BERT モデルを用いた障害レポートに対する重要箇所抽出

本間 広樹 [1]　小町 守 [1]　真鍋 章 [2]　谷本 恒野 [2]

[1] 東京都立大学　[2] 富士電機株式会社
homma-hiroki@ed.tmu.ac.jp, komachi@tmu.ac.jp
{manabe-akira, tanimoto-kouya}@fujielectric.com

1 はじめに

世の中はあらゆる機器にあふれている．これらの機器が不具合を起こすたびに，保全担当員によって障害レポートが作成される．日々作成されている障害レポートから自動で重要な部分を抽出することは，保全担当員による異常箇所の特定と対処や，設備運用者による適切な設備運転対応，コールセンタオペレータによる迅速・適切な顧客対応に役立つと考えられる．障害レポートは主に，**状況**，**要因**及び**措置**の 3 種類の内容から構成される文章である．レポート内の各文は，前述の 3 種類と**その他**の内のどれか，あるいは複数に分類できる．本研究では分類済みの 3 種類の文章である，**状況**，**要因**及び**措置**から，それぞれ重要部分を抽出することを目的とする．

近年ニューラルネットワークを用いた大規模なモデルの研究が盛んに行われ，様々な自然言語処理タスクで高い精度を出している．しかし，本研究で扱うデータは数百件程度と規模が非常に小さく，大規模なモデルを一から学習するのは困難である．そこで，本研究では，小規模なデータに対する有効性が確認されている事前学習モデルに着目し，障害レポートを対象とする重要部分抽出タスクに対する有効性を検証する．

さらに，本研究では質問応答形式によるマルチタスク学習手法を提案する．抽出元文章には**状況**，**要因**及び**措置**の 3 種類のソースがあるが，提案手法では，既存の質問応答形式の fine-tuning 手法を流用し，抽出元文章，ソース及び重要箇所をそれぞれ，コンテキスト，質問及び解答としてモデルに入力し，全ソースを併せて学習を行う．これによってモデルがソースの違いを認識し，より高い精度の抽出ができるようになることが期待できる点や，既存の fine-tuning 手法を用いるため，低コストで実装できる点が利点として挙げられる．

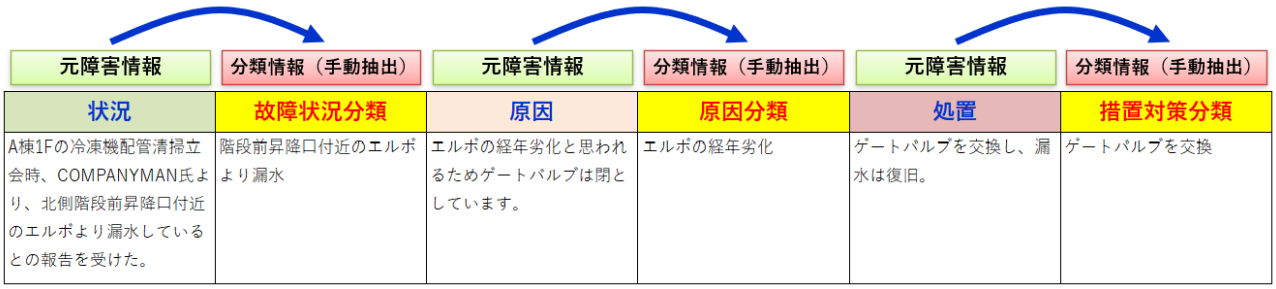
また，複数ソースが存在する場合の学習方法について考えると，各ソースで抽出モデルを学習するか，全ソースを併せた抽出モデルを学習するかの 2 通りが存在する．しかし，それぞれの方法には次のような問題が存在する．各ソースで抽出モデルを作成する場合には，モデルが増える問題や，別ソースに含まれる有用な情報を学習できなくなる問題が考えられる．また，全ソースを併せて抽出モデルを学習する場合には，あるソースの抽出精度が，別ソースに含まれる情報によって低下する可能性が考えられる．ここで，「事前学習モデルを用いて抽出する際に，複数ソースデータを同一モデルに学習するマルチタスク学習がどの程度有用か」という疑問が生じる．そこで，本研究では，「異なるソースのデータでマルチタスク学習を行う際の有効性は，ソース間類似度に依存する」という仮説を立て，マルチタスク学習の有効性とソース間類似度の関係性を検証する．

2 関連研究

2.1 事前学習モデル

近年，事前学習モデルが様々な自然言語処理タスクで最先端の性能を達成している．これは，大規模なニューラル言語表現モデルを事前学習し，下流タスクごとにそのモデルに出力層を追加して fine-tuning することで，幅広いタスクを高い精度で解くことができるというものである．

特に，2019 年に Devlin らによって考案された Bidirectional Encoder Representations from Transformers (BERT) [1] というアーキテクチャが当時様々なタスクで最先端の性能を達成し，広く使われている．このモデルは Transformer [2] を双方向に接続することで構成されている．Transformer [2] は従来の一般的



状況	故障状況分類	原因	原因分類	処置	措置対策分類
A棟1Fの冷凍機配管清掃立会時、COMPANYMAN氏より、北側階段前昇降口付近のエルボより漏水しているとの報告を受けた。	階段前昇降口付近のエルボより漏水	エルボの経年劣化と思われるためゲートバルブは閉としています。	エルボの経年劣化	ゲートバルブを交換し、漏水は復旧。	ゲートバルブを交換

図 1　障害レポートに対するアノテーション例

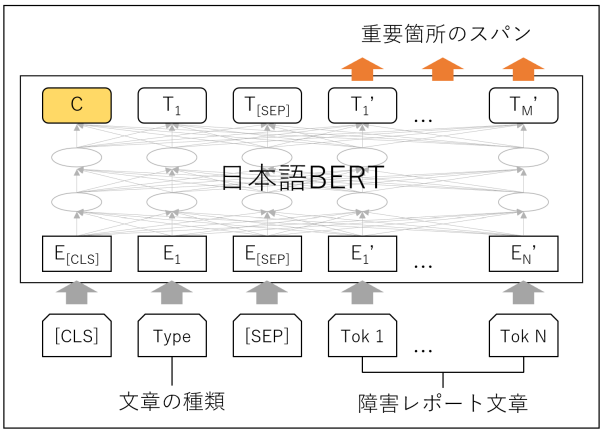


図 2　質問応答形式による fine-tuning のモデル

なニューラルモデルで使われる畳み込みニューラルネットワークや再帰ニューラルネットワークを用いず，注意機構のみで構成されたモデルである．

Devlin らの研究 [1] は英語を用いて行われているが，日本語版の事前学習済み BERT モデルもいくつかの機関から公開されている．本研究では事前学習モデルを提供する Hugging Face の `Transformers` [1)] を通して簡易に利用可能である，日本語の Wikipedia のデータを用いて学習された，東北大学が公開しているモデル[2)] を用いる．

2.2　重要箇所の抽出タスク

本研究と類似した形式の，障害レポートからの重要箇所抽出タスクを行っている研究に，Encoder-Decoder モデルを用いて系列ラベリング問題として解いている先行研究 [3] が存在する．この先行研究では事前学習モデルを用いず，Long short-term memory (LSTM) を用いている．このモデルをベースラインとして事前学習モデルの有効性を評価する．

また，この重要箇所抽出タスクは，出力の形式が文や文章ではなくその一部という違いはあるものの，抽出型要約タスクと類似している．抽出型要約はドキュメント内の重要な文を識別することによって行われており，最近ではニューラルモデルを用いて文分類の問題として解かれている．SummaRuNNer [4] はニューラルモデルを用いた初期の抽出型要約システムであり，エンコーダには再帰ニューラルネットワークを用いている．また，BanditSum [5] はドキュメントを context，要約文選択を action とした contextual bandit 問題として解いている．そして，SWAP-NET [6] はドキュメント内の際立った文とキーワードの両方を異なるエンコーダで探し，文とキーワードを各デコードステップで選択するスイッチ機構によりそれらを組み合わせて抽出要約を形成するシステムである．さらに，事前学習モデルを用いた要約タスクの研究も行われており，その有効性が確認されている [7]．これらの研究は英語を対象としているが，本研究ではより小規模な，日本語の障害情報レポートに対する事前学習モデルの有効性を確認する．

3　障害情報レポートの重要箇所抽出

3.1　タスク設定

本研究で用いるデータは，1 件の障害レポートから人手で抽出された，**状況**，**原因**及び**措置**の 3 文章に対してそれぞれ重要箇所を 1 箇所以上手動抽出したものである．重要箇所のアノテーション例を図 1 に示す．本研究は，**状況**，**原因**及び**措置**の 3 種類の文章からそれぞれ重要箇所を自動で抽出することを目的とする．

3.2　手法

ベースライン: LSTM を用いた文圧縮　ベースラインとして小平らによる関連文章を考慮した LSTM による文圧縮手法 [3] を用いる．この手法は，Filippova らの手法 [8] を元にしており，LSTM を用

1）https://github.com/huggingface/transformers
2）https://github.com/cl-tohoku/bert-japanese

いて文圧縮を系列ラベリングタスクとして解いている．さらに，3 種類の文章を同時に別々のエンコーダに入力し，それらの出力を連結し，次元数を揃えるための線形変換を行い，種類ごとのデコーダに入力する操作を行っている．この操作により複数の文章全体を考慮した重要箇所の抜き出しに取り組んでいる．

提案手法: 日本語 BERT を用いた文圧縮　本研究では日本語データで事前学習済みの言語表現モデルである日本語 BERT に対して質問応答形式の fine-tuning を行うことで文章抽出モデルを作成する．この時，抽出を行う文章の種類をモデルに理解させるために，**状況**，**原因**及び**措置**のラベルを質問として入力する．質問応答形式の fine-tuning のモデルを図 2 に示す．文章の種類のラベルを入力することで，複数種類によるマルチタスク学習の性能の向上を図る．実装には `Transformers` に含まれるスクリプト，`run_qa.py` を用いる．

4　実験

4.1　データ

本研究では，富士電機（株）が保有する設備保全の障害レポート 194 件に対して，3.1 節で説明したアノテーションを付与したものを使用する．また，元データに含まれる，人名，企業名及び場所はそれぞれ `MAN`，`COMPANY` 及び `PLACE` にマスキングを施している．

4.2　実験設定

比較手法として以下 2 つの手法を用いる．

- 小平ら [3] の手法 (MultiLSTM)
- 質問応答形式で日本語 BERT を fine-tuning する提案手法 (BERT-QA)

各手法のハイパーパラメータの設定は以下の通りである．

MultiLSTM　ドロップアウト率を 0.4，学習率を 0.001，勾配クリッピングを 6 に設定する．その他のハイパーパラメータの設定は先行研究 [3] に従う．

BERT-QA　事前学習済みモデルとして，東北大学が公開している日本語 BERT の内，whole word masking を適用して学習させているモデル[3)]を用いる．学習率を 0.005，バッチサイズを 12，エポック数を 10，保存ステップを 10 に設定し，最良のモデルを開発セットによって選択する．各文章の種類のラベルとして，**状況**，**原因**及び**措置**の 1 単語をそのまま用いる．その他のハイパーパラメータの設定は先行研究 [1] に従う．

また，どちらの手法においてもデータを 5 等分に分割し，そのうち 3 つを併せたものを学習セット，1 つを開発セット，残りの 1 つを評価セットとして用いる形で，5 分割交差検証を行う．さらに，このタスクはユーザに重要箇所をなるべく多く提示することが重要なため，評価手法には適合率に 2 倍の重みを付ける評価尺度である F_2 スコアを用いる．

3）https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

表 1　各モデルの適合率，再現率及び F_2 スコア

	適合率	再現率	F_2 スコア
MultiLSTM	0.331	0.270	0.280
BERT-QA	**0.460**	**0.570**	**0.544**

4.3　先行研究との性能比較

表 1 に各モデルの適合率，再現率及び F_2 スコアを示す．提案手法である BERT-QA が先行研究の結果を大幅に改善していることが見て取れる．このことより，数百件程度の極めて小規模なデータを用いた重要箇所抽出タスクにおいても事前学習モデルが有用であることが分かる．

4.4　ソース間類似度とマルチタスク学習

マルチタスク学習の有効性を調べるために，まず，ソース間類似度を計測する．計測には本実験と同様の日本語 BERT を用い，文の先頭に挿入する `[CLS]` トークンの最終隠れ状態（図 2 の入力から `Type` 及び `[SEP]` トークンを除いた場合の `C`）を文ベクトルとして利用する．ここで，文章から重要箇所を抽出するタスクにおけるソース間類似度を捉えるために，重要箇所の文ベクトルから元の文章の文ベクトルを引いたベクトルを事例ごとの抽出ベクトルとして定義する．

表 2 に全事例の抽出ベクトルのソースごとの平均をコサイン類似度により比較した結果を示す．**状況–原因**及び**原因–措置**の対は比較的類似度が高く，**状況–措置**の対の類似度が低くなっていることが確認できる．

次に，学習データの種類を制限する実験を行う．3 種類のソースが存在し，その内 1 種類あるいは 2 種類のソースのデータを利用してモデルを学習し，

表 2 ソース間のコサイン類似度

ソース対	類似度
状況–原因	0.752
状況–措置	0.565
原因–措置	0.789

表 3 各モデルのソース毎の F_2 スコア

	すべて	状況–原因	状況–措置	原因–措置	状況	原因	措置
状況	0.622	**0.660**	0.622	–	0.647	–	–
原因	0.445	0.492	–	0.460	–	**0.534**	–
措置	0.561	–	0.572	0.522	–	–	**0.581**

表 4 各モデルの出力結果の比較

	種類	抽出元文章（重要箇所の参照）	MultiLSTM	BERT-QA
例 1	状況	●●階段前▲▲▲より漏水	前▲▲▲より漏水	階段前▲▲▲より漏水
例 2	原因	圧縮機単体で絶縁測定実施し 0M Ωだった為、圧縮機不良と思われます。圧縮機本体の交換を要します。	を要します。	圧縮機単体で絶縁測定実施し 0M Ωだった為、圧縮機不良
例 3	措置	■■内部のドレン配管が詰まっており、そこから溢れた水が漏水していた為、ウェットバキュームにて清掃実施。通水良好の為済とします。	–	ウェットバキュームにて清掃実施

合計 6 種類のモデルを新たに作成する．

表 3 にソースの種類を絞って学習した場合の結果を示す．各列が利用した学習データ，各行が評価データのソースを示している．学習データが「すべて」であるのは先行研究との比較実験で使用したモデルと同一のものである．基本的には少ないソースで個別にモデルを学習したほうが抽出精度が高いことが見て取れる．

ここで表 2 に示すソース間類似度に着目し，単一ソースで学習した場合に対して 2 ソースでマルチタスク学習した場合を比較する．類似度が高かった**原因**と**措置**及び**状況**と**原因**はそれぞれ-0.074 と-0.059 及び+0.013 と-0.042 ポイントの差があり，その平均は-0.041 ポイントである．それに対し，類似度が低かった**状況**と**措置**は-0.025 と-0.009 ポイントの差であり，その平均は-0.017 ポイントであり，ソース間類似度が高い場合よりも良い結果となっている．つまり，質問応答形式を用いた抽出タスクを行う場合，BERT モデルを用いて算出したソース間類似度が高いほどマルチタスク学習の効果が高くなるわけではないことが言える．

4.5 ケーススタディ

表 4 に各モデルの出力結果を示す[4]．抽出元文章内の赤いゴシック体の箇所は重要箇所の参照である．例 1 は抽出にある程度成功している例である．どちらのモデルに対しても抽出元文章が短い場合は精度が高くなる傾向が見受けられた．例 2 の各モデルの出力を比較すると，MultiLSTM の出力が非文になっているのに対し，BERT-QA の出力は参照の範囲を大きく逸脱しているものの，意味が通る出力になっていることが分かる．この BERT-QA の出力が MultiLSTM に比べて流暢であるという傾向は全体を通して見られた．また，例 3 において，MultiLSTM は重要箇所が存在しないと判断している．これは学習データに含まれる類似する事例が少ないことが原因の 1 つとして考えられ，MultiLSTM では全体を通して散見された．それに対して BERT-QA は正しく出力できており，MultiLSTM に比べてデータセットが小規模な場合の抽出にも有効であると考えられる．

また，BERT-QA における出力を分析すると，抽出元の長い文章全てを重要箇所として判定してしまっている誤りがいくつか存在した．これに関しては，重要箇所の最大系列長をヒューリスティックに決定し，上位の出力候補から条件を満たすものを選択することでさらなる精度の向上を図ることが可能と考えられる．

5 おわりに

本研究では BERT モデルに対する質問応答形式の fine-tuning を利用した重要箇所抽出抽出手法を提案した．この手法では既存の実装を用いて容易に抽出型要約を行うことができ，LSTM を用いた既存手法に比べて F_2 スコアで 0.26 ポイントと大幅に精度が向上することを示した．さらに，ソース間類似度に着目したマルチタスク学習の有効性を分析し，データが少ない場合，ソース間類似度が高いほどマルチタスク学習の効果が高くなるとは一概に言えないことを発見した．

4） データプライバシーのため一部表現及び固有名詞を●，▲及び■でマスキングしている．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL 2019*, pp. 4171–4186, 2019.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS 2017*, pp. 5998–6008. 2017.

[3] 小平知範, 宮崎亮輔, 小町守. 障害情報レポートに対する同時関連文章圧縮. 言語処理学会年次大会 2017, pp. 186–189, 2017.

[4] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In Satinder P. Singh and Shaul Markovitch, editors, *AAAI 2017*, pp. 3075–3081, 2017.

[5] Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. BanditSum: Extractive summarization as a contextual bandit. In *EMNLP 2018*, pp. 3739–3748, 2018.

[6] Aishwarya Jadhav and Vaibhav Rajan. Extractive summarization with SWAP-NET: Sentences and words from alternating pointer networks. In *ACL 2018*, pp. 142–151, 2018.

[7] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *EMNLP-IJCNLP 2019*, pp. 3730–3740, 2019.

[8] Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. Sentence compression by deletion with LSTMs. In *EMNLP 2015*, pp. 360–368, 2015.