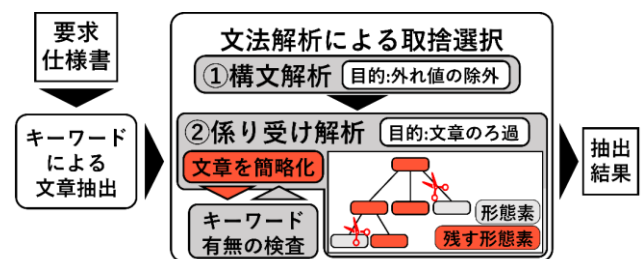


図 1 手法概要図

3.2 ヒューリスティクスに基づいた係り受け解析の利用概要

キーワードを使った抽出を行っている先行研究[2]において、誤検出となった文章を分析すると、以下の 3 つの特徴のいずれかが該当することがわかった。

- <特徴1> 後に記述する内容を参照する文章
- <特徴2> 文献や文書、書籍などを参照する文章
- <特徴3> 要求の補足説明を含む文章

[†] 立命館大学大学院情報理工学研究科

本研究では、この 3 つの特徴による誤検出を防ぐために、係り受け解析による構文解析と文章の簡略化を利用する。

3.3 構文解析による外れ値の除外

本節では、上記の特徴 1・2 に該当する文章の誤検出に対する解決法を提案する。

特徴 1 の文例は、「～を以下に示す」、「～は下記を満たす」などが挙げられる。こういった文章では、「以下」や「下記」という語が参照する先の内容には要求が含まれることが多い。しかし、この文章自体が直接的に要求を説明しているとは考えられない。また、特徴 2 の文例は、「～は別紙に示す」、「～を参照」などが挙げられる。こういった文章においても、「別紙」などの表現で参照する文書や文献には要求が含まれることが推測される。しかし、この文章もそれ自体が直接的に要求を説明しているは考えられない。したがって、これらの文章は間接的に要求に関わる可能性がある特殊な文章として扱うこととした。

手法としては、本項で挙げたような例文に出現する語について、係り受け解析の結果を利用して特定の処理を行うことで構文を判定した[1]。

3.4 係り受け解析による文章のろ過

本節では、主に上記の特徴 3 に該当する文章の誤検出に対する解決法を提案する。要求の補足説明が記述されている文章では、要求と直接には関係しない補足部分においてキーワードが使われていることで誤った抽出がされてしまうという問題がある。これに対する解決策として、文章の簡略化を行うことで、本質的な要求部分だけを取り出すことを試みた。簡略化した文章にキーワードが含まれているかを確認することで、文章を取捨選択した。

手法としては、文章を係り受け解析し、得られた形態素と係り受け関係の木構造の ROOT となる語から指定した係り受け関係のみをたどっていくことで、形態素を取捨選択することにより、文章の簡略化を実現した[1]。

4. 語と係り受けを用いた要求抽出・分類ツール

[1]にて開発ツールの内容や実装方法に触れているため、本論文では改善箇所のみを記述する。

4.1 ツール実装

まず、5 つの要求仕様書から 12 種類のソフトウェア品質特性それぞれにキーワード群を作成するさいに、異なる品質特性で同一の単一キーワードが発現した場合の差別化ができない問題があったため、組み合わせキーワードによって差別化ができるように改善した。

次に、文章を簡略化するために取得する係り受け関係を表 1 に示すように改善した。

表 1 取得する係り受け関係

係り先の語	指定する係り受け関係		
①「こと」が ROOT	acl	advcl	advmod
② ROOT(VERB), ①に acl で係る語	advcl	advmod	nsubj
	obj	obl	

4.2 抽出精度評価と考察

[1]では保守性要求にて 2 文の誤検出があったが、前節の改善によりこの誤検出を排除することができた。

また、[1]にて誤検出を排除しきれなかった外部とのインタフェース要求と、論理データベース要求、信頼性要求に関しては、まだ誤検出を排除することができていない。文章の簡略化で省略できない文章の根本部分で使われていることや、文章が人の動作を表していることが要因であると考えられる。

5. 発展方針

[1]でも言及したが、抽出精度を高めることに加え、機能要求を適用対象とすることと表や図の分析を可能にすることが挙げられる。ソフトウェア要求において機能要求は不可欠な要素である。また要求仕様書は表や図を用いた要求の記述が多い。したがって、これらの分析を可能にすることで、提案ツールの価値を高めることができると考える。

また、提案ツールを利用して要求仕様書から抽出した要求とテストを行う際のテスト仕様書のテスト内容との対応付けを行うことで網羅的に要求をテストできているかを確認できるように拡充する方針を考えている。

6. おわりに

本研究では要求関連文書から 12 種類のソフトウェア品質特性に関わる文章を抽出することを目的とした。要求仕様書中の要求文を抽出、分類する手法を提案し、手法に基づくツールを開発した。研究の結果、先行研究の課題であった「キーワードが想定とは違う意味合いで使われていることや、文章において要求とは関係のない部分で使われていることによる誤検出の発生」を、係り受け解析の技術を用いることで一部防ぐことができた。

今後は第 5 章に記した発展方針に基づいて研究を進める。

参考文献

- [1] 小川貴史,大森隆行,大西淳,“キーワードと文法解析を用いた日本語要求関連文書からのソフトウェア要求文の抽出”,電子情報通信学会技術報告,vol. 121, no. 334, KBSE2021-39, pp. 20--25, (2022).
- [2] 木下貴文,大森隆行,大西淳,“日本語要求文書からの非機能要求抽出と検証”,電子情報通信学会技術報告,Vol,121, KBSE2021- 21, pp.46--51, (2021).
- [3] 北川貴之,長岡武志「深層学習による要求仕様自動分類手法の提案と評価」,情処論,vol.61, no.4,pp.842-852,(2020).
- [4] 古川実歩,松尾美穂,村瀬摩由加,佐伯元司,“機械学習を用いた非機能要求の分類法の評価実験”,南山大学紀要『アカデミア』理工学編 第 22 巻,198-202,(2022).
- [5] J. Slinkas and L. Williams, "Automated extraction of non-functional requirements in available documentation," 2013 1st International Workshop on Natural Language Analysis in Software Engineering (NaturaLiSE), pp. 9--16, (2013).
- [6] M. A. Haque, M. Abdur Rahman, M. S. Siddik, "Non-Functional Requirements Classification with Feature Extraction and Machine Learning: An Empirical Study," 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), pp. 1--5, (2019).
- [7] 須堯大喜,寺井あすか,“係り受け解析を用いた比喩生成支援システム”,情処論, Vol.2019-CH-119, no.1, pp.1--6,(2019).
- [8] 石原 颯人,入部 百合絵,北岡 教英,“語彙と係り受け構造に着目した雑談対話からの認知症傾向検出”,情報処理学会第 82 回全国大会,pp.459--460,(2020).