平成29年度 情報工学コース卒業研究報告要旨

山本 研究室	氏　　名	石 野　慎 弥
卒業研究題目	Word2Vecを用いたソフトウェアドキュメント検索の支援にむけた調査	

ソフトウェア開発において, 要求仕様書, 基本設計書といった, 多くのソフトウェアドキュメントが用いられる. プログラマやテスターはソフトウェアドキュメントに従ってコーディングやテストなどの作業を進める. プログラマやテスターは, 必要な情報を得ようとしてソフトウェアドキュメントを読み返すときがあるが, これはプログラマやテスターにとって手間であると考えられる.

ソフトウェアドキュメント内から必要な情報を得るためには, 通常ドキュメント内をキーワード検索する. しかし, 検索の手間が入力キーワードとして用いる単語のドキュメント内での登場回数に依存するという問題が存在する.

本調査では, Word2Vecを用いることでソフトウェアドキュメント検索の支援を行うことができるかを調べることを目的とする. Word2Vecとは, 2層から成り, テキスト処理を行うニューラルネットワークのモデルである. テキストコーパスを入力すると, 出力としてベクトルのセット, すなわちコーパスにある単語の特徴量ベクトルを返す. Word2vecの目的および有用性は, 類似語のベクトルをベクトル空間にグループ化することであり, すなわち数値に基づいて類似性を検知する. Word2vecは, 十分なデータ, 利用例, コンテキストが与えられれば, ある単語の意味の推測を, 過去の出現例を基に, かなり高い精度で行うことが出来る. Word2vecは, 分散した単語の特徴の数値表現であるベクトルを作成する. これらの推測により, ある単語と他の単語との関連性を確立することができる.

本調査では, 形態素解析ツールを利用して分かち書きされたソフトウェアドキュメントをテキストコーパスとして用いた. また, 調査にはWord2Vecの機能の1つである「word-analogy」を用いた. これは学習した単語のベクトルから, 類似の関係性をもつ単語を推測する機能である.

まず, 自然言語で記述された2つのソフトウェアドキュメントを対象に, ソフトウェアドキュメント中の本文あるいは表に含まれる内容からWord2Vecに与える単語の質問の組を作成できるか, 作成できた質問の組から特定の構成パターンが存在するかを調べた. 調査の結果, 2つのソフトウェアドキュメントからは, 本文と表からWord2Vecに与えるための質問の組を作成することができ, 実際に作成した質問から構成パターンを4つ, 対象のソフトウェアドキュメントからは作成できなかったが表の構成上作成できると考えられる構成パターン1つを作成することができた. これにより, ソフトウェアドキュメントからWord2Vecに与えるための質問の組を作成できるということが確かめられた.

次の調査として, 自然言語で記述された6つのソフトウェアドキュメントを対象に, 実際にWord2Vecに質問を与え, 得られた結果とキーワード検索による検索との比較を行った. この調査の目的は, キーワード検索よりもWord2Vecを用いた検索の方が効率的に検索を行うことができる場合があるかを調べることである. 調査の結果, 6つのソフトウェアドキュメントおよび前の調査で作成した5つの構成パターン全てにおいて, キーワード検索よりもWord2Vecを用いた検索の方が効率的に検索を行うことができる場合があることが分かった. これにより, Word2Vecをソフトウェアドキュメント検索の支援に利用できる可能性を示すことができたと考えられる.