要求仕様書におけるドメイン情報を使った多義語の意味の特定

2020SE054 山堂 美空
指導教員：佐伯 元司

1 はじめに

要求分析はシステム開発の最初の工程であるため，要求分析以降の工程に大きな影響を及ぼす．特に要求仕様書の曖昧性によって引き起こされる問題は，システム開発の工数やコストに大きな影響を及ぼすと考えられる．

要求仕様書の曖昧性には様々なものがあるが[1]，特にドメイン知識の差異について，読み手側のドメインでの意味と書き手側のドメインでの意味とが異なることによって発生する曖昧性がある．また，単純な要求文の構文解析などでは検出が難しい曖昧性であるため検出するための新しい手法を開発する必要がある．

我々が使用している自然言語の単語は本来「多義語」であるが，ドメイン知識に差異がない場合は読み手と書き手でどの意味かを1つに決めることができる．しかし，差異がある場合は，各々が解釈した意味が異なってしまう可能性がある．このような語句をここでは「曖昧な多義語」と呼ぶ．本論文では多義語と共起する単語の意味的な特徴の分析手法[2]を用いて，要求仕様書中に出現している「曖昧な多義語」を検出する手法を提案する．これにより，ドメイン知識に差異のある場合でも解釈を取り違えないようにすることができる．

2 多義語の語義分析

遊佐らは単義語と共起関係にある多義語の語義分析を，cosine 類似度と k-means クラスタリングアルゴリズムを用いて調査した[2]．これにより単義語と共起する多義語の語義識別に有効であることが分かった．この研究においては日本語全体を一つのドメインとして扱っているため要求仕様書にこの手法を適用するめには複数のドメインについて考える必要がある．

3 アプローチ

3.1 提案概要

Wikipedia から異なるドメインを記述した文書を収集する．それらの文書に出現する単語を収集し，その中で共通に出現する語を抽出して意味を分析する．共通する語の中で異なる意味で使用されていると判断されたものを多義語とする．得られた多義語と共起する語の各ドメインでの cosine 類似度を計算しクラスタリングする．これによって意味的に近い単語が一つのクラスタに入り，共起する語はそのクラスタが表す意味で使用されているとみなす．クラスタ内の語から，その語が使用されている要求文のドメインが判別可能であると考える．しかし図1のようにドメインを判別可能な語がないクラスタができた場合，そのクラスタ内の単語と共起する多義語がどのドメインで使用されているか特定できなくなり，「曖昧な多義語」であることになる．この「曖昧な多義語」を検出する．

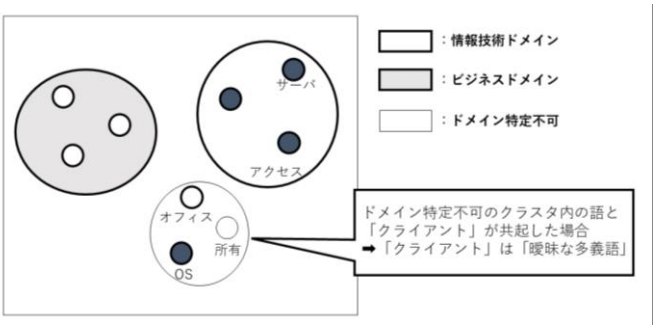

図 1 「曖昧な多義語」の判別方法

3.2 多義語の収集

各ドメイン内での単語の意味ベクトルを算出し多義語を収集するために，以下の処理をドメイン毎に行う．

1. 特定のドメインに該当する Wikipedia 内カテゴリから文書情報を取得
2. MeCab によって文書情報から形容詞，動詞，名詞のみを抽出
3. 加工した文書情報から Word2Vec によってドメイン内での単語の意味ベクトルを算出
4. ドメイン毎共通にかつ頻出している単語を収集し，それらの中で異なる意味で使用されているものを抜き出す．意味が異なるかどうかは，その単語の意味ベクトルと類似度が高い語を各ドメインから 10 語ずつ抽出しそれらの語の意味を比較することで判断する．

3.3 クラスタリング

収集した多義語に対して k-means クラスタリングアルゴリズムを用いて以下の処理を行う.

1. 多義語が持つ複数の意味を調査し, どのドメインでどの意味が用いられているのかを確認する.
2. 多義語に対して共起する語を収集し多義語と共起する語との cosine 類似度をドメイン毎に算出する.
3. 算出された cosine 類似度を k 個のクラスタに分け, クラスタごとの単語の意味にどのような特徴があるかを多義語がドメイン毎に持つ意味と照らし合わせて分析する. 分析の結果, あるクラスタ内で複数のドメインの単語を含むものになっていれば, 多義語がそのクラスタ内の単語とともに使用された場合「曖昧な多義語」であると判定する.

4 実験の実施

提案手法の有効性を調査するため, ドメイン固有の単語を用いて先述した処理で実験を行う.

「情報技術」「ビジネス」「食品」「日本の行政」の4つのドメインを対象として要求仕様書中の 1357 単語のうち情報技術ドメインに関する 239 単語から多義語を収集した結果, 次の表1の語が得られた. これらの単語に対してここでは k=10 とし k-means クラスタリングを行った結果を 5 節で説明する.

表 1 収集した多義語の一覧

アクセス	クライアント	クラス	対応	管理
リスト	プログラム	運用	基本	機能
テーブル	パッケージ	出力		

5 結果と考察

表 2 には「クラス」に対してクラスタリングを行った結果を示す. 「クラス」の意味は次の表 2 である.

表 2 「クラス」のドメインごとの意味

	意味	意味に対応するドメイン
1	オブジェクト指向	情報技術ドメイン
2	学級	(tf*idf の数値により対象外)
3	等級, 階級	ビジネスドメイン

クラスタは全部で 10 個であるが, ここではそのうちの 4, 6 のクラスタについて説明する. 表 3 に示す通りクラスタ ID4 番のクラスタには「単位」や「サイズ」といった階級の意味を示す語が集まっている. また, 6 番には情報技術ドメインの語が集まっている.

この結果から, 「クラス」という単語は 4 番のクラスタに属する単語と共起すると「階級」の意味となる可能性が高く, 6 番のクラスタに属する単語と共起するとオブジェクト指向で用いられる意味になると考えられる.

表 2 クラスタ ID とクラスタ内の単語

クラスタ ID	4	6
単語	それぞれ	ログ
	単位	修正
	サイズ	反映

クライアントについても同様な手法で, クラスタ0={ソフトウェア, システム, サーバ}, クラスタ 4={オフィス, OS, 所有}が得られた. 情報技術の意味で使われていることは判別可能であるが, ビジネスの意味では OS は入っているためこれと共起すると判別不能となる可能性があることがわかった.

実験の結果から提案手法が大まかな意味特定に有用であることが分かったが, 共起する単語の中には意味特定につながらない単語も含まれていた. この要因として共起する語が文章全体を対象としており多義語からの距離が遠いものも含まれていることが考えられる.

6 今後の課題

現在の手法では意味特定につながらない単語を含むクラスタが存在しているため, 意味特定の精度向上のために共起する語の多義語からの距離を加味することが必要だと考えられる.

参考文献

[1]D.M. Berry, et. al, ”From Contract Drafting to Software Specification:Linguistic Sources of Ambiguity”, November 2003, https://cs.uwaterloo.ca/~dberry/handbook/ambiguityHandbook.pdf

[2]遊佐他, 「単義語と共起する多義語に対する分散表現を利用した語義分析」, 言語資源活用ワークショップ発表論文集 p.216-222(2017)/ http://doi.org/10.15084/00001522