E-032

自動作成された類義語抽出ルールによる類義語抽出精度向上手法
Improving Synonym Extraction Accuracy Using Automatically Created Rules

我妻 正太郎† 照屋 絵理† 竹内 理†
Shotaro Agatsuma Eri Teruya Tadashi Takeuchi

1. はじめに

医用機器メーカ等の製品保守現場では、日々の保守作業の内容を記録した保守報告書が蓄積されている。保守報告書には機器の故障内容、原因、処置が記載されており、各メーカでは、これらの情報を活用した保守作業の効率化を試みている。保守報告書活用の実例として、機器の故障が発生した際、故障原因の特定や処置方法の立案を迅速に行うために、保守員が類似の故障事例が記載されている過去の保守報告書を検索する事例が挙げられる。

しかし、実際の保守報告書には、故障機器名の正式名称と略称が混在するなど表記ゆれが多く、単純なキーワード検索では求める報告書がヒットしない。このような表記ゆれに対しては類義語辞書を導入することで対処することが多い。しかし、一般的な文書向けの表記ゆれ対策用の類義語辞書には故障機器名などの保守に関する専門用語の記載が無いため、医用機器の保守報告書のような専門用語が多い文書に適用するには不十分である。このため専門用語を含む類義語辞書の作成が求められる。

既存の類義語抽出手法では、機械学習により各単語の周辺に出現する単語の出現分布 (コンテキスト)を学習し、コンテキストが類似する単語同士を類義語とする手法が一般的である [1][2][3]。しかし、医用機器メーカ等の保守報告書等の社内文書ではコンテキストを十分に学習できるほどの文書量が無く、既存手法では抽出精度が低い。例えば、既存手法 [1]では、10 億語を含むニュースデータセットを学習に用いるが、多くの場合これほどのデータ量を確保することは困難である。学習データ量が少ない場合、学習済みモデルの転移学習が一般的に有効である。しかし、公開されている学習済みモデルの多くは一般的な文書を元に学習されており、専門的表現を多く含む社内文書に適用しても効果が小さい [4]。

このように、機械学習を用いた既存手法では、学習データ量が少なく、かつ専門用語の類義語抽出が求められる場合、抽出精度が低い。これに伴い、類義語抽出結果の修正工数が大きくなる。よって、実用的な時間で類義語抽出精度を向上させることが課題となる。

本研究では、類義語抽出精度向上のために、既存の機械学習手法に加え、ユーザの一部の類義語候補に対する正誤判定結果から自動作成した類義語抽出ルールを類義語抽出に用いることで抽出精度の向上を図る手法を提案する。そして、提案手法を実際の社内文書に適用して類義語抽出を行い、その精度と結果修正工数を既存手法の結果と比較した。その結果、提案手法により、精度が実用レベルまで向上出来ることを確認した。

† (株) 日立製作所 Hitachi, Ltd.

2. 課題解決方針とその実装

2.1 解決方針

1 章で述べた通り、既存手法では、学習データ量が少なく、かつ専門用語の類義語抽出が求められる場合の類義語抽出精度向上が課題となる。そこで、本研究では、既存の機械学習手法では文書数が少ない場合、抽出のための閾値設定が困難なこと、専門用語の類義語には文字の一部が一致している (部分列一致)単語が多いことの 2 点に着目し、これらに対する類義語抽出ルールをユーザの一部の類義語正誤判定結果から自動作成する。さらに、機械学習手法と類義語抽出ルールとを組み合わせ、抽出精度向上を図る。

閾値設定に関して、既存の機械学習手法では、各単語のコンテキストを学習した結果得られた単語のベクトル表現間の距離を単語間の類似度とし、類似度がある閾値を超えれば類義語としている [3]。閾値は通常人が抽出結果を見て設定するが、学習データ量が少ない場合、コンテキストが偏り、類似度の高い単語が複数出現するため、閾値の設定が困難となる。そこで本研究では、ユーザの類義語正誤判定結果とそれぞれのペアになっている類義語 (類義語ペア)に紐づけられる機械学習が算出する類似度などの特徴量から、特徴量ごとに適切な閾値を自動作成する。

部分列一致に関して、専門用語の類義語ペアは「X ノズル」と「Y ノズル」のように文字の一部が一致していることが多い。ここから、ユーザの類義語正誤判定結果から部分列一致である箇所を自動的に抽出する。さらに、部分列一致情報を用いた抽出ルールを自動作成し、作成したルールを用いることで類義語を抽出する。

2.2 類義語抽出精度向上手法の概要

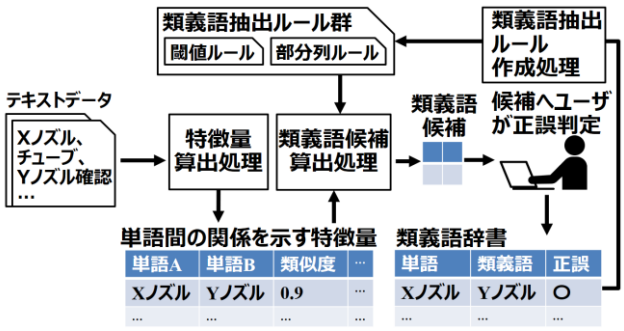

図 1　類義語抽出精度向上手法概要

提案手法の概要を図 1 に示す。まず、特徴量算出処理により、既存の機械学習手法を用いて、テキストデータ中の単語間の関係を示す特徴量 (機械学習が算出する類似度など)を算出する。その後、特徴量を元に、類義語候補算出処理にて類義語候補を出力する。また、類義語候補をユーザに提示し、ユーザは一部の候補に対して類義語かどうかの正誤判定を行い、正判定の候補を類義語辞書として保存

する。さらに、類義語抽出ルール作成処理にて、類義語辞書の単語間の関係を示す特徴量の閾値と単語間の部分列一致情報を算出し、類義語抽出ルールを作成する。作成したルールを類義語候補算出処理時に正誤判定が行われていない単語間に適用することで、単語間の中から尤もらしい類義語候補を効率的に抽出する。ルールによって算出された類義語候補へ必要に応じてユーザが再度正誤判定することで、類義語抽出ルールを更新し、さらに精度良い類義語抽出が可能になる。

2.3 類義語抽出ルールの概要

類義語抽出ルールは閾値算出ルールと部分列ルールで構成する。

(1) 閾値算出ルール

既知の類義語ペアと非類義語ペアから、類義語ペア間の特徴量と非類義語ペア間の特徴量それぞれの分布を集計する。その後、既知の類義語ペアについて、精度と再現率の積が最大となる閾値を特徴量ごとに算出する。一般に精度と再現率はトレードオフの関係にあり、精度を優先すると誤抽出が減少し、未抽出が増加する。反対に再現率を優先すると誤抽出が増加し、未抽出が減少する。出来る限り未抽出を減らしつつ、類義語抽出の精度を向上するため、上記のように閾値を設定した。なお、特徴量としては、単語の文中での出現回数、単語間の編集距離 [5]、Word2Vec [1]で算出した単語間の類似度、部分列一致単語の類義語抽出に優れる既存手法 FastText [2]で算出した単語間の類似度の 4 種類を用いた。

(2) 部分列ルール

まず、既知の類義語ペア、非類義語ペア間で部分列が一致していた場合、部分列が一致していない箇所の類似性を判定する。例えば「X ノズル」と「Y ノズル」が類義語なら、部分一致列「ノズル」を除いた「X」と「Y」に類似性があるとする。部分列ルール適用時は、例えば「X」と「Y」に類似性がある場合、「X ソレノイド」「Y ソレノイド」のような類似性ある部分列を保有するペアを類義語とする。このルールにより、部分列一致単語から効率的に類義語抽出を行う。

3. 評価と考察

本章では、提案手法と既存手法である Word2Vec [2]を用いて類義語を抽出し、類義語抽出精度と誤抽出修正工数を評価した。また、目標である少ないデータ量から高精度な類義語抽出が達成できたのか、実用上の観点から考察した。

3.1 評価

評価医用機器関係の報告書 27538 件 (約 200 万語)を対象に提案手法と既存手法で類義語抽出を行い、精度を比較した。また、専門知識を有する保守員 1 名が誤抽出修正を行い、かかった工数を比較した。提案手法については、類義語抽出に当たり、既知の類義語ペア 113 ペアと非類義語ペア 224 ペアから前章で述べた類義語抽出ルールを算出した。また、Word2Vec など機械学習に用いるハイパパラメータは 2 つの手法で同一のものを用いた。

表 1 に評価結果を示す。提案手法により抽出された類義語候補 711 ペア中、類義語であったのは 481 ペアであり、正答率は 67.6%であった。また、誤抽出修正工数は約 2 時間であった。一方、既存手法により抽出された類義語候補は全 13388 ペアであった。このペア全体の誤抽出修正を行うには工数がかかるため、今回はランダムに抜き出した 711 ペアから、既存手法の精度を推定した。その結果 711 ペア中類義語であったのは 51 ペアであり、既存手法の正答率は 9%であると推定できる。また、誤抽出修正工数は、711 ペアに約 2 時間かかることから、約 38 時間かかると推定される。以上の結果から、提案手法により、精度は約 59%上昇し、辞書修正工数は約 1/19 になると推定される。これにより、提案手法の類義語抽出精度向上と誤抽出修正工数削減の効果が認められた。

表 1　既存手法と提案手法の性能比較結果

評価尺度	類義語抽出手法	
	既存手法 [1]	提案手法
精度(正解率)	9% (推定)	67.6%
誤抽出修正工数	38 時間 (推定)	2 時間

3.2 考察

類義語抽出を実際の保守業務に適用するにあたって重要なことは、誤抽出修正も含めた実用的な時間での類義語抽出の実現である。評価結果から、提案手法により約 2 時間程度で社内文書からの類義語抽出が可能になった。実運用の形態にもよるが、ある一定の期間ごとに類義語抽出を行う場合においては十分実用的な処理時間である。この結果から、本手法により、実用上では十分な精度と作成工数で類義語抽出が可能になったと考えられる。

4. おわりに

本研究では、既知の類義語情報から自動作成した類義語抽出ルールと既存の機械学習手法を組み合わせることで、社内文書のようなデータ量が少ない、かつ専門用語を多く含む文書から、高精度かつ省工数で類義語抽出を行える手法を提案した。既存手法を比較した結果、精度は約 59%上昇、辞書修正工数が約 1/19 削減した。本結果より、提案手法が類義語抽出精度向上に有用であることが分かった。今後は医用機器関係とは別種類の社内文書にも提案手法を適用し、類義語抽出可能か調査する。また、既知の類義語情報から取得できるルールの拡充を行う。

参考文献

[1] Tomas Mikolov et.al, “Distributed Representations of Words and Phrases and their Compositionality” Proceedings of the 26th International Conference on Neural Information Processing Systems, pp.3111–3119 (2013).

[2] Piotr Bojanowski et.al., “Enriching Word Vectors with Subword Information,” Transactions of the Association for Computational Linguistics, Volume 5, pp.135-146 (2017).

[3] Derry Jatnika et.al., “Word2Vec Model Analysis for Semantic Similarities in English Words”, Procedia Computer Science, Vol.157, pp.160-167 (2019).

[4] Lili Mou et.al., “How Transferable are Neural Networks in NLP Applications?” Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp.479-489 (2016).

[5] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” (in Russian), Doklady Akademii Nauk, Vol.163, no. 4, pp. 845–848 (1965).