7L-03

LDAによるトピック抽出に着目した非機能要求の要約手法の提案

多田　一仁†　　長岡　武志‡　　北川　貴之‡　　位野木　万里†

工学院大学†

東芝デジタルソリューションズ株式会社‡

1. はじめに

大規模情報システムの開発では，関連する情報システムの要求仕様書の理解も必要である．初級技術者にとり，効率的に関連する情報システムの理解を進めることは容易ではない．文書の理解の促進のために，自然言語処理に基づく要約技術を活用することは有効であると考えられる[1]．文書の要約技術は抽出型の要約と抽象型の要約に分けることができる．例えば，抽出型の要約手法では，TextRank[2]等，文書中から重要と考えられる代表文を選択する方法がある．また，文章のトピックを導出し，そのトピックに沿った文を選択するという要約手法もある．選択方法としては，重要なトピックに近い文の選択，各トピックの代表文を選ぶ方法が考えられる[3]．中でも，Latent Dirichlet Allocation（以下，LDA と略す）はトピック分布にディリクレ事前分布を仮定しベイズ推定する手法として注目されているが，要求仕様書からのトピック抽出への適用実績は明らかになっていない．そこで，本研究では，トピック分類技術である LDA を用いることで，非機能要求の要約と仕様理解支援手法を提案する．

以下，本稿は次のように構成する．2 章では研究課題と解決へのアプローチを示し，3 章では LDA を用いた非機能要求の要約方法を提案する．4 章では実際に行ったケーススタディを示し，5 章で考察を述べ，6 章で本稿をまとめる．

2. 研究課題と解決へのアプローチ

要求仕様書には，機能要求と非機能要求が含まれる[4]．非機能要求は，可用性，運用・保守性，拡張性，移植性等のタイプに分類される．このようなタイプをトピックとして特定できれば，要求仕様書に定義された非機能要求の要約情報を把握可能と期待できる．そこで，本研究では，要求仕様書のうち非機能要求に着目し，トピック抽出することで要約情報を生成し，技術者の理解の支援に貢献する．

Proposal of Summarization Method of Non-Functional Requirements Focusing on Topic Extraction by Latent Dirichlet Allocation

†Kazuhito Tada, Mari Inoki, Kogakuin University

‡Takeshi Nagaoka, Takayuki Kitagawa, Toshiba Digital Solutions Corporation

本研究でとりあげる LDA とは，トピックモデルの一種である[5]．トピックモデルとは，文書が複数の潜在的なトピックから確率的に生成されると仮定したモデルである．また，文書内の各単語はあるトピックが持つ確率分布に従って出現すると仮定する．トピックモデルでは，トピックごとに単語の出現頻度分布を想定することで，トピック間の類似性やその意味を解析できる．

3. LDA を用いた非機能要求の要約方法の考え方

LDA を用いた非機能要求の要約手法の概要を図 1 に示す．提案手法は以下の（1）～（3）で構成する．

（1）LDA モデルの構築：（1a）学習対象のデータセットを入力し，（1b）モデルの構築には Python ライブラリの gensim[6]を用いて，機能要求および非機能要求が記述されたデータセットを学習する．

（2）LDA によるトピック抽出：（2a）要約対象の文書を mecab[7]を用いて形態素解析し，（2b）上記（1a）で作成したモデルによりトピックを抽出する．

（3）トピックから非機能要求の要約を生成：（3a）（2）の結果である各トピックを構成する，単語，当該単語の品詞，出現頻度，確率を参照し，設計要素別に整理分類し，アクター，データ，振る舞いの視点で要約情報を生成する．（3b）トピック横断で要約情報を生成する．

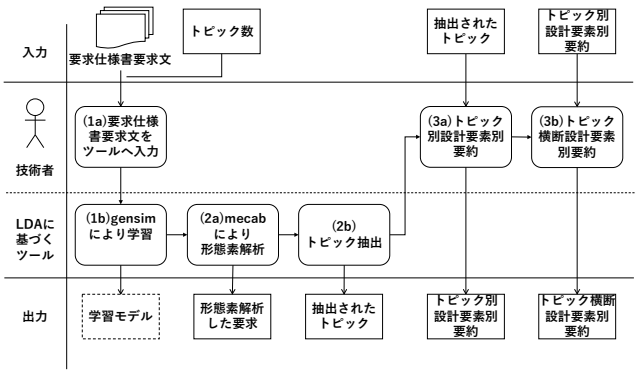

図 1　LDA によるトピック抽出に基づく要約手法の概要

4. ケーススタディ

提案手法を具体事例に適用し有効性を評価した．

4.1　ケーススタディの手順

非機能要求のデータセット要求文に対して，LDA を活

用した非機能要求の要約情報が全体の理解にどの程度有効かを確認する．対象とする非機能要求は，データセット要求文 PROMISE[8]を著者側で日本語要約した 625 件である．実際に入力するデータの一部を表 1 に示す．

評価手順は次の通りである．(1)機能要求と非機能要求が書いてあるデータセット要求文 625 個を学習させ LDA を用いてトピックごとに分類する．(2)トピックごとに出現頻度の高い上位 20 個を抽出する．(3) 出力された単語をアクター，対象，振る舞い，条件他の設計要素別に分ける．

表 1　入力する非機能要求データ

SEQ	TEXT	カテゴリ
1	システムは60秒ごとにディスプレイの表示を更新する。	性能
2	アプリケーションは、国土安全保障省によって定められたカラー・スキーマに準拠すること。	レイアウト
3	投影する場合、データは読み取り可能にすること。10x10の投影スクリーンにおいて、視聴者の 90%が30の視聴距離からイベントまたはアクティビティのデータを読み取ることがでること。	ユーザビリティ

4.2 ケーススタディの結果

LDA によるトピック抽出を行った結果を表 2 にまとめる．また，設計要素別に分けた単語を図 2 にまとめる．設計要素別に分類したものをみるとユーザや顧客などのアクター，製品やデータなどの対象，管理や使用などの振る舞い，時間や機能などの条件他からデータセット要求文の非機能要求の内容を，俯瞰的に理解することが可能である．

表 2　出力されたデータ(トピック 1)

SEQ	単語	単語品詞	出現頻度	確率
1	製品	名詞,一般	216	0.056
2	必要	名詞,形容動詞語幹	67	0.025
3	ユーザ	名詞,一般	138	0.012
4	使用	名詞,サ変接続	60	0.012
5	2	名詞,数	56	0.011
6	データ	名詞,一般	61	0.01

図 2　表 2 で出力された単語を分類した結果

5. 考察

LDA を使ってトピックごとに分類をするだけでは，要求仕様書に記載されている非機能要求の内容について理解することは有効ではないが，トピック全体を設計要素別に分類することで．非機能要求の内容について俯瞰した内容を理解するのに有効である．また設計要素別に分類することで要求仕様書内の非機能要求について俯瞰した事柄の概念について理解することができるため，非機能要求について理解するのに妥当であると考えられる．本手法は俯瞰した事柄の概念の理解を目的とし，トピック全体を設計要素別に分類すれば様々な要求仕様書内の非機能要求文に適用可能であると考えられる．

6. まとめ

本稿では，トピック分類技術である LDA を用いて，非機能要求の要約を提案した．本手法を用いることで俯瞰した事柄についての概念の理解に一定の効果を示すことができた．現状の提案手法では学習データとして非機能要求が定義されたデータセット要求文を用いたが，データセット要求文だけではなく様々な要求仕様書内の非機能要求の俯瞰した事柄を理解するのに有効であると考えられる．

謝辞

本研究は JSPS 科研費 JP19K11907 の助成を受けた．

参考文献

[1] M. Day and C. Chen, "Artificial Intelligence for Automatic Text Summarization," 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, pp. 478-484, 2018

[2] Rada Mihalcea and Paul Tarau, TextRank: Bringing Order into Texts Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp.404–411, 2004

[3] Makbule Gulcin Ozsoy and Ferda Nur Alpaslan, Ilyas Cicekli, Text summarization using Latent Semantic Analysis, Journal of Information Science, 37(4), pp.405–417, 2011

[4] ISO/IEC/IEEE 29148:2018, Systems and software engineering — Life cycle processes — Requirements engineering

[5] David M, Andrew Y, Michael I, Latent Dirichlet Allocation, Journal of Machine Learning Research Vol.3, pp.993-1022，2003

[6] GENSIM，https://radimrehurek.com/gensim/ （参照 2021-09-28）

[7] MeCab，https://taku910.github.io/mecab/ （参照 2021-09-28）

[8] PROMISE Software Engineering Repository Dataset http://promise.site.uottawa.ca/SERepository/ （参照 2021-04-23）