日本語 T5 モデルを用いた障害レポートからの重要箇所抽出

山下 郁海 [1] 岡 照晃 [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
yamashita-ikumi@ed.tmu.ac.jp, {teruaki-oka, komachi}@tmu.ac.jp
{manabe-akira, tanimoto-kouya}@fujielectric.com

概要

本研究では日本語 T5 モデルを用い、機器の障害情報レポートからの重要箇所抽出を質問応答形式のタスクとして実験を行う．実験の結果，BERT を用いた先行研究を上回る性能で，障害レポートからの重要箇所抽出を実現した．また T5 に入力する質問形式を変更することで抽出結果が変わることを確認し，質問の内容が重要箇所抽出の性能に影響を与えていることを明らかにした．また適切な質問を用いることで重要箇所抽出の性能が向上することを示した．

1 はじめに

現代社会は様々な機器に溢れており，そうした機器は時に不具合を起こすこともある．不具合が発生すると機器の保全担当者たちは障害情報のレポートの作成を行う．しかし，活用の際に蓄えられたレポートを人手でくまなく確認し，重要箇所を探し当てるのは労力の大きい作業である．自動で重要箇所の抽出が可能になれば，例えば，故障箇所の迅速な特定と対処や，それを踏まえた顧客への早期対応などの業務改善に役立つと考えられる．本研究で扱う障害レポートは，主に障害の**状況**，**原因**，そしてそれに対する**措置**を示す文と，それらに含まれないその他の文によって構成されている．障害レポートの状況・原因・措置のラベルが付与された文と重要箇所の抽出例を図 1 に示す．本研究は障害レポートの各文が上述の 3 種類に分類されている状況下において，各文から重要箇所を抽出することを目的として研究を行う．

本研究で扱う障害レポートはデータが少なく，追加でのデータの確保も権利の関係で難しい．そのためタスクに合わせた大規模モデルを一から作成することは困難である．そこで本研究では，学習に使用可能なデータが少量である場合でも有効性が確認されている事前学習モデル [1, 2, 3] を用いた実験を行う．本研究と同様に障害レポートからの重要箇所抽出を行った本間ら [4] の先行研究では，Transformer [5] の Encoder 部分を用いる事前学習モデル Bidirectional Encoder Representations from Transformers (BERT) [1] により，LSTM を用いた小平らの研究 [6] からの性能向上を達成した．本研究では本間ら [4] とは異なり，Encoder-Decoder モデルの事前学習モデルである Text-To-Text Transfer Transformer (T5) [7] を用いて実験を行い，障害レポートからの重要箇所抽出への有効性を検証し，BERT を用いた先行研究と比較して大幅な性能の向上を確認した．また，本研究では本間ら [4] によって提案された障害レポートからの重要箇所抽出のタスクを抽出型の質問応答形式のタスクとして扱う手法にならい，T5 モデルを用いた質問応答タスクとして実験を行った．その際，質問形式の違いによる重要箇所抽出の性能比較も実施した．結果として，適切な質問を用いることで重要箇所抽出の性能が向上することを示した．

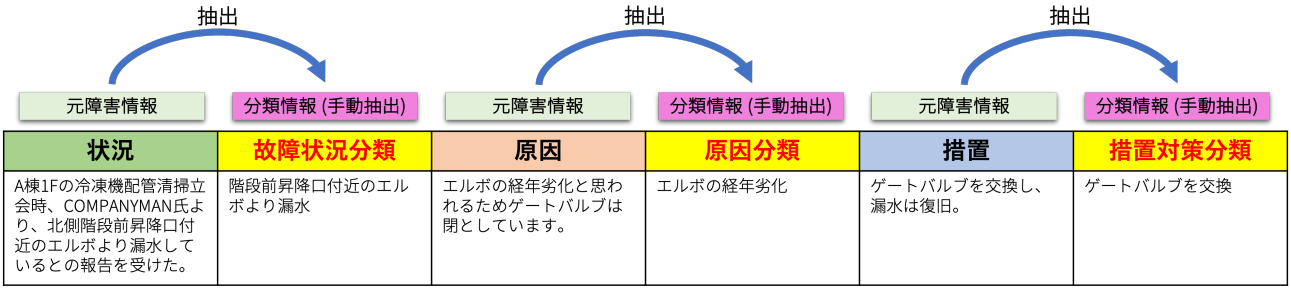
2 関連研究

2.1 事前学習モデル T5

近年，大規模なラベルなしデータで学習を行う事前学習モデルの研究が進み，様々なタスクで高い性能を発揮している．事前学習モデルの 1 つである T5 は自然言語処理に関する様々なタスクを全て系列から系列への変換とみなし，大規模な事前学習済み Transformer モデルを用いてタスクを解くことで性能向上を実現している．T5 の事前学習時の入力と出力の例を以下に示す．

入力 今日はいい<X>で<Y>やすいですね。
出力 <X>天気<Y>過ごし

ここで<X>と<Y>はマスクのための特殊トークンで



状況	故障状況分類	原因	原因分類	措置	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の例

ある．T5 の事前学習は Masked Language Model に基づいており，入力側は文書の一部のトークンがマスクされており，出力側はマスクされたトークンの予測をすることで学習を行う．

2.2　重要箇所抽出タスク

本研究と類似した形式で障害レポートからの重要箇所抽出を行っている研究として，LSTM を用いて系列ラベリングタスクとして解く小平らの手法 [6] と，BERT モデルを用いて質問応答タスクの形式を応用して解く本間らの手法 [4] が存在する．これらの研究は本研究と同様のタスクについて研究を行っているが，本研究で用いている事前学習済みの Encode-Decoder モデルを用いた実験や，質問応答形式における質問を変更した際の影響に関する分析は行っていない．

本研究で扱う障害レポートからの重要箇所抽出タスクは対象とするデータのドメインの違い，抽出すべき箇所が文や文章ではなくその一部であるなどの違いはみられるものの，抽出型要約タスクや Stanford Question Answering Dataset (SQuAD) [8] に代表される抽出型の質問応答タスクに類似している．抽出型要約の研究としては，BERT を用いて fine-tuning を行う手法の有効性が報告されており [9]，正解の要約と要約候補の BERT から得られる文書表現がより近くなるように学習を行う手法 [10] も提案されている．抽出型の質問応答の研究としては，BERT をはじめとした様々な事前学習モデルを fine-tuning する手法 [1, 2, 3] や，スパン予測に特化した学習で性能向上を図る SpanBERT [11] が存在する．また Raffel ら [7] は T5 モデルを用いて抽出型要約や質問応答の実験を行っており，BERT モデルよりも高い性能を示している．これらの研究は本研究と類似したタスクであるが，対象は英語であり．一方，本研究ではより小規模な日本語の障害レポートを対象に T5 モデルを用いた重要箇所抽出の有効性を確認する．

2.3　入力形式の変更による性能向上

事前学習モデル GPT-3 を提案した Brown ら [12] はモデルを fine-tuning するのではなく，prompt と呼ばれる入力形式を変えることで望むタスク出力を得られることを示し，注目を浴びた．また分類タスクにおいて，入力の形式を適切に変化させ fine-tuning を行うことで性能向上が可能であることが報告されている [13, 14, 15]．これらはいずれもタスクに合わせた適切な入力形式での性能向上を示唆している．本研究でも先行研究にならい質問応答の入力となる質問の形式を複数用意し比較を行った．本研究では先行研究で扱っていない重要箇所の抽出タスクを行うだけでなく、Encoder-Decoder モデルに対する入力形式の変化の有効性について，日本語での検証に取り組んでいる。

3　障害レポートからの重要箇所抽出

3.1　タスク設定

障害レポートには本研究で着目する状況や原因，措置にあたる文章以外にも，これらに含まれないその他の文章も存在する．本研究では，このような各障害レポートを人手で分類し，その他にあたる文章を除いた上で取り出された状況・原因・措置それぞれの文章を入力とし，それらから重要箇所を抽出することを目的とする．実際の人手分類後の文章（入力）と抜き出された重要箇所（目的とする出力）の例を図 1 に示す．

3.2　手法

本研究では質問応答形式で重要箇所抽出を行う手法 [4] に基づき，日本語 T5 モデルを用いた質問応答タスク形式での重要箇所の抽出を行う．本研究で扱う質問応答形式での重要箇所の抽出タスクの例を

表 1　重要箇所抽出の実験結果．太字はその列内で最もスコアが高いものを示し，下線はその列内で 2 番目にスコアが高いものを示す．質問種類がラベル別の BERT のスコアは本間ら [4] のものである．

		状況			原因			措置			全体		
質問種類	モデル	Prec.	Rec.	F_2	Prec.	Rec.	F_2	Prec.	Rec.	F_2	Prec.	Rec.	F_2
ラベル別	BERT	56.3	63.9	62.2	34.9	47.8	44.5	47.7	58.7	56.1	46.0	57.0	54.4
質問なし	T5	**71.4**	**69.2**	**69.6**	61.6	59.3	59.7	57.9	58.9	58.7	64.5	63.1	63.3
ラベル別	T5	70.2	68.7	69.0	**63.6**	**63.2**	**63.3**	**63.4**	**64.8**	**64.5**	**66.2**	**66.1**	**66.1**
統一	T5	70.0	67.7	68.1	60.9	60.9	60.9	59.9	62.8	62.2	64.1	64.1	64.1
無意味	T5	67.7	68.2	68.1	59.4	62.1	61.5	53.9	61.3	59.6	61.0	64.3	63.6

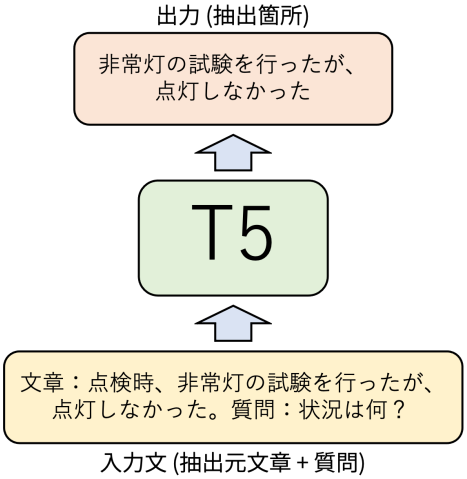


図 2　T5 を用いた質問応答形式での重要箇所抽出の例．

図 2 に示す．図 2 では，質問として各ラベルごとの文を用意し，抽出元文章と組み合わせることで質問応答のタスクとしている．抽出元文章の始まりには“文章：”をつけ，質問の始まりには“質問：”をつけた．また T5 は Encoder-Decoder の構造の生成モデルであり，出力は抽出箇所の系列がそのまま出力される．本研究では質問を 4 種類用いて，質問形式による結果の比較を行った．以下に，質問の種類とその概要を示す．

質問なし　質問を一切用いず抽出元文章のみを入力とする形式

統一　状況・原因・措置のラベルに関わらず統一の“重要箇所はどこ？”という質問を行う形式

ラベル別　ラベルごとに“状況/原因/措置 は何？”という質問を行う形式

無意味　タスクとは無関係な“今日は晴れですか？”という質問を行う形式

4　実験

4.1　実験設定

本研究では、富士電機（株）の保有する設備保全に関する障害レポートのデータを用いる．障害レポートは 194 件存在し，データ中の人名，企業名及び場所はそれぞれ MAN，COMPANY 及び PLACE へマスキング処理を施している．また一部データでアノテーション誤りにより抽出元文章内に正解の重要箇所が含まれていないものがあったためそれらを除き，状況・原因・措置を合わせて 577 件の抽出元文章と重要箇所のペアを作成した．実験の際はこのデータを訓練：開発：評価=3:1:1 として 5 分割交差検証を行う．評価手法は本間ら [4] にならい適合率に 2 倍の重みを付ける評価尺度 F_2 スコアを用いる．

本研究で用いる T5 モデルは Hugging Face [1] の自然言語処理ライブラリ Transformers [2] に基づく，Megagon Labs 公開の事前学習済み日本語 T5（語彙サイズ 32K のモデル）[3] を使用した．実装には Transformers のリポジトリに含まれるスクリプト run_translation.py を本研究に合わせて一部書き換えたものを用いた．学習時のパラメータは学習率 5e-5，バッチサイズ 32，エポック数 100 として最も性能の高いモデルを開発データで選択した．

4.2　実験結果

実験結果を表 1 に示す．結果を見ると T5 を用いたモデルはどのような質問の形式でも，全てのラベルにおいて BERT を用いた先行研究よりも大幅に性能が向上している．T5 の結果内で質問形式を比較

1）https://huggingface.co/
2）https://github.com/huggingface/transformers
3）https://huggingface.co/megagonlabs/t5-base-japanese-web

表 2　モデルの出力例.

	種類	抽出元文章（重要箇所の参照）	モデル出力
例 1	状況	9:57COMPANY にて消火栓ポンプ起動信号発報。	消火栓ポンプ起動信号発報
例 2	状況	監視業務時 COMPANY にて PA-1 故障発報、現場確認した所アラームコード 41「冷房過負荷」が発報していた。	異常発報
例 3	原因	臨時口階段 No.1（上側）バッテリー不良	バッテリー不良
例 4	原因	AC-8 奥の壁面より漏水	奥の壁面より漏水
例 5	措置	本日 COMPANY にてモーターの交換実施。問題なく運転し異常ない為、済とする。	モーターの交換
例 6	措置	電力監視設備更新した為済とする。	電力監視設備更新 電力監視設備更新

すると，質問なしのモデルとラベル別またはラベル間で統一の質問を用いるモデルの結果の比較から，質問を用いることで性能が向上していることがわかる．一方，無意味な質問を用いるモデルは質問なしのモデルとほとんど同程度の性能であり，単に質問を追加するだけでは性能は向上せず，質問の内容が重要であることがわかる．また最もスコアの高いラベル別の質問を行うモデルとラベル間で質問を統一したモデルを比較すると，ラベルごとに質問を分けることが性能向上に寄与していることもわかる．

5　分析

5.1　出力例の分析

実際のモデルの出力例を表 2 に示す．抽出元文章の赤色で書かれた箇所は重要箇所の参照である．出力に用いたモデルは，表 1 (3) の（全モデル中で最も性能の高かった）ラベル別の質問を文章の後ろにつけて入力を行ったモデルである．例 1，例 3，例 5 を見ると，それぞれのラベルに対して適切な出力が行われていることがわかる．また，例 4 を見ると，完全一致ではなく抽出範囲に誤りがあるものの，抽出すべき箇所の判断は正しく行われており正解に近い出力ができていることもわかる．一方，T5 は抽出モデルではなく生成モデルのため，抽出モデルでは起こりえない誤りも存在している．例 2 と例 6 は生成モデル特有の誤りであり，例 2 では抽出元の文章には存在しない単語を出力している．例 6 では抽出箇所は正しいものの，繰り返しが起きてしまっていることがわかる．

5.2　人手評価

5.1 節で見た表 2 の例 4 にもあるように T5 を用いたモデルの出力は，正解との完全一致ではないが部分一致であるようなものが多くみられた．そこで

表 3　人手評価の結果.

種類	評価者 1（件）	評価者 2（件）	割合（%）
完全一致	-	-	46.1
i	103	112	18.6
ii	30	43	6.3
iii	125	140	23.0
iv	53	16	6.0

表 1 の (3) の出力の人手評価を行った．具体的には，全 577 件のデータのうち，モデルの出力と正解が完全一致であったものを除く 311 件について，2 人の評価者が下記の 4 段階での評価を行った．

i. 正解と出力は完全一致ではないが、問題なく出力は受け入れられる
ii. 正解と出力は完全一致ではないが、正解の方に問題があり出力は受け入れられる
iii. 正解と出力は完全一致ではなく、出力は受け入れられない
iv. 正解と出力は完全一致ではなく、正解に問題があるが、出力も受け入れられない

それぞれの評価の具体例を付録 A に示す．表 3 に人手評価の定量評価結果を示す．出力が受け入れられると評価されているもの（i と ii）の割合が全体の 24.9%となっており，完全一致のものと合わせて 71.0%のモデル出力が受け入れられるものとなった．

6　おわりに

本研究では日本語 T5 モデルを用いた障害レポートからの質問応答形式での重要箇所抽出の研究に取り組み，日本語 BERT モデルを用いた先行研究と比較し高い性能が得られることを示した．また質問の形式を適切なものに変えることで，質問を用いない場合や適切ではない質問を用いる場合と比較し，性能向上を確認した．今後はより適切な質問形式に関して詳細な研究を行っていきたい．

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**, pp. 4171–4186, 2019.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach, 2019.

[3] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite bert for self-supervised learning of language representations. In **ICLR**, 2020.

[4] 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, pp. 189–193, 2021.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **NIPS**, Vol. 30, pp. 5998–6008, 2017.

[6] 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 第 23 回年次大会, pp. 189–193, 2017.

[7] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **JMLR**, Vol. 21, pp. 1–67, 2020.

[8] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In **ACL**, pp. 784–789, 2018.

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

[10] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. In **ACL**, pp. 6197–6208, 2020.

[11] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. **TACL**, Vol. 8, pp. 64–77, 2020.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In **NeurIPS**, Vol. 33, pp. 1877–1901, 2020.

[13] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In **EACL**, pp. 255–269, 2021.

[14] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. GPT understands, too. **arXiv:2103.10385**, 2021.

[15] Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. In **NAACL-HLT**, pp. 2339–2352, 2021.

付録

A　人手評価の各評価の例

出力の人手評価における各評価の例を表 4 に示す．評価 1 は，正解の重要箇所が過不足なく必要な情報を持っており，重要箇所とモデル出力との差が，モデル出力の方が正解の重要箇所より長い場合は余計な要素がなく，モデル出力の方が正解の重要箇所より短い場合は差分にあたる箇所の重要度が低いものである．評価 2 は，正解が必要最低限の情報を持っていないなどの理由で問題がある一方で，モデル出力が必要な情報を抽出できている場合である．評価 3 は，評価 1 と同様に正解の重要箇所が過不足なく必要な情報を持っている一方で，モデル出力に余計な要素が存在している，必要な情報まで削ってしまっている，そもそも抽出箇所が異なる，などのものである．評価 4 は，評価 2 と同様に正解が必要最低限の情報を持っていないなどの理由で問題があるもので，かつモデル出力も評価 3 と同様の理由で問題があるものである．

表 4　人手評価における各評価例.

評価	種類	抽出元文章（重要箇所の参照）	モデル出力
1	状況	監視時、COMPANY にて AC-1 差圧異常が発報。	差圧異常
2	措置	瞬時の通信異常でオフラインとならないように発報まで 60 秒のタイマーに設定しております。設定変更後、発生していない為、一度済とします。	瞬時の通信異常でオフラインとならないように発報まで 60 秒のタイマーに設定
3	原因	いずれも排煙口は動作していなかった。防災盤からの信号に異常があると思われる。業者診断要す。	排煙口は動作していなかった
4	措置	垂れ壁復旧し、念のため感知器交換実施。責任者には感知器に養生する際は埃をたてないよう注意しています。	垂れ壁復旧