Vol.2012-MPS-90 No.3
2012/9/19

仕様書における文脈情報を考慮した同義語の抽出支援手法

川井 康示[1]　吉川 大弘[1]　古橋 武[1]　平尾 英司[2]　久野 綾子[2]　五藤 智久[2]

概要：仕様書を正確・明確に記述することは常に求められてきたことである．しかし，大規模なプロジェクトや業務などでは，複数の担当者が仕様書を分担執筆することが多く，言葉の定義の不統一性などから，読み手が，書き手の意図とは異なった解釈をしてしまうことが問題となっている．そこで，本研究では，語句の表記は異なるが，同じ意味で使われる語句“同義語”に着目し，文書における同義語候補をユーザに提示し，仕様書の作成支援を行う手法を提案する．従来の同義語検出手法として，語句の文脈情報をベクトルとして表し，語句間の類似度をコサイン類似度などの類似度指標を用いて定量化を行うものがあるが，本研究では，それに対し複数の手法を取り入れることで，仕様書中からより適切に同義語候補が抽出可能となることを示す．性能評価には，実際に公開されている仕様書において，特定の語句について，その半分を表記の異なる別の語句に置換することで人工的に正解同義語を作成し，提案手法の有効性を検討する．

キーワード：仕様書，同義語，文脈情報，語句の集約，支配力

Abstract: To describe specification document exactly and clearly has been required in any fields and scales. Specification documents are used for the technical transfer and inheritance of manufactures and services. However, the description and the meaning of the component words in specifications are often inconsistent or multiple, because a specification document is made by the persons in charge of various parts. Then the readers may misunderstand the contents of specifications by them. This paper focuses on synonyms, multiple description of words for a meaning or a word, in specification documents and proposes an extraction method of them considering the co-occurrence words of component words. This paper applies the proposed method to a test data, in which some words in an actual specification document are replaced with another words, and studies the effectiveness of the proposed method.

Keywords: Specification document, Synonyms, Contextual information, Aggregation of Words, Domination Weight

1. はじめに

仕様書を正確・明確に記述することは，その分野や規模に関らず，常に求められている．しかし，大規模なプロジェクトや業務などでは，複数の担当者が仕様書を分担作成 (執筆) することが多く，表記の違いなどから，誤解を与えてしまう仕様書が作成されてしまう事例が後を絶たない．例えば，言葉の定義が不統一となり，異なる表記の語句が同じ意味として使われることや，逆に同じ表記の語句が複数の意味で用いられてしまうことがある．読み手は，そのような表記の曖昧性から，書き手の意図とは異なった解釈をしてしまう可能性がある．

[1] 名古屋大学
Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
[2] 日本電気株式会社
NEC Corporation, 1753, Shimonumabe, Nakahara-Ku, Kawasaki, Kanagawa 211-8666, Japan

本稿では，語句の表記は異なるが，同じ意味で使われる語句“同義語”に着目する．同義語を自動的に抽出する研究は，これまで数多く行われてきた [1] [2]．その中のアプローチとして，シソーラスなどの辞書を用いて類似語を抽出する方法がある．しかし仕様書では，シソーラスに含まれない専門用語や複合語が多く用いられ，むしろそれらの辞書では未知語として扱われる語句が同義語となる場合が多い．また別の方法として，同じ意味として用いられている同義語ならば，類似した文脈情報を持つという仮定に基づいた抽出方法がある [3][4]．この方法では，各語句の文脈情報を，文書中での語句間の共起関係を表した“共起ベクトル”により表現する．そして，語句間の類似性を，コサイン類似度などの類似度指標を用いることで定量化する．しかし，意味や用法の類似性を考慮しないで，語句の単純な一致のみしかみていないため，文脈情報の適切な定量化が困難という問題点がある．

本研究では，上記で説明した文脈情報の類似性を用いて，

ある文書中のみでしか出現しない語句間の類似性を抽出する手法を提案する．このとき，仕様書特有の複合語については，意味を持つ最小の単位である形態素に分解する．さらに，意味の類似性を考慮するために，共起ベクトルの要素となる各語句を，シソーラスに基づき集約する．複合語については，各形態素の持つ意味の強さを考慮することで，文脈情報をより適切に定量化する．

本稿では，厚生労働省による公開仕様書において，出現頻度の異なる特定の語句について，その半分を表記の異なる別の語句に置換することで人工的に正解同義語を作成し，提案手法の有効性を検討する．

2. 関連研究

同義語を自動的に抽出する研究は，これまで数多く行われてきた．[5] では，文脈情報として，語の前後の局所情報を用いて，略語とその原型語の対応付けを目的とした同義語検出を行っている．[6] では，4 種類の文脈情報特徴（動詞に係る主語，動詞に係る目的語，名詞に係る形容詞，および前後連続している隣接語）に異なる重みを加える手法の有効性を検討するとともに，単語類似度ネットワークを用いて，一度算出した単語対をリランキングする手法を提案している．また，[7] では，大規模コーパスから類語を抽出する際に，広範囲の語と共起する語が類似度計算におけるノイズとなるという前提のもと，このノイズの低減手法を提案している．これらの研究は，同義語を抽出するという部分は本研究と同じであるが，対象が特定分野の専門語の抽出ではなく，そのための対応を行っていないという点が異なる．特定分野の専門語の抽出を目的としたものには [8] がある．[8] は，分野独特の同義語の抽出を目的として，システムにクエリを与えることで，同義語候補語を提示するツールを提案している．この研究は [5] の論文を日本語に拡張し，文脈情報の使用に工夫を加えたものであり，本研究と同じく，特定分野の専門語の抽出を目的としているが，複合語を 1 つの語として扱っている点が異なっている．

3. 提案手法

3.1 語句ベクトルの作成

同義語検出の対象となる全ての語句に対して，それと共起する語句のベクトルを作成する．本稿では，ここで作成した共起ベクトルを，“ 語句ベクトル ” と呼ぶ．ここで共起の範囲としては，同じ文中に出現する自身以外の語句とする．また，ベクトルの要素としては，文中で出現した全ての名詞，動詞，形容詞とし，各要素の値は，それらの出現回数とする．このとき，専門用語などに代表される仕様書に特有の語句などは，複数の形態素で構成される複合語として出現することが多く，それらの語は形態素に分割されてしまうことで，文書中の本来の意味とは異なる語句となってしまう．そのため，本手法では名詞が連続した場合には，形態素を再結合し，“ 複合語 ” とする．

3.2 支配力の算出

複合語を，意味を持つ最小の単位である形態素で分割し，それぞれの形態素が持つ影響力の算出を行う．本稿では，この影響力の強さの指標を，“ 支配力 ” と定義する．支配力は，ある形態素が複合語の一部となることによる意味（共起情報）のばらつき度によって表され，「支配力が強い」＝「どのような形態素と複合しても意味（共起）が同じ」，逆に「支配力が弱い」＝「複合語ごとに意味（共起）が異なる」ということを基本的な指標としている．式 (1) に，支配力の算出式を示す．対象となる形態素が含まれる全ての複合語に対する語句ベクトル群において，2 つ以上の複合語に対して出現（共起）した共起語（ベクトルの要素）の数（重複共起語句数）を，いずれかの複合語 1 つに対しても共起した全共起語句数（語句ベクトル群の要素数）で割ったものとなる．このように，ある形態素の語句ベクトル群において，重複共起語句数と全共起語句数の比をみることで，ある形態素が含まれることによる語句ベクトルのばらつき度合い，すなわち，ある形態素の影響力の強さを定量的に表すことが可能となる．

$$\text{支配力} = \frac{\text{重複共起語句数}}{\text{全共起語句数}} \tag{1}$$

表**表 1** に，形態素「変更」が含まれる複合語の語句ベクトル群の例を示す．この例では，重複共起語句数は 5（管理，伴う，提案，検討，実施），全共起語句数は 6 となり，「変更」の支配力は約 0.83 と計算できる．以下に，「変更システム」という複合語において，「システム」の支配力を 0.25 として，複合語の重みを計算した例を示す．ここでは，重みの合計が 1 となるように正規化を行っている．

$$\begin{aligned} \text{“変更”} : \text{“システム”} &= \frac{0.83}{0.83+0.25} : \frac{0.25}{0.83+0.25} \\ &= 0.77 : 0.23 \end{aligned} \tag{2}$$

表 1 語句ベクトル群の例
Table 1 Examplse of word vectors

複合語	管理	伴う	設定	提案	検討	実施
変更内容	1	3	0	1	1	1
変更システム	0	0	0	1	0	0
追加変更	0	1	0	0	0	1
変更禁止期間	0	0	1	1	0	0
組織変更	1	1	0	0	0	1
変更要望	0	0	0	0	1	1

3.3 概念ベクトルの作成

語句ベクトルの要素となる各語句を，意味のまとまりで集約し，新たな共起ベクトルを作成する．ここで，シソーラスにおける意味のまとまりを基に，語句ベクトル中の語句を集約したものを“概念”[9] と定義し，概念を要素として構成された共起ベクトルを“概念ベクトル”と呼ぶ．本研究では，「日本語大シソーラス」[10] を用いる．本シソーラスは 3 つの階層的グループを持っており，最も上位の階層のもの（“経済”など）を“大概念”，中位の階層のもの（“価格・コスト”，“お金・資本”など）を“中概念”，下位の階層のもの（“価格”，“高価”など）を“小概念”と呼ぶ．小概念は語句が集約される最小の単位のグループであり，小概念とそれに属する語句は，その上位の概念である中概念に属している．また，語句を大概念，中概念，小概念でそれぞれ集約した共起ベクトルを“大概念ベクトル”，“中概念ベクトル”，“小概念ベクトル”と呼ぶ．

本稿で用いるシソーラスでは，多くの場合，「変更システム」のような複合語は登録されていない．そこでここでは，概念ベクトルの要素としての複合語と，所属する概念との対応について述べる．複合語に対しては，3.2 で算出した支配力を用いて概念ベクトルを作成する．3.1 では，複合語を 1 つのベクトル要素として扱ったが，ここでは，複合語の構成要素となる形態素がそれぞれ属する概念に，3.2 で算出した各構成語の重みを考慮して値を付与する．例えば， 3.2 における「変更システム」という構成語（変更：システム=0.77：0.23）においては，「変更」は“改める”1 個，「システム」は“コンピュータ”，“情報科学”などといった 20 個の小概念にそれぞれ属しているため，ある語句が「変更システム」と 3 回共起した場合，“改める”という概念に 3×0.77=2.31，“コンピュータ”，“情報科学”などといった概念には 3×0.23=0.69 という概念ベクトルの要素の値を付与する．

3.4 類似度計算

共起ベクトル間の類似度計算式として，コサイン類似度に調整項を加えた式 (3) を用いる．式 (3) において，k は低頻度語の考慮のための調整項であり，低頻度の語句が高い類似度になり易いという傾向が強く見られることから付与している．k は，極めて低頻度の語句には強く影響を与える（類似度の値を下げる）が，高頻度の語句においては，得られる類似度の値は通常のコサイン類似度（式 (3) において k のない形）とほぼ同値となる．また，式 (3) の意味合いとしては，比較する 2 つのベクトルに，長さ k の直交するベクトルを付与した場合のコサイン類似度と同様のものとなる．

$$\frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\sqrt{||\boldsymbol{x}||^2 + k^2}\sqrt{||\boldsymbol{y}||^2 + k^2}} \quad (3)$$

3.5 同義語候補のフィルタリング

本手法では，様々なルールを加えることで，同義語候補となる語句のペアの絞込みが可能となっている．以下にルールの一覧を示す．

(a) 同文中に出現するペアの除外

同文中に 1 度でも出現したペアを同義語候補から除外する．同文中に属する語句は，その性質上，類似した共起語を持つことになるため，共起ベクトルが類似しやすくなる傾向がある．一方で，同義語（同じ意味で異なる表記）が同文中に出現するケースは，略語表現（CD (Compact Disk)，など）として出現する以外ではほとんど生じ得ないと考えられる．そのため，同文中に出現する語句を候補から除外することで，類似度が上位かつ不必要な同義語候補ペアの多くを除外することが可能となる．

(b) 品詞細分類が一致するペアの除外

品詞細分類が一致しないペアを同義語候補から除外する．品詞細分類の例としては，“名詞－一般”，“名詞－サ変接続”，“名詞－形容動詞語幹”などといったものがあり，これらが一致するペアはより用法が近いと考えられ，逆に品詞細分類が異なるペアが同義語とはなりにくいと考えられる．なお複合語の場合は，複数の名詞が組み合わさっているため，末尾の構成語の品詞細分類を用いる．

4. 実験

4.1 同義語の埋め込み

本手法の有効性を検討するため，仕様書の例として厚生労働省により公開されている運用業務委託仕様書 [11] を用いて同義語抽出実験を行った．本稿では，正解情報となる同義語を人工的に生成するために，出現頻度の異なる特定の語句について，50 %の確率で別の語句に置換した．表 2 に，置換処理前の語句とそれらを置換することによって作成した同義語を示す．表 2 で，() 内は文書中での出現回数を表す．ここでは，高頻度の語句として「管理」と「運用受託者」，中頻度の語句として「作成」と「障害」，低頻度の語句として「検討」と「保管場所」のそれぞれ 2 つずつを用いた．以下の実験では上述のように，ランダムに置換することで正解同義語を作成し，評価することを 10 試行繰り返して評価を行った．

表 2 埋め込み同義語
Table 2 Inserted synonyms

	置換処理前	作成した同義語	
A	管理 (164)	管理	監督
B	運用受託者 (117)	運用受託者	運営受託者
C	作成 (39)	作成	策定
D	障害 (33)	障害	弊害
E	検討 (14)	検討	思案
F	保管場所 (15)	保管場所	保存場所

4.2 概念の集約方法の比較

最初に，3.3 で述べた概念による集約の効果を検討するため，語句ベクトルのまま類似度計算を行った場合と，概念ベクトルで集約を行った場合との，同義語抽出性能の比較を行った．ここでは，式 (3) の類似度指標における k の値を 0 としている．また，支配力は考慮せず，複合語には各構成語に等しい重みを用いている．表 3 に，表 2 で示した正解同義語のそれぞれの類似度と，すべての候補ペアに対して，類似度の高い順に並べた際の順位を示す．ここで，左の列の A〜F は表 2 で示した各正解同義語のアルファベットと対応しており，表中の上の数字は式 (3) で算出される類似度の値の 10 試行の平均，下の値は順位の平均となっている．このとき，全てのペア数は 2,593,503 であり，これは，例えば“語句ベクトル”において，正解同義語 A,“管理”と“監督”は 2,593,503 の候補の中で，平均して 100,408 番目に同義語候補として出現したことを意味する．

表 3 より，小概念ベクトルを用いた場合では，語句ベクトルと比較して順位が上昇していない場合もあるものの，中概念ベクトル，大概念ベクトルでは，すべての頻度の同義語について，順位が上昇していることがわかる．実際に，同義語のチェックを行う場合は，候補となるペアについて，順位の高い順にみていくことになるため，類似度ではなく，順位が上昇することが重要となる．このように，概念による集約を行うことで，語句そのものではなく，意味の上で類似した共起情報，すなわち文脈の類似性を抽出することができ，今回のように文書データ量が少ないスパースな共起ベクトルに対しても，適切に共起ベクトルの類似性が抽出できていることがわかる．以降は，表 3 において，最も安定して高い順位で同義語候補が得られている大概念ベクトルを用いて実験を行う．また，類似度は省略し，対象ペアの順位の平均のみを示す．

表 3 概念の集約方法の比較
Table 3 Comparison of aggregation method by concepts

	語句ベクトル	小概念ベクトル	中概念ベクトル	大概念ベクトル
A	0.724 100408	0.962 1508	0.965 1489	0.984 1422
B	0.926 2597	0.975 1218	0.975 1232	0.989 1120
C	0.511 103675	0.938 2278	0.943 2282	0.976 2174
D	0.716 11869	0.878 12326	0.883 8377	0.958 7937
E	0.702 14714	0.817 60437	0.887 7859	0.937 12919
F	0.491 107743	0.723 166343	0.785 75773	0.880 106108

4.3 支配力の効果の検討

次に，3.4 及び 3.5 で示した手法の効果を検討するため，以下の条件で比較実験を行った．

条件 1：式 (3) の $k = 0$
条件 2：式 (3) の $k = 3$
条件 3：式 (3) の $k = 3$，3.5 のフィルタリングルール (a),(b) の適用

表 4 に，支配力を用いずに，複合語の各形態素に等しい重みを与えた場合の平均順位を示す．このとき，条件 1，条件 2 における全てのペア数は 2,593,503，またその中で条件 3 を満たすペア数は 618,297 であった．表 4 から，条件 2 では，低頻度語については順位が低下したが，高頻度語については大きく順位が上昇していることがわかる．今回，調整項 k の値として 3 を用いたが，抽出する同義語の出現回数に応じて適切な値を設定することで，低頻度な候補の順位も上げられることが期待され，調整項 k の適切な値の決定法が今後の課題として挙げられる．また，表 4 から，条件 3 により，すべての同義語候補において順位が大きく向上していることがわかる．このことから，提案手法によるフィルタリングルールが有効であると考えられる．

表 4 支配力を考慮しない場合の平均順位
Table 4 Average rank order without consideration of domination weight

	条件 1	条件 2	条件 3
A	1422	223	5
B	1120	73	2
C	2174	844	82
D	7937	5504	857
E	12919	24624	4524
F	106108	149019	29892

表 5 に，3.2 で示した支配力を用いて，複合語の各形態素に重みを考慮した場合の平均順位を示す．括弧内の数字は表 4 と比べた際の順位の変化を表している．表 5 より，支配力を考慮することで，多くの正解同義語において，順位が上昇していることがわかる．このように順位が上昇した理由として，複合語を形態素に分割してしまうことで意味が失われてしまう語句に対して，適切な重みを表現することができたことが考えられ，本稿で用いた支配力が有効であったと考えられる．

5. まとめ

本稿では，シソーラスを基にした場合では抽出することが難しい，仕様書中の同義語の検出手法を提案した．また本稿では，厚生労働省による公開仕様書において，出現頻度の異なる特定の語句について，その半分を表記の異なる別の語句に置換することで人工的に正解同義語を作成し，

提案手法の有効性を検討した．今後の課題としては，他の支配力の算出手法や，調整項 k の適切な値の決定法，また効果的なノイズ同義語候補の除外法などに対する検討が挙げられる．また，他の仕様書についても，提案手法を適用し，その有効性を検討していく予定である．

表 5 支配力を考慮した場合の平均順位
Table 5 Average rank order considering domination weight

	条件 1	条件 2	条件 3
A	1382(-40)	152(-71)	4(-1)
B	1130(10)	73(0)	3(1)
C	2147(-27)	1184(340)	33(-49)
D	5307(-2630)	2522(-2982)	671(-186)
E	19575(6656)	12458(-12166)	2873(-1651)
F	93282(-12826)	51768(-97251)	16768(-13124)

参考文献

[1] Dekang Lin: Automatic retrieval and clustering of similar work, Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational linguistics (COLING-ACL'98), pp.786-774 (1998).
[2] P. D. Turney: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, ECML-2001, pp.491-502, Freiburg, Germany (2001).
[3] Harris Z: Distributional structure, Word 10 (23), pp.146-162 (1954).
[4] Lee, L.: Measures of distributional similarity, Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pp.25-32 (1999).
[5] Akira Terada, et al.: Automatic expansion of abbreviations by using context and character information, Inf. Process. Manage., 40(1), pp.31-45 (2004).
[6] 王玉馨，清水伸幸，吉田稔，中川裕志：単語類似度ネットワークを通じた自動同義語獲得，自然言語処理研究会報告 2008(46)，pp.7-14 (2008).
[7] 相澤 彰子：大規模テキストコーパスを用いた語の類似度計算に関する考察, 情報処理学会論文誌, 49(3), pp.1426-1436 (2008).
[8] 寺田昭，吉田稔，中川裕志：同義語辞書作成支援システム，自然言語処理 15(2)，2008-04，pp.39-58 (2008).
[9] 小林大輔，吉川大弘，古橋武：概念を用いた HK Graph によるテキスト解析支援，日本感性工学会論文誌，Vol. 11，No. 2(Special Issue)，pp.159-165 (2012).
[10] 山口翼：日本語大シソーラス〜類語検索大辞典〜，CD-ROM 版，大修館書店 (2006).
[11] 厚生労働省：平成 21 年度がん対策情報センターシステム運用業務委託仕様書,
http://www.mhlw.go.jp/sinsei/chotatu/chotatu/kankeibunsho/090123/