5N-04

word2vec を用いた同義語辞書自動作成手法の提案と適用評価

伴　凌太†　　高橋　宏季†　　位野木　万里†

工学院大学†

1. はじめに

著者らの研究グループでは，自然言語で記述された要求仕様書の高品質化のために，あいまい表現となる同義語を自動的に指摘するツールの開発を試みてきた[1][2]．その開発過程において，エンタープライズ系のあるドメインに特化したシステムの要求仕様書を分析し，当該分野の同義語辞書を作成した．同義語辞書の作成にあたり，著者らは実際の要求仕様書から設計要素用語を抽出し，それらを手作業により整理分類して同義語の候補を洗い出す作業を行った．膨大な量の要求仕様書から，人手による同義語辞書の抽出は，多大な開発コストがかかることや，同義語とするかどうかの判断にはノウハウが必要であり初級の技術者には困難であるという課題がある．

そこで，本課題に対し著者らは，同義語辞書の作成作業を，単語のベクトル表現化ツールである word2vec[3]のサポートによって負荷を軽減する手法の開発を試みた．同義語が各ページに渡り複数回出現する要求仕様書に，単語の関連度を出力する word2vec を組み合わせることで，見落としや正誤の判断ミスのリスクを抑え，コスト軽減が可能だと考えられる．

以下，本稿は次のように構成する．2 章では同義語辞書の自動作成にあたり前提知識となる，同義語と要求仕様の一貫性検証支援ツールについて説明する．3 章では，word2vec を用いた同義語辞書自動生成手法を提案する．4 章にて同義語辞書自動生成手法を提案する．5 章では，提案手法の考察を述べる．6 章で本稿のまとめを示す．

2. 要求仕様書における同義語と同義語表現

2.1 同義語について

同義語とは，意味は同じだが表現が異なる用語である．実際の要求仕様書で使用されていた同義語の例を図 1 に示す．「格納」「保管」「保存」「書き込む」は，一般的には異なる意味を持つが，実際の要求仕様書中で同一の意味で使われていた．このように同一の意味で使われていたとしても，読み手の違いにより，a)～d)をすべて違う表現だと解釈する可能性がある．

2.2 要求仕様の一貫性検証支援ツール

著者らの研究グループでは，自然言語で記述された要求仕様書の高品質化のため，要求仕様書の一貫性検証支援ツール（以下，一貫性検証支援ツールとする）を開発した[1]．これは，要求仕様の品質特性である「一貫性」に着目し，「アクター」「データ」「画面」「振る舞い」の設計要素が，要求仕様書で一貫した定義で記述されているかを支援する．

一貫性検証支援ツールの開発過程において，人手により要求仕様書を分析して同義語辞書を作成した．人手による同義語辞書の作成には，多大な開発コストがかかることや，

Proposal of Automatic Construction Method for Analogous Term Dictionary Using word2vec and its Evaluation
†Ryota Ban, Hiroki Takahashi, Mari Inoki, Kogakuin University

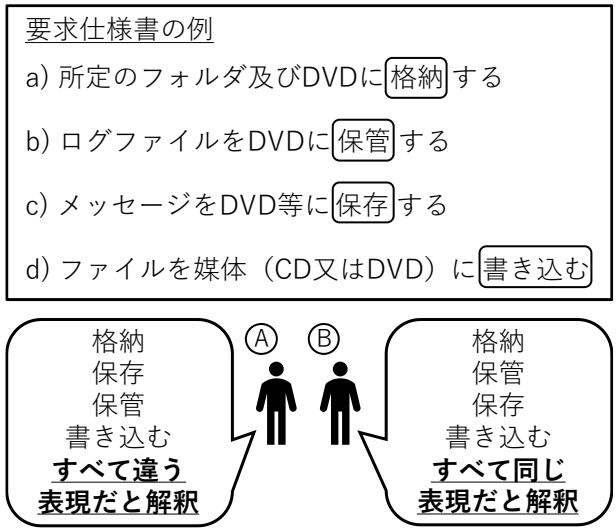

図 1　要求仕様書のあいまい表現と解釈

同義語とするかどうかの判断にはノウハウが必要であるため，初級技術者にはそのような辞書の作成は困難であると考えられる．

3. 同義語辞書自動生成手法の提案

word2vec を用いて同義語辞書を作成する手順は以下の通りである．

(1) 同義語辞書作成の素材となる振る舞い用語を抽出するため，一貫性検証支援ツールで分析を行い，振る舞い用語を抽出する．
(2) 振る舞い用語をベクトル表現化するため，word2vec に要求仕様書の学習データをインプットする．
(3) 近い単語を検索するプログラムファイル Distance に(1)で抽出した振る舞い用語を入力し，ベクトル表現化された振る舞い用語の関連語を出力する．
(4) 同義語には，類似している振る舞い用語を登録するため，出力した振る舞い用語の関連語でベクトル計算を行い，振る舞い用語のグループ間の類似度を抽出する．
(5) 各用語グループごとに，類似度が高い用語グループを紐づけし同義語候補とするため，対象とする用語グループに対し，グループ間の類似度の閾値が 0.4 以上の用語グループを同義語候補として抽出する．
(6) 対象となる用語グループに紐づけされた用語グループに紐づけされた用語グループまでを抽出し，同義語とする．
(7) (6)の作業を繰り返し行い，全ての用語グループに対して処理を行ったものを同義語辞書とする．

上記（1）～（7）の手順に従い，同義語辞書を作成した．

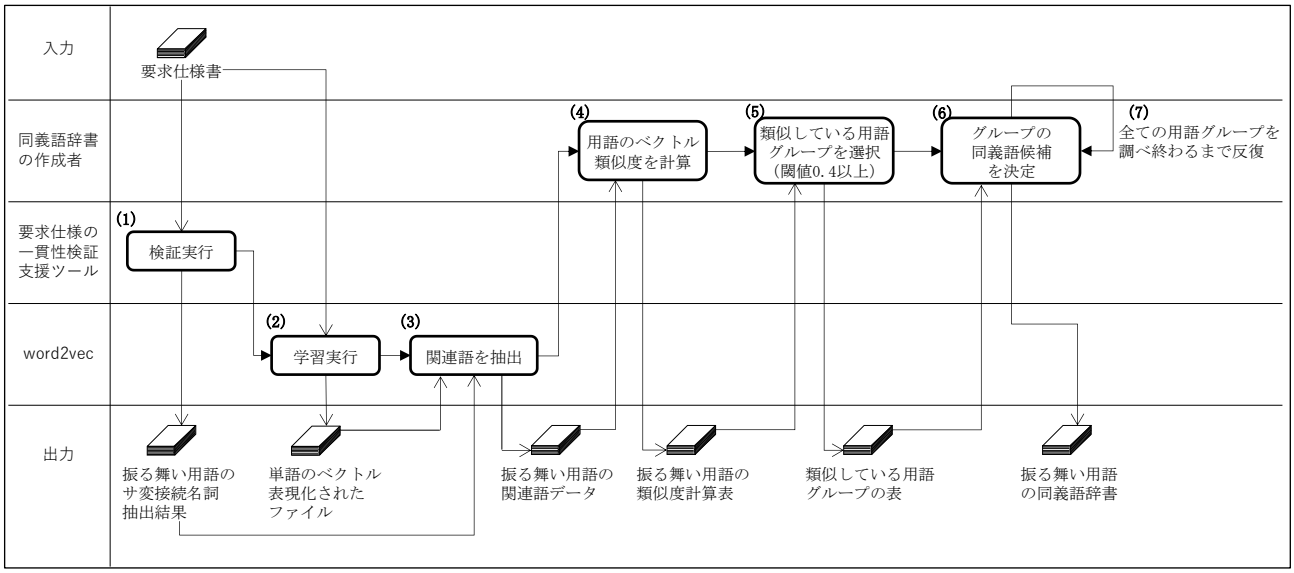

図 2　同義語辞書作成の手順

4.　結果

同義語辞書を作成した対象はエンタープライズ系システムの要求仕様書[4]（26 ページ，24,783 文字）のシナリオ記述を分析して抽出した，振る舞い用語である．なお本研究では，振る舞い用語の中でもサ変接続の名詞の単語のみを取り扱う．振る舞い用語の抽出は，MeCab を形態素エンジンとしている一貫性検証支援ツールを用いた．同ツールによる抽出の結果，振る舞い用語としては，延べ 2,026 件，正味 205 件を特定した．

表 1 に作成した同義語辞書の抜粋を示す．本同義語辞書は，同義語の種類としては 24 件，同義語用語数 181 件である．

表 1　振る舞い用語の同義語辞書（抜粋）

No.	同義語
1	依頼，通知，登記，報告
2	格納，登録，保管，保存，バックアップ
3	証明，背系，送付，納付，返却

5.　考察

今回の手法により構築した同義語辞書には，人手により構築した同義語辞書の内容を含んでいた．例えば，No.1 の「通知」「発信」や，No.2 の「格納」「保存」「保管」を特定できた．加えて，今回の手法により構築した同義語辞書は，人手での同義語辞書作成では指摘できなかった用語を指摘できた．例えば，表 1 に示す，No.2 の「バックアップ」や，No.3 の「送付」，「返却」は，実際の要求仕様書中でも同義語として利用されており，読み手の違いにより異なる解釈を誘発するリスクのある記述であった．

なお，今回の手法により構築した同義語辞書には，同一の意味ではないと考えられる用語も含まれている．これは，同義語として判断する際の類似度の閾値や，学習用のデータを変更することで調整できる可能性があるため，今後，継続して改善を行う．

6.　まとめ

本研究では，振る舞い用語の同義語辞書の作成方法に焦点を当て，自動作成手法の提案と適用評価を行った．提案手法により，手作業による同義語辞書作成では見落としていた用語を自動抽出するという効果があった．今後は，同義語辞書の内容の精度向上のために，同義語グループを判断するための閾値の調整，学習用データの拡張を行い，提案手法を改善する．

謝辞

要求仕様の一貫性検証支援ツール開発に関わる研究は，独立行政法人情報処理推進機構技術本部ソフトウェア高信頼化センター（SEC: Software Reliability Enhancement Center）が実施した「2015 年度ソフトウェア工学分野の先導的研究支援事業」の支援を受けたものである．また，本研究開発の一部は，2016 年度科研費「要求定義の高品質化のためのシナリオの一貫性検証・シナリオ生成手法」JSPS 科研費 JP16K00105 の助成を受けて実施した．

参考文献

[1] 高橋 宏季，野村 典文，近藤 公久，位野木 万里，要求仕様書における派生系アクター自動抽出手法：組織変更による影響対応への効果，情報処理学会，ソフトウェアエンジニアリングシンポジウム 2018 論文集，pp.121-129，2018

[2] 位野木 万里，近藤 公久，省略と修飾パターンを用いた用語不一致検証による要求仕様の一貫性検証支援ツールの実現と適用評価，日本ソフトウェア科学会，コンピュータソフトウェア，Vol.35，No.3，pp.109-127，2018

[3] word2vec，GitHub リポジトリ https://github.com/dav/word2vec（参照 2019-01-10）

[4] 厚生労働省，労働保険適用徴収システムに係るシステム運用業務一式 http://www.mhlw.go.jp/sinsei/chotatu/chotatu/shiyousho-an/120329-2.html（参照 2019-01-10）