対照学習による文ベクトルを使用した障害レポートのクラスタリング

小林 千真[1] 山下 郁海[1] 岡 照晃[1] 小町 守[1] 真鍋 章[2] 谷本 恒野[2]
東京都立大学[1] 富士電機株式会社[2]
{kobayashi-kazuma1@ed., yamashita-ikumi@ed., teruaki-oka@, komachi@}tmu.ac.jp,
{manabe-akira, tanimoto-kouya}@fujielectric.com

概要

本研究では、機器が故障した際に作成される障害レポートを文ベクトル化しクラスタリングを行なった。SimCSE は現在最高性能の文ベクトル獲得手法の一つであるが、日本語での有効性が明確に示されていない。そこで述部意味関係データを使用して、日本語文ベクトルの評価を行なった。その結果、教師ありの SimCSE は日本語でも有効であることを確認した。また文ベクトルを使用し、障害レポートのクラスタリングを行ない、ドメインごとの結果の違いを事例に基づいて分析した。その結果、入力文やドメインの特性により文ベクトルが有効な場合とそうでない場合が確認された。

1 はじめに

自動販売機や ATM に挙げられるように、現代社会には様々な場所に様々な機器が設置され我々の暮らしを支えている。それらの機器が故障するたびに障害レポートが作成され、日々増え続けている。障害レポートを分析し、障害の全体像を捉えることで、対策や業務の効率化が期待できる。本研究で扱う障害レポートは大きく **状況**、**原因**、**措置** の 3 要素からなる文章であり、重複する内容のレポートもある。図 1（上）が障害レポートの具体例である。それらを集約することで確認すべき障害データ数は削減され、数千以上の障害レポートを有効に分析することができる。

山下ら [1] は障害レポートを集約するにあたり、障害レポートの各文が状況・原因・措置・その他のどれに該当するかの分類タスクを行なった。本間ら [2] は障害レポートの各文から状況・原因・措置に関する重要箇所を抽出する重要箇所抽出タスクを行なった。本研究はこれらの下流タスクとして

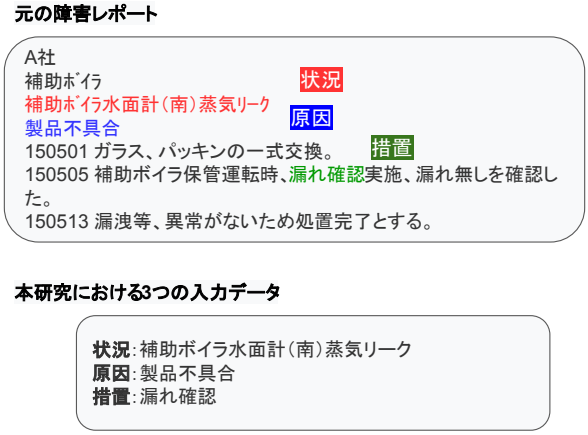

図 1 障害レポートと本研究で想定する入力の具体例。[1)]

位置付けられ、抽出された状況状況・原因・措置のフレーズを入力とし、それらを基に段階的な集約に取り組む。本タスクでの入力の具体例を図 1（下）に示す。

本研究では似た意味の文を集約するにあたって、文の意味関係を反映した文ベクトルを作る技術に注目した。大規模事前学習モデル (PLM) によって Semantic Textual Similarity (STS) の性能が大幅に向上している。STS で特に高い性能を出しているのは PLM に 2 文を入力し、実数値で類似度を出力するように fine-tunig する手法であり、現在最高性能の SMART [3] もこの類の手法である。この手法では、類似度計算のたびに PLM による計算を実行する必要があり、繰り返し類似度計算を要するクラスタリングには計算時間の観点で向かない。対して、PLM から文ベクトルを獲得する手法の場合、文ベクトルを一度獲得すれば、以降ベクトル類似度の計算のみで済むため、計算時間の問題が大幅に改善される [4]。代表的な PLM のひとつに BERT [5] があ

1) 本研究では人手で作成されたデータを使用するが、実際に適用する際には本間ら [2] により、障害レポートの各文から状況・原因・措置が抽出された状態が本研究の入力データとなる。

る。BERT は文書分類のような応用タスクのために fine-tuning をすることを前提としたモデルである。BERT から文ベクトルを得ることもできるが、 文ベクトルは性能が低いことが知られている [4]。BERT から高性能の文ベクトルを獲得するための手法がいくつか提案されているが、 SimCSE [6] が最も性能の高い文ベクトルを獲得できる手法である。しかし SimCSE は英語でしかその性能を保障されていない。

本研究では、障害レポート（日本語）から抽出した重要箇所を SimCSE で追加学習した BERT に入力し、文ベクトルを獲得する。その文ベクトルをクラスタリングすることで集約を行なった。本研究の貢献を以下に示す。

- 障害レポートから抽出した重要箇所に対して文ベクトルを用いたクラスタリングを試みた。
- 日本語において SimCSE による文ベクトルを使用した。また、その評価実験を行なった。

2 障害レポートの集約

前述の通り、想定する入力は状況・原因・措置の 3 つの要素に分けられており、図 1 （下）が入力となることを想定している。本研究では 3 段階のクラスタリングによる集約を行う（付録 A で具体例を使用し、その手順を図示した）。各レポートは $r_i = (s, c, a)$ の形で入力される。ただし、s は状況、c は原因、a は措置である。1 つ目のステップでは全レポートを親クラスタとして、s の情報のみでクラスタリングを行い、クラスタ集合 $\{C_1^s, C_2^s, ..\}$ を得る。2 つ目のステップでは C_x^s をそれぞれ親クラスタとして、c の情報のみでクラスタリングを行う。その結果 C_x^s ごとにクラスタ集合 $\{C_1^c, C_2^c, ..\}$ を得る。3 つ目のステップでは、C_x^c をそれぞれ親クラスタとして、a の情報のみでクラスタリングを行う。その結果 C_x^c ごとにクラスタ集合 $\{C_1^a, C_2^a, ...\}$ を得る。[2)]

集約を行う際、状況・原因・措置の各段階でそれぞれの文字列でクラスタリングを行う。文字列にはフレーズのように短いものもあるが、本研究では文とみなし、BERT を SimCSE で訓練したモデルを使用して文ベクトルを作成する。集約タスクの各段階では文ベクトル同士のユークリッド距離を使用して DBSCAN [7] でクラスタリングする。

2） このような流れから、原因・措置では、同一の文字列であっても親クラスタが異なる場合は、次段のクラスタリングは別々に行われる。

3 SimCSE

SimCSE は PLM に対照学習を行うことで、STS においてより性能の高い文ベクトルを獲得する手法であり、BERT や RoBERTa に適用することで、STS で最高性能のスコアを記録する文ベクトルを獲得できる。STS では 、2 つの文とそれに対する意味的な類似度を人手でスコアリングした値があり、その値との相関によりモデルの性能が評価される。

SimCSE には Natural Language Inference (NLI) データを使用した教師ありの手法と生の文 を使用した教師なしの手法がある。教師ありの手法では NLI データセット[3)]から、ある文に対して含意の関係にある文と矛盾の関係にある文を抽出し、元の文と含意の関係にある文の文ベクトルが近づくように、矛盾の関係にある文の文ベクトルは遠くなるように対照学習を行う。教師なしの手法では生の文集合（例えば、 Wikipedia） のある文に対して、同じ文の異なるドロップアウトの文ベクトルを生成し、その文ベクトルと元の文ベクトルが近づくように、ランダムに選んだ別の文の文ベクトルが遠ざかるように対照学習を行う。

Gao ら [6] は教師ありの手法が教師なしの手法よりも英語 STS において良い結果となることを示した。しかし日本語で公開されている NLI データセットは SNLI （英語の NLI データセットのひとつ）を機械翻訳し、人手でフィルタリングしたものであり、流暢な翻訳とはなっていない。このため、教師ありの手法は英語ほどの性能が得られるとは限らない。それに対して、 流暢な日本語で書かれているテキスト（例えば、Wikipedia）は利用可能であるから、教師なしの手法では英語と同等の性能が期待される。このことから英語同様に教師ありの手法がより良い結果になることは自明ではないので、本研究では教師ありの手法と教師なしの手法の両方を使用し比較した。

4 日本語 SimCSE

4.1 実験設定

本研究では、SimCSE によるモデルの学習は東北大が公開している日本語 BERT [4)] に SimCSE を適用することで訓練した。また教師ありと教師なしの手

3） 文に対して「矛盾」「中立」「含意」の 3 つの関係である文が用意されているデータセット。

4） https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/

表 1 原文と述部意味関係コーパスの文ベクトルのコサイン類似度。

	同義（↑）	無関係（↓）
SimCSE （教師あり）	**0.9081**	0.7648
SimCSE（教師なし）	0.8361	**0.6962**
BERT	0.8653	0.7775

法でそれぞれ訓練を行なった。教師ありの学習には京都大学が公開している日本語 NLI データセット[5]を使用した。教師なしの学習には日本語 Wikipedia 100 万文を使用した。SimCSE の教師ありでは 10 epoch、教師なしでは 5 epoch 学習を行なった。

4.2 文ベクトルの評価

[6] では獲得した文ベクトルを英語 STS データセットで評価した。しかし日本語で公開されている STS データセットは存在しないため、京都大学が公開している述部意味関係コーパス[6]を使用し評価を行なった。このコーパスには「同義」、「無関係」、「反義」、「含意」の 4 種類の文対が用意されている。「反義」と「含意」は文ベクトルによる評価は自明ではないため、本研究では「同義」と「無関係」のみを使用した。評価には文ベクトルのコサイン類似度を使用し、「同義」ではより高く、「無関係」ではより低くなることを基準とした。

4.3 実験結果

結果を表 1 に示す。SimCSE の教師ありでは、BERT と比べ「同義」で類似度が 4 ポイント上がり、「無関係」では 1 ポイント下がった。つまり、どちらでも BERT よりも改善している。一方、SimCSE の教師なしでは類似度が「同義」で 3 ポイント、「無関係」で 8 ポイント、それぞれ下がってしまった。この結果から、流暢でない日本語 NLI データを使用した教師ありの SimCSE で性能向上することが確認できた。教師なしと教師ありの優劣はこの実験結果から述べることができない。

5 障害レポートのクラスタリング

5.1 実験設定

ベースラインとして入力文字列同士の編集距離を距離とみなす手法を採用する。また SimCSE を使用しない東北大の日本語 BERT に直接 1 文を入力し、文ベクトルを獲得する手法もベースラインとする。これらの比較により、そもそも文ベクトルを使用することで性能が向上するか否か、SimCSE で学習した文ベクトルを使用することが障害レポート文のクラスタリングに効果があるか否かを検証できる。

5） https://nlp.ist.i.kyoto-u.ac.jp/
6） https://nlp.ist.i.kyoto-u.ac.jp/?PredicateEvalSet

表 2 各ドメインのデータ数と段階ごとの正解クラスタ数。

	データ数	状況	原因	措置
冷凍	4,556	822	3,170	3,777
自販機	1,130	578	949	986
火力	960	793	852	933

DBSCAN のハイパーパラメータについて、ϵ（近傍とみなす最大距離）は [0.1, 1, 2, 3, 4, 5, 6, 10] の間でハイパーパラメータの探索を行なった。min-sample （密集領域を形成する最低サンプル数）は正解データの最低クラスタ数が 1、2 のものが散見されたため 2 とした。外れ値として扱われたデータ点は要素数 1 のクラスタとみなした。SimCSE による文ベクトルの獲得には 4 節で最も良い結果を得たモデルを使用した。

クラスタリングの評価には富士電機（株）の保有する食品流通分野（冷凍）、食品流通分野（自販機）、火力発電分野の障害レポートに、機器障害の実務に携わった経験のある社員がアノテーションを行い、状況、原因、措置に分類した正解データ（NFKC 正規化済み）を作成し、使用した。以後、食品流通分野（冷凍）を「冷凍」、食品流通分野（自販機）を「自販機」、火力発電分野を「火力」とする。使用する障害レポートのデータ数、正解クラスタ数を表 2 に示す。全体としてかなり細かいクラスタを形成することを期待しているデータセットである。

クラスタリングの評価指標の 1 つである Adjusted Rand Index (ARI) [8] を使用し、状況・原因・措置をそれぞれ評価した。ARI は正解クラスタと同一の場合 1 となり、一致率が下がると値も小さくなる。本研究では、前の段階（原因の時は状況、措置の時は原因）の正解データのクラスタをもとに各段階のクラスタリングを行い、評価した。

5.2 実験結果

結果を表 3 に示す。ドメインごとにそれぞれ異なる結果となった。自販機ドメインでは、状況において SimCSE（教師あり）が圧倒的に良い結果となっ

表 3　クラスタリング結果に対する ARI スコア。

	冷凍			自販機			火力		
	状況	原因	措置	状況	原因	措置	状況	原因	措置
SimCSE （教師あり）	0.8757	**0.9275**	**0.9448**	**0.8366**	0.9098	0.9583	0.6592	0.9842	**0.8614**
SimCSE（教師なし）	0.8757	0.9236	0.9363	0.6550	0.8957	0.9583	0.6692	0.9842	0.8298
BERT	0.8757	0.9229	0.9363	0.6085	0.8944	0.9479	**0.6767**	0.9842	0.8299
編集距離	0.8757	0.9270	0.9363	0.6312	**0.9107**	**0.9850**	0.6297	0.9842	0.8555

た。しかし原因と措置ではベースラインである編集距離が最も良い結果となった。冷凍ドメインでは、状況では全ての手法で全く同じ結果となった。原因と措置においては SimCSE（教師あり）がわずかに良い結果となった。火力ドメインでは、状況において BERT が最も良い結果となった。原因では全て同じ結果となった。措置では SimCSE （教師あり）が最も良い結果となった。次節では具体的な事例を確認しながら、このようにドメインごとに異なる結果になった原因を探る。

6　事例分析

本節では、状況のフレーズに注目して分析を行なった。また教師ありの SimCSE を単に SimCSE として記述した。

ドメインごとの結果の違い　自販機ドメインは「50 円玉が 10 円のメックに入る」のように文に近い入力が多く、火力・冷凍ドメインでは、「給水シリカ計指示上昇」のようなフレーズ的な入力が多かった。入力が文に近いほど SimCSE による文ベクトルの性能向上の恩恵を得たと考える。逆にフレーズ的な入力が多かった火力や冷凍ドメインでは高性能の文ベクトルが効かなかったと考える。冷凍ドメインでは状況も高いスコアとなっているが、完全一致した文字列だけを同一クラスタとみなした場合と同じスコアだった。このことから冷凍ドメインはもともと統一の取れた表記がなされており、文ベクトルを使うことで完全一致でない文字列を余計に同一クラスタとしてしまったと考える。火力ドメインでは「GT」 のような独自の省略記号を区別できずに失敗する事例（後に紹介）が散見された。

モデル間の結果の違い　BERT と SimCSE でクラスタリング結果が異なった事例を比較し、フレーズ的な入力では、わずかな違いで異なる基準でクラスタを成していることがわかった。しかしその基準は明確ではなく、定性的な判断でもその評価は難しい。具体的な事例を付録 B に示した。文に近い入力では、「 購入不可」と「 購入できない」のような違いを BERT では別のクラスタとしたが、SimCSE では同じクラスタとして分類できた。

モデルの結果と正解データの違い　「GT」のような独自の短縮単語が付くか否かといった違いを正解データでは区別しているが、 SimCSE では区別しないという事例が散見された。これらを別クラスタとみなすか否かは、この障害レポート独自の基準であり、実務に携わっていない場合、人間でも判断が分かれるところであると考える。SimCSE はあくまでも 一般的な文ベクトルの性能向上を考えた手法であり、本研究で対象としたような特定の意図に基づくクラスタリングには別のアプローチが必要である。また反義的なフレーズを、SimCSE では同じクラスタと見なす事例があった。反義的なフレーズ同士にどのような類似度を設定すれば適切かは自明ではないため、このようなケースが文ベクトルの類似度を使った手法の難しさであると考える。具体的な事例を付録 B に示した。

7　おわりに

本研究では、日本語において SimCSE を使用して文ベクトルを学習し、その評価を行なった。その結果、教師ありの手法では性能の向上を確認した。次に障害レポートから抽出されたフレーズに対して文ベクトルを使用したクラスタリングを試みた。3 つのドメインを使用したが、ドメインごとに異なる結果となった。分析として、文ベクトルの類似度を使う手法では反義的なフレーズの区別が難しいという点が分かった。今後、さらなる性能向上のためにはタスク特化のデータ（分けたい事例、分けたくない事例）を使って SimCSE を学習する方法や、文ベクトル以外の部分に注目した方法で性能向上を目指したい。

参考文献

[1] 山下郁海, 小町守, 真鍋章, 谷本恒野. 隠れ層補間によるデータ拡張を用いた障害レポート分類. 言語処理学会第 27 回年次大会 発表論文集, pp. 1794–1798, 2021.

[2] 本間広樹, 小町守, 真鍋章, 谷本恒野. BERT モデルを用いた障害レポートに対する重要箇所抽出. 言語処理学会第 27 回年次大会 発表論文集, pp. 189–193, 2021.

[3] Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2177–2190, Online, July 2020. Association for Computational Linguistics.

[4] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[6] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[7] M Ester, H P Kriegel, J Sander, and Xu Xiaowei. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In **Proceedings of the Second International Conference on Knowledge Discovery and Data Mining**, pp. 226–231, Portland, 1996. AAAI Press.

[8] Lawrence Hubert and Phipps Arabie. Comparing partitions. **Journal of Classification**, Vol. 2, No. 1, pp. 193–218, 1985.

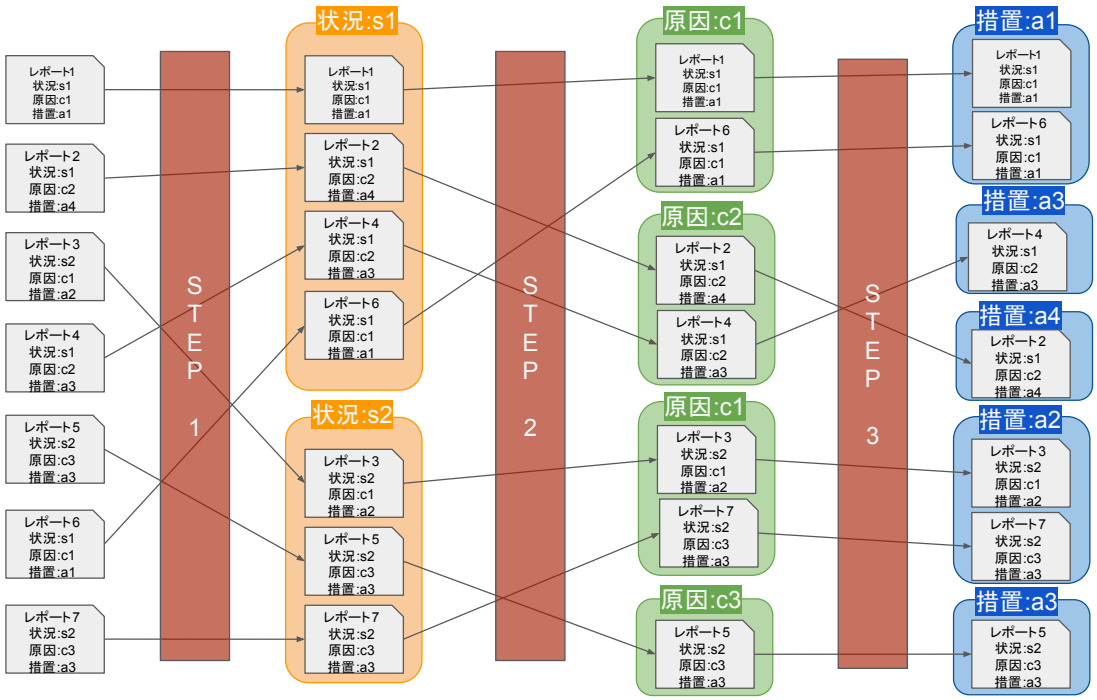

図 2　本研究における集約の概略図である。まず、状況の情報を用いてクラスタリングし、似た状況ごとに障害レポートをまとめる。次に原因の情報を用いてクラスタリングし、障害レポートを原因の情報を用いてクラスタリングし、原因ごとのクラスタを作る。最後に措置の情報を用いてクラスタリングし、措置ごとのクラスタを作る。

A　集約の具体例

集約は、図 2 に示すように、Step1〜3 の 3 段階のクラスタリングを行うタスクである。

B　失敗事例の具体例

障害レポートの ID（実際の ID とは異なる）と原文通りのフレーズを「ID: フレーズ」という形式で示し、それらに対するクラスタリングを集合の形式で示した。

モデル間の違い

- 1: 排熱回収ボイラ　布製伸縮継手より漏水
- 2: 排熱回収ボイラ布製伸縮継手より漏水
- 3: 排熱回収ボイラ布製伸縮継手 (出口) より漏水

上記の文字列に対し、BERT では{{1,2,3}}、 SimCSE では{{1,2}, {3}}とクラスタリングした。

- 4: 失火・燃焼異常
- 5: 失火・燃焼異常発生

また上記の文字列に対し、BERT では{{4}, {5}}、 SimCSE では{{4,5}}とクラスタリングした。

モデルの結果と正解データの違い

- 6, 7: GT 排ガス温度センサー異常
- 8: 排ガス温度センサー異常

上記の文字列に対して、SimCSE では{{6,7,8}}とクラスタリングし、正解データ は{{6,7},{8}}であった。

- 9: 高圧給水流量指示不良
- 10: 低圧給水流量指示不良

上記の文字列に対して、SimCSE では{{9, 10}}とクラスタリングし、正解データは {{9}, {10}} であった。