Vol.2019-NL-241 No.2
2019/8/29

共起情報を利用した不具合事象の同義表現獲得

川村晋太郎[†1]

概要:製造業の品質保証業務において，不具合の原因究明，対策内容検討や再発防止に活用する為，コールセンターログ（お客様からの問い合わせやクレームにおいて使用された表現をそのまま書き留めたもの），保守記録，保守ナレッジ共有サイトなどの多様なデータソースから解決の糸口になりそうな過去の情報を活用している．数百万件規模の多様なデータから所望のデータを検索する際に，その検索精度を向上させる為，対象製品の部品名，箇所名や不具合現象などの同義（言い換え）表現を得る必要がある．同義表現については，コーパス全体から分布仮説に基づいた類似性指標（コサイン類似度など）で獲得する手法がよく知られているが，同じ単語の同義表現であっても，実際に文書内で使用される言い回しや表現方法は，不具合事例により異なることも多い．本稿はこのような同義表現獲得の問題点に鑑み，ある単一の用語に対する同義表現を獲得するのではなく，不具合やトラブルを表す“事象”の表現が「箇所名+現象名」で成り立つことに着目し，それらをセットにした際の互いの関連度・共起度によって，「箇所名」及び「現象名」の同義表現を同時に獲得していく手法の開発を試みた．

Acquisition of synonyms of trouble events by co-occurrence information

SHINTARO KAWAMURA[†1]

1. はじめに

製品開発において，その市場品質を確保することは製造業における重要なタスクとなっている．具体的な課題としては，不具合がどういう傾向にあるか（予兆・影響範囲の予測），その可視化と対応，コールセンターの内容把握，過去トラブルの再発防止等が挙げられる．そのために，社内に蓄積されている様々なデータソース（市場保守記録/コールセンターログ/トラブル報告書/設計書 etc.）から得た有益な情報を効率的に活用することで，品質向上や信頼性の高い製品開発を支援するソリューションが考えられている．

しかし，上記データは一般的に自然言語で記述されており，専門用語知識や特定ドメインの同義（言い換え）表現が必要となる．例えば，設計書をチェックリストや不具合事例と照合して，関連不具合情報や注意事項を自動提示する設計品質向上支援システムでは，チェックルールを作成する際に，対象製品の仕様書や不具合報告書中に記載される部品名や不具合現象などの専門用語辞書が必要になる[1]．高橋ら[2]は要求仕様書の品質低下につながるリスクのある振る舞い用語の同義語について分析し，同義語辞書を作成している．同様にデータソースから所望のデータを検索する際において，文書とクエリで異なる語が現れる問題は語彙の不一致として知られ，シソーラスを使ったクエリ拡張によって対処されてきた[3].

同義表現の獲得には様々な方法が提案されているが，同義語名詞なら似ている文脈情報を持つという分布仮説[4][8][9]に基づく手法がよく知られている．同義語獲得では，一般的に以下の３ステップでこの仮説は実行される[5]．まず，第一段階の処置では，コーパスから抽出された重要度の高い単語の文脈情報における統計情報を抽出し，各単語はこれらの文脈特徴のベクトルによって表される．第二段階の処理では，コサイン類似度などの類似性量度を選び，それをクエリ単語と同義語候補の単語対に適用して，類似度を計算する．類似度の降順で各クエリ単語の同義語候補リストを作る．最後に同義語リストからトップ候補を選んで，クエリ単語の同義語と認定する．

我々は市場障害が発生した際に市場への影響・規模等を調査する目的で，類似する過去の不具合事例や保守案件を検索する為に必要な同義表現について，分布仮説に基づいた獲得を試みた．対象として製品の保守情報（コールセンターログや保守記録に記載されたテキスト情報）をコーパスとした分布仮説による同義表現獲得実験を行った結果，テキスト内で実際に用いられる同義表現は対象とする不具合事象により異なることを認識した．具体的には不具合事象が「箇所名+現象名」で構成されるとした場合，箇所名または現象名の同義表現は，その対となる現象名または箇所名によって変化する．これは分布仮説による単なるクエリ拡張では検索ノイズが増大することを意味している．このような問題に鑑み，本稿では「箇所名+現象名」のペアを入力とし，「箇所名」及び「現象名」の同義表現をセットで獲得していく手法を紹介する．

†1 株式会社リコー
RICOH COMPANY, Ltd.

以降 2 章では同義表現の自動獲得及び不具合事象に関する用語抽出の先行研究を紹介し，3 章で不具合事象を対象とした同義表現獲得手法について説明し，4 章で参考データとして人手で同義表現を選定する手法，分布仮説（word2vec）による手法及び提案手法による同義表現獲得の実験結果を示し，5 章でまとめる．

2. 関連研究

近年，同義表現の自動獲得に関する研究が盛んに行われている[3][6]．中でも Bannard & Callison-Burch らの手法[7]は，まずアライメントのとれた二言語対訳コーパスを用意して，同じ単語とアライメントのとれた単語を同義表現と見なした．例えば日本語の「二酸化炭素」と「炭酸ガス」は，両方とも英文中で「carbon dioxide」とアライメントがとられることが多い．このとき「carbon dioxide」をピボットとして，「二酸化炭素」と「炭酸ガス」が同義表現になっていると見なせる．対訳コーパスから同義表現を得る手法は，人手で辞書を整備する必要がない上，言い換えらしさを示す言い換え確率付きで大量に同義表現を得ることができる．

“類似する文脈でよく使われる表現は似た意味を持つ”という分布仮説に基づく同義表現獲得においては，係り受け関係のある名詞と動詞の対を抽出し，名詞ごとに係り受け関係にある動詞の頻度を数え，共起語ベクトルを作成し，与えられた名詞に対し，共起語ベクトル間の類似値が高い順に名詞を出力する手法がある[10]．Qiu & Frei[11]や Schutze ら[12]は，語の共起関係からシソーラスを構築してクエリ拡張する手法を提案している．これらの手法は同一トピックの語によるクエリ拡張であり，Mandala ら[13]はこうした共起を元にしたクエリ拡張を，WordNet などの人手で構築したシソーラスによるクエリ拡張と組み合わせることでより精度の高い検索結果を達成している．

特に分布仮説について，荻原ら[14]は役に立つ文脈情報を調べ，係り受け関係（動詞の主語と目的語）および名詞の修飾語などの有効性を示している．寺田ら[15]は名詞の隣接語を文脈情報として用い，日本語の航空安全報告レポートから日本語の同義語を自動的に得ている．

さらに不具合情報から用語を抽出する技術に関して，大和ら[16]は故障及び不具合を表す表現は発生個所名と現象名で成り立つ[17]という考えに基づき，自然言語処理技術により「箇所名+現象名」で故障の情報を自動抽出する方法を導入している．不具合に関する文章では，係り受け関係に注目すると，不具合が発生した箇所とそこで起こった現象を表す文節のペアを発見できる（ルールがある）ことがわかっている．本稿においては，人手で同義表現を選定する際この手法を参考にしている．

また大森ら[18]は，不具合事例文から，不具合に関与する製品や部品の記述を特定・抽出する手法について提案している．ここでは，小町ら[19]による名詞の事態性の判別を利用している．事態性とは名詞が事態を示す用法で使われている状態のことであるコト（事態）を指すかモノ（物体）を指すかという意味的な違いに対応する．本研究で抽出しようとする語も大森ら[18]と同様，不具合にまつわる事態に登場する製品や部品を表す語であり，モノを表す名詞であるので，事態性をもたない名詞である．したがって，注目語の後ろのサ変名詞有無といった事態性なし判別に効果が高いパターンを利用する．

3. 提案手法

本節では，「箇所名」及び「現象名」の対を不具合事象として，それらの同義表現をセットで獲得していく手法を詳述する．

3.1 同義表現獲得プロセス

「箇所名」及び「現象名」の同義表現獲得のプロセス概要を以下に示す．

(1) テキストデータ全体に対して，word2vec[20]で学習する
(2) 元となる「箇所名」及び「現象名」の単語ベクトルを(1)の結果から獲得し，コサイン類似度降順で同義表現候補を取得する
(3) 取得した同義表現候補に対して，「箇所名」及び「現象名」の抽出ルールを適用する
(4) (3)で絞り込まれた同義表現候補に対して，「箇所名」×「現象名」でベクトル内積を算出する
(5) (4)で算出されたベクトル内積値上位を対象不具合事象の同義語表現とする

(1)におけるテキストデータは特に不具合事象の表現が多く含まれていると思われるコールセンターログ，保守記録，保守ナレッジ共有サイトなどを想定している．(2)では単語のベクトル表現化手法である word2vec を用いて，自動で文脈特徴のベクトルを獲得する．以降の節で(3)～(5)の各プロセスについて詳述する．

3.2 不具合事象の箇所名及び現象名抽出

社内のコールセンターログ及び保守記録から同義表現候補を人手で選定する際（4.2 節参照），データ解析結果から“現象情報は箇所に付随している”ことが判明し，手がかりとして 名詞+動詞/サ変名詞/形容動詞語幹 というルールを与えて処理した．尚，図 1 に示すように本ルールでは前方の名詞が箇所，後方の動詞/サ変名詞/形容動詞語幹が

現象の候補となる．

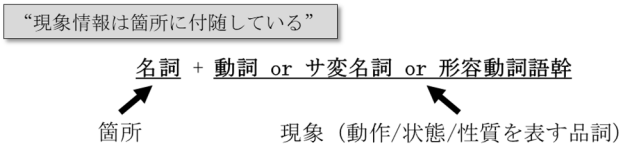

図 1　不具合事象の箇所名及び現象名抽出ルール

3.3 箇所及び現象の共起度計算

本稿では，同義表現獲得に“箇所”及び“現象”の共起を考慮することが特徴となっている．単語の共起度を表す良く知られた方法として，2 つの要素間の関連度合いを測る尺度である自己相互情報量（Pointwise Mutual Information, PMI）がある[21]．ここでxとyを単語とすると，

- x, yがそれぞれ出現する確率は$P(x), P(y)$
- x, yが同時に出現する確率は$P(x, y)$
- x, yがそれぞれ出現する回数は$C(x), C(y)$
- x, yが共起する回数は$C(x, y)$
- Nをコーパス全体の単語数

と表され，PMI の式は以下となる．

$$\begin{aligned} PMI(x,y) &= \log_2 \frac{P(x,y)}{P(x)P(y)} \\ &= \log_2 \frac{\frac{C(x,y)}{N}}{\frac{C(x)}{N} \cdot \frac{C(y)}{N}} \\ &= \log_2 \frac{C(x,y)N}{C(x)C(y)} \end{aligned} \quad (1)$$

分母の$C(x)$及び$C(y)$はどの文中にも現れるような頻出単語の影響を軽減する効果がある．

一方で，PMI は負例サンプリングに基づく skip-gram と深く関係しており，以下のように word2vec で得られた単語x, yのベクトル表現$\boldsymbol{v}_x, \boldsymbol{v}_y$の内積で近似できることが知られている[22]．本稿でも単語間の共起度指標として内積値を用いている．最終的に箇所及び現象の同義表現候補の中から内積値が高いペアを同義表現として獲得する．

$$PMI(x,y) \approx \langle \boldsymbol{v}_x, \boldsymbol{v}_y \rangle \quad (2)$$

4. 実験

3 節で説明した手法について，不具合事象に含まれる箇所及び現象の共起度を利用することが同義表現を獲得する上で有用であることを検証する．比較として人手により選定した同義表現，分布仮説（word2vec）による同義表現の獲得（3 節(1)及び(2)に相当）も実施する．

4.1 データ

弊社製品の保守情報を対象ドメインとした．保守情報は保守案件毎に「コールセンターログ」及びサービスパーソンによる「現象内容」「原因内容」「依頼内容」がフリーテキストとして記載されているものである．その例を図 2 に示す．

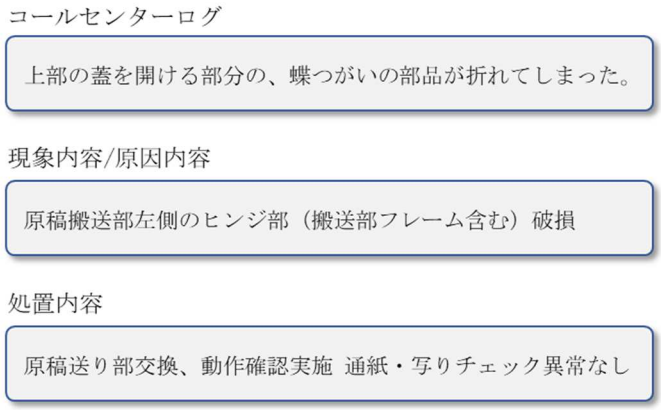

図 2 保守案件に含まれる保守情報内容例

4.2 人手による同義表現の選定

対象とする不具合事象に対して，実際のデータ上で使用されている同義表現を，比較用参考データとして人手で獲得する．

仮に不具合事象に含まれる箇所名及び現象名を検索クエリとして得られた保守案件群から，人手で各保守案件をチェックし同義表現を獲得しようとすると膨大な時間的コストが掛かってしまう．また，同一内容の案件を重複してチェックするケースも多く非効率的である．

このような課題に鑑み，3.2 節で紹介した箇所名及び現象名の抽出ルールを用いて同義表現候補を単語リストとして自動抽出し，最終的なフィルタリング（同義表現か否かの判断）のみ人手で行うことにより高精度で同義表現を獲得しようとするものである．図 3 にそのプロセスを紹介する．

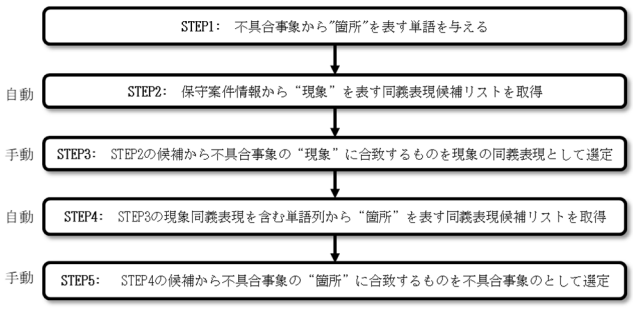

図 3 人手による同義表現獲得手順

4.3 実験結果

対象とする不具合事象を“ヒンジ破損”とした．尚，この事象は実在するものである．与えられた不具合事象に対し，箇所を“ヒンジ”，現象を“破損”としてこれらの同義表現を獲得することになる．人手による同義表現の選定結果を表 1 に示す．

表 1　人手選定により得られた同義表現

箇所：ヒンジ	現象：破損
根元	折れる
蝶番	壊れる
付け根	割れる
	外れる
	ぐらつく

次に，分布仮説（word2vec）によって得られた箇所及び現象の同義表現を，単語“ヒンジ”及び“破損”のコサイン類似度降順（Top20）で表 2 に示す．表 2 において，色付けされたものは（人が見て）誤りであることが明白な単語である．

表 2 分布仮説により得られた同義表現(類似度降順)

箇所：ヒンジ	現象：破損
支柱	割れ
支点	割れる
ワッシャー	われ
ヒンジステー	損傷
保持力	差損
カバーヒンジ	折損
フィーダーヒンジ	変形
フック	欠損
ベースカバー	はずれ
カバーセンサーフィラー	壊れる
付け根	ぐらつく
ワッシャ	ひび割れ
蝶番	付け根
ストッパギヤ	脱落
ヒンジピン	ワレ
ヒンジビス	破る
レバープルアウト	こわれる
支え	外れ
カバーセットピン	はずれる
ヒンジスペーサ	削れる

“ヒンジ破損”における word2vec の結果（コサイン類似度上位語）を見ると，箇所については“ヒンジ”の同義表現ではない語が多く含まれている．一方現象については“破損”が広義の単語であり，高頻度であることから文脈からその意味が比較的捉えられていると考えられる．しかし，表 1 と比較すると“ヒンジ”とセットで使用される表現は極限られることが分かる．

最後に，箇所及び現象の共起及び 3.2 節で示した箇所名及び現象名の抽出ルールから絞り込みを行った同義表現を表 3 に示す．表 3 は事前に得られた分布仮説（word2vec）による箇所及び現象それぞれのコサイン類似度 Top300 を同義表現候補として共起度計算（内積算出）を行い，ベクトル内積値 Top100 に現れる単語を最終的な同義表現として並べたものである．重複単語もある為，抽出された箇所及び破損の同義表現数は異なっている．太字部分は人手により獲得した同義表現と一致する単語であり，色付けされたものは（人が見て）誤りであることが明白な単語である．

表 3 提案手法により得られた同義表現

箇所：ヒンジ	現象：破損
支え	引っ掛ける
蝶番	支える
付け根	ひっかける
根元	引っかける
フック	突起
支柱	**割れ**
付根	**割れる**
支点	接合
つけ根	出っ張る
接合	脱落
台座	**ぐらつく**
マジックテープ	折損
	位置決め
	外れる
	はずれる
	われる
	おれる

表 3 を分布仮説（word2vec）によって得られた箇所及び現象の同義表現（表 2）と比較すると，箇所については人手による選定で得られた同義表現（表 1）を含む，“ヒンジ”の同義表現として相応しい語が多く含まれていることが分かる．現象については，箇所との共起を考慮すると“引っ掛ける”“ひっかける”“引っかける”とのベクトル内積値が比較的高く，結果的には誤りが増加しており，現象の同義表現獲得については明確な効果が見られなかった．即ち現象との共起情報を加味することで，箇所の同義表現が精度よく獲得できたとも言える．

表 4-1 及び 4-2 に“ヒンジ破損”と同一の不具合事象か否かを判定する例を示す．人が該当と判定した不具合事象

例が表 4-1，非該当と判定した不具合事象例が表 4-2 である．各不具合事象に対して，人手選定による同義表現（手法 1），分布仮説により得られた同義表現（手法 2），提案手法により得られた同義表現（手法 3）を基に，箇所及び現象相当の同義表現が含まれていれば○，それ以外は×としてマークしている．尚，分布仮説により得られる同義表現は類似度上位 Top20（表 2 に記載分）を対象にしている．

該当とされた不具合事象（表 4-1）については，提案手法により箇所の同義表現を与えることで，検索漏れを防ぐことが可能であることが分かる．一方で，分布仮説によって得られた表現では検索ノイズが発生する場合がある．例えば“ヒンジビス”と“外れ”は共に分布仮説によって得られる表現ではあるが，内積値では Top100 に入らない為，表 3 の箇所（最終的な同義表現）からは除外されている．

しかし，“カバーヒンジ”については正しい同義表現ではあるものの，共起度（内積値）が低いことで除去されてしまっている．

対処不具合事象：“ヒンジ破損”
手法 1：人手選定による同義表現（表 1）
手法 2：分布仮説により得られた同義表現（表 2）
手法 3：提案手法により得られた同義表現（表 3）

表 4-1　各手法による不具合事象判定例（該当ケース）

手法1	手法2	手法3	保守案件に含まれる不具合事象
×	○	○	「上の蓋の支えが割れた」
○	○	○	「カバーの蝶番が割れた」
○	×	○	「原稿送り部ユニット根元破損」
×	○	○	「上の抑えるフタの支点の片方が割れている」
×	○	×	「前カバーヒンジ部外れ」

表 4-2 各手法による不具合事象判定例（非該当ケース）

手法1	手法2	手法3	保守案件に含まれる不具合事象
×	×	×	「原稿フィーダヒンジ異常」
×	×	×	「手差し台ヒンジピン折れ」
×	○	×	「原稿ヒンジビス外れ」

尚，現象における“引っ掛ける”“ひっかける”“引っかける”は類似度上位 Top20（表 2）には含まれていない．Top300 までを同義表現候補とせずに，この時点でさらに絞り込む方法もあるが，一方で網羅性を上げるには分布仮説による手法で同義表現候補を得ておく必要があり，本実験においては箇所名では多く，現象名では少なく取得した方が良いことが分かる．

本稿で紹介した箇所及び現象の共起度を考慮する手法を適用することにより，高精度で同義表現が得られる可能性があるものの，word2vec 等から得られる類似度及びベクトル内積値に対する適切な閾値設定が課題であると言える．

5. おわりに

本稿では，不具合事象を構成する「箇所名+現象名」のペアを入力とし，「箇所名」及び「現象名」の同義表現をセットで獲得していく手法を紹介した．また，よく知られた分布仮説による文脈情報を基にした同義表現獲得に加え，箇所及び現象の共起度と抽出パターンで同義表現を抽出する実験を行った．

今後は，一般公開されている自動車のリコール情報や家電製品の不具合/点検/修理/自主回収データ等を対象として実験することで手法の有用性を検証することを考えている．また，応用として検索システムにおけるクエリ拡張での効果検証も実施予定である．

参考文献

[1] 今村誠, 高山康博, 三上崇志, 岡田康裕. 技術文書からの用語知識の自動獲得方式の検討. 情報処理学会研究報告情報学基礎. 2007, 34 号, p. 25-32.
[2] 高橋宏季, 井上昇, 伴凌太, 位野木万里. 要求仕様をあいまいにする同義語の特性分析と同義語辞書の自動作成手法の提案. 情報処理学会研究報告ソフトウェア工学. 2019, 1 号, p. 1-7.
[3] 海野裕也, 宮尾祐介, 辻井潤一. 自動獲得された言い換え表現を使った情報検索. 言語処理学会第 14 回年次大会. 2008, p. 123-126.
[4] Zellig Harris. Distributional Structure. The Philosophy of Linguistics. Oxford University Press. 1985, p. 26-47.
[5] 王玉馨, 清水信幸, 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. 情報処理学会研究報告自然言語処理. 2008, p. 7-14.
[6] “言い換え技術の研究動向”. http://paraphrasing.org/~fujita/publications/mine/fujita-paraphrase-201806.pdf, (参照 p. 67-83.).
[7] C. Bannard and C. Callison-Burch. Paraphrasing with bilingual parallel corpora. In Proc. of ACL ’05, 2005.
[8] Harris, Z. Distributional structure. Word, 10(23), 1954, p. 146-162.
[9] John Rupert Firth. Papers in Linguistics 1934-1951. Oxford University Press, 1957.
[10] 平原一帆, 難波英嗣, 竹澤寿幸, 奥村学. 言い換えを用いたテキスト要約の自動評価. 情報処理学会論文誌データベース. 2010, Vol.3, No.2, p. 91‐101.
[11] Y. Qiu and H. P. Frei. Concept based query expansion. In Proc. of SIGIR ’93, 1993.
[12] H. Schutze and J.O. Pedersen. A cooccurrence-based thesaurus and two applications to information retrieval. Information Processing and Management, Vol. 33, No. 3, 1997.
[13] R. Mandala, T. Tokunaga, and H. Tanaka. Combining multiple evidence from different types of thesaurus for query expansion. In Proc of SIGIR ’99, 1999.
[14] Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiro Toyama. Selection of Effective Contextual Information for Automatic Synonym Acquisition. Proc. Of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, 2006, p. 353-360.

[15] 寺田昭, 吉田稔, 中川裕志. 文脈情報による同義語辞書作成支援ツール. 情報処理学会研究報告自然言語処理. 2006, p. 87-94.
[16] 大和裕幸, 稗方和夫, 辻本翔, 松野二郎. 駐在監督報告書からの不具合知識獲得手法に関する研究. 日本船舶海洋工学会論文集. 2009.
[17] 小路悠介, 來村徳信, 加藤義清, 筒井良夫, 溝口理一郎. 相互運用性を指向した機能・不具合知識の統合とその概念写像に基づく知識変換. 人工知能学会論文誌. 2007, vol.22, No.1, p. 78-92.
[18] 大森信行, 森辰則. 不具合事例文からの製品・部品を示す語の抽出―語の実体性による分類―. 電子情報通信学会論文誌. 2012, vol.J95-D, No.3, p. 697-706.
[19] 小町守, 飯田龍, 乾健太郎, 松本裕治. 名詞句の語彙統語パターンを用いた事態性名詞の項構造解析. 自然言語処理. 2006, vol.17, no.1, p. 141-159.
[20] word2vec GitHub リポジトリ https://github.com/dav/word2vec
[21] “自然言語処理における自己相互情報量”. http://camberbridge.github.io/2016/07/08/自己相互情報量-Pointwise-Mutual-Information-PMI-について/
[22] S Arora, Y Li, Y Liang, T Ma, A Risteski. A latent variable model approach to PMI-based word embeddings. Transactions of the Association for Computational Linguistics 4, 385-399, 2016.