CE-003

災害発生時のクロノロジーに対する優先度推定
Estimating Priority of Tasks in Japanese Chronology Logs in Disasters

孝壽 真治[1)]　竹内 孔一[2)]　渡邉 暁洋[3)]　平山 隆浩[4)]　中尾 博之[4)]
Shinji Koju　Koichi Takeuchi　Akihiro Watanabe　Takahiro Hirayama　Hiroyuki Nakao

1　はじめに

災害発生時には県や保健医療調整本部，病院などで災害対策本部が設置され，医療機関や消防などの様々な組織と連携して保険医療を提供する．災害対策本部には，連携している医療機関や避難所などから医療ニーズや資源・物資の不足，ライフラインの状態のような様々な情報や要請，報告が集まる．災害対策本部では，このような情報から連携先の各種機関や医療チームなどに適切な指示を出す必要がある．しかし，災害対策本部に届く情報は膨大であるため，混乱を防止し，情報を管理し，適切な指示を出すことを目的として，クロノロジーと呼ばれる経時活動記録が作成される．

クロノロジーには，災害対策本部に届けられた情報と災害対策本部から連携先の各種機関へと発信した情報の全てが時系列形式で記録される．そのため，クロノロジーとして記録されている情報には，人命に関わる対応の優先度が高いものと，人命に関与しない単なる報告のような対応の優先度の低いものが混在している．災害対策本部の職員は，クロノロジーに記録された情報を精査して，人命に関わるものや，緊急性の高いものを判別し，その優先度を考慮して適切に対応するという業務を行なっている．

そのため，システムがクロノロジーの優先度を自動的に推定することができれば，災害対策本部職員の負担を軽減して災害対策に時間を割り当てることが可能になる．災害対策本部で受信，発信される情報は，クロノロジーとして時間順に短い文書で電子的に記録されている．そこで，この 1 件 1 件のクロノロジーに対して，機械学習を利用して優先度を推定するシステムを構築する．システムの構築には，災害医療のトリアージの観点から高・中・低の 3 段階の優先度を定義し，人手により分類された評価データを使用した．

本研究では Decoder モデルである GPT-NeoX をテキストの Embedding 手法として用いて構築したモデルの優先度推定精度を検証した．検証では Bag-of-Words (BOW) や，BERT などのテキストの Embedding 手法を用いてモデルを構築した場合と優先度推定精度を比較した．実験の結果として，GPT-NeoX モデルが優先度の識別に対して高い F 値を示すことを明らかにする．

1）　岡山大学大学院環境生命自然科学研究科　Graduate School of Environmental, Life, Natural Science and Technology, Okayama University
2）　岡山大学学術研究院　Academic Research Assembly, Okayama University
3）　兵庫医科大学危機管理医学講座
4）　岡山大学学術研究院医師薬学域災害医療マネジメント講座　Okayama Uniersity, Dept. Disaster Medicine and Manage- ment (DMMA), Academic Field of Medicine, Dentistry and Phrmaceutical Sciences

2　関連研究

大きく分けて災害時に関係する文書処理の研究と，自然言語処理における文書分類の研究が関連する．まず，災害時における文書処理に関連した研究では実システムと結びつけた研究が展開されている．クロノロジーのように時系列で記録する方法をベースに災害情報を共有するシステム「災害ネット」[1]，災害時における Twitter などを含む SNS の情報を活用するシステム「DISAANA」[2][1)]，「高度自然言語処理プラットフォーム」[3] が提案されている．

上記の 3 件の研究は災害に関連したテキストを処理しているが，後者の 2 件は SNS の文書であるため，本研究が対象としているクロノロジーと文書の質が異なる．また，前者の研究はクロノロジーを対象にしているが本研究のように優先度の識別に関する詳細な分類手法は提案されていない．

自然言語処理において文書をある目的に分ける文書分類タスクでは，近年，機械学習を利用した手法が提案されている [4]．機械学習を利用した枠組みでは，分類クラスを定義した後に，文書を人手により目的とするクラスに分類した学習用のデータを作成して機械学習を適用する．文書分類に利用されてきた機械学習のモデルとしては，サポートベクターマシン (SVM)[5] や XGBoost[6] などが挙げられる．

これらの機械学習のモデルを利用するためには文書をベクトル化する必要がある．基本的な手法として文書を語彙空間でベクトル化する BOW がある．また，深層学習モデル Transformer[7] の Encoder 部分を利用して大規模テキストを学習することで文脈を考慮した文書ベクトルが作成できる BERT[8] が提案されている．ベクトル化手法と文書分類モデルの組み合わせにより様々な文書分類手法が可能となる．また，Transformer の Decoder 部分を利用して主に文書生成などのタスクで利用できる GPT-3[9] や，オープンソースのモデルとしては GPT-NeoX-20B[10] がある．これらのモデルは Few-shot learning などの手法により様々なタスクに利用可能となる．

本研究のクロノロジーに対する優先度分類の問題に対して文書分類手法を適用した予備的な研究 [11, 12] が行われている．クロノロジーの優先度を分類したデータ構築においてこれらの手法を参考にする．一方で，優先度分類モデルでは先行研究では SVM，XGBoost，およびニューラルネットワークを比較し，ニューラルネットワークが有効であることが示されている．本研究では先行研究を参考にニューラルネットワークを利用した文書分類モデルを適用する．一方で，先行研究では BOW や Encoder モデルである BERT が用いられているが，本研究では Decoder モデルである GPT-NeoX を用いた文書分類モデルを構築しているためこれらの先行研究と異

1）　現在 DISAANA は 2023 年 12 月 28 日に終了した．

なる．

3　データ

本研究で使用するクロノロジーの特徴と優先度，データセットの形式について説明する．

3.1　クロノロジー

本論文では，西日本豪雨災害において実際に作成されたクロノロジーのデータを使用する．クロノロジーは，災害発生時に災害対策本部でやり取りされる情報に対して，日時・発信者・受信者・報告内容などの項目をホワイトボードに記録したものであり，これは表 1 のように電子化されている．

クロノロジーの特徴として，Patient (患者) を「Pt」と表記するといった略語や，透析・DMAT (災害派遣医療チーム：Disaster Medical Assistance Team) などの専門用語，「→」のような記号，体言止めや短い表現などが多く使用される．その他にも，1 つの項目に複数のやり取りが含まれる場合もある．

表 1　電子化されたクロノロジーの例[1)]

年月日	時刻	発信者	受信者	内容
2018/7/9	09:50	○大 ○○さん	岡山○○ Dr.	昨日までの流れをプレゼン 今日の方針を話し合い
2018/7/9	14:15	本部 DMAT ○○○	県 DMAT 本部	DMAT ○○ 1 チーム 2 チーム DMAT を要請
2018/7/9	16:59	○○医療 センター	KuDRO 本部	[○○病名の略称○○] ○○消防救急車で○大へ ○○才女性

3.2　優先度

本研究で使用したデータでは，先行研究 [11] と同様に災害医療のトリアージの考え方を参考に重要度と緊急度の 2 つの観点から優先度を分類している．重要度は生命の危機に関することであれば高いとし，緊急度は時間的猶予が無い場合に高いとする．つまり，表 2 に示すように，重要かつ緊急度が高いクロノロジーを優先度を高とし，どちらか一方でも高いものを中，どちらも低いものを低として分類する．

表 2　優先度の付与基準

優先度	重要度	緊急度
低	低い	低い
中	高い	
高		高い

この基準をもとに，2 人のアノテーターが同じクロノロジーに対して優先度を付与し分類の一致度を測定した．4463 件のクロノロジーに対して一致率は 0.947 であり，Kappa 値 [13] は 0.598 であった．

3.3　データセットの特徴

実験に用いるクロノロジーデータでは，上述の 2 名が付与した優先度を平均したもの使用している．実験ではクロノロジーデータの内，クロノロジーの内容と付与した優先度ラベルのみを使用するため，実験で使用するデータセットは 表 1 の内容と優先度ラベルの組となる．

クロノロジーデータを表 3 に示すように学習，検証，テストデータに分割する．それぞれの優先度の分布を見ると，優先度の低が非常に多く，優先度の高・中が少なく，優先度が偏ったデータになっていることがわかる．特に，優先度高は全体の 1%程度である．本研究では，偏りの大きいクロノロジーの優先度の高・中に対して推定精度の向上を目指す．

1)　文章の一部を記号に置き換えている

表 3　実験に使用するデータの優先度の内訳

モデル	学習 (割合)	検証 (割合)	テスト (割合)	全件 (割合)
低	2444 (0.913)	794 (0.889)	802 (0.898)	4040 (0.905)
中	211 (0.079)	85 (0.095)	79 (0.088)	375 (0.084)
高	22 (0.008)	14 (0.016)	12 (0.013)	48 (0.011)
合計	2677	893	893	4463

4　提案手法

本研究で構築したモデルおよび，学習手法について説明する．

4.1　構築したモデル

本研究ではテキストの Embedding 手法として，Bag-of-Words (BOW) を用いる手法，BERT を用いる手法，GPT-NeoX を用いる手法の 3 種類の手法を用いた．これらを用いて構築したモデルについて以下で説明する．

4.1.1　Bag-of-Words (BOW) モデル

BOW によるベクトル化には，形態素解析器として MeCab を使用し，辞書には NEologd[14] を用いる．BOW の語彙として，出現した形態素のうち品詞が名詞・形容詞・動詞であり，かつ，形態素が使用されている文章数が 10 個以上かつ 3 割未満のものを使用する．この結果，BOW の語彙数は 1177 となった．

優先度の推定では図 1 のように，BOW のベクトルを 3 層ニューラルネットワークに与えている．ネットワークの中間層のユニット数は，入力層に近い側から 300，100，3 とした．

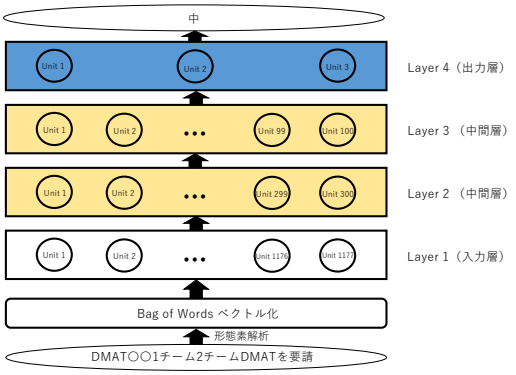

図 1　BOW を利用したモデルの構成図

4.1.2　BERT モデル

東北大学が作成した事前学習済み BERT[2)] にファインチューニングを適用して使用する．BERT のファインチューニングでは，最終層から 4 層のみを学習させて，その他の層は学習時に更新しないようにした．

優先度の推定では図 2 のように，BERT の [CLS] トークンに対応するベクトルのうち，最終層から 4 層のベクトルを連結して出力層に入力として与える．

2)　https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking 2023/5/21 アクセス

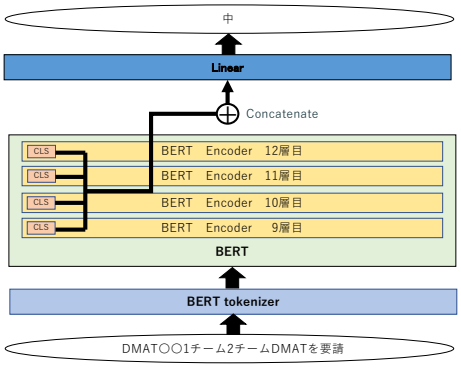

図 2　BERT を利用したモデルの構成図

4.1.3　GPT-NeoX モデル

rinna 株式会社が作成した事前学習済み GPT-NeoX[3] にファインチューニングを適用して使用する．使用する VRAM サイズを削減するため GPT-NeoX に LoRA [15] を適用してファインチューニングを行う．

優先度の推定では図 3 のように，GPT-NeoX の [EOS] トークンに対応する最終層のベクトルを出力層に入力として与える．

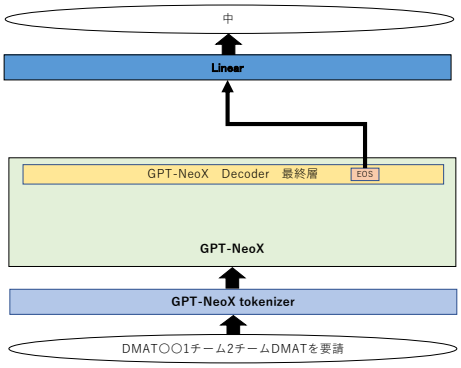

図 3　GPT-NeoX を利用したモデルの構成図

4.2　学習手法

本研究におけるクロノロジーの優先度ラベルは，高・中・低の 3 つであり，3 つのラベルは順序関係がある 3 段階のクラスとなっている．そのため，優先度推定は順序回帰の手法を用いて解く．順序回帰として解く際にはいくつかの手法が存在するが [16]，本研究では損失関数の目標出力にソフトラベルを与える手法 [17] を用いた．よって，学習時の損失関数の目標出力には，正解クラス $\overline{k}$ とあるクラス k $(k = 1, 2, ..., K)$ との距離を $|\overline{k} - k|$ として，式 (1) で求められる表 4 のようなソフトラベルを与える．

$$d_k = \frac{\exp(-|\overline{k} - k|)}{\sum_{i=1}^{K} \exp(-|\overline{k} - i|)} \tag{1}$$

表 4　各優先度における目標出力

優先度	ハードラベル	ソフトラベル
低	[1, 0, 0]	[0.665, 0.245, 0.090]
中	[0, 1, 0]	[0.212, 0.576, 0.212]
高	[0, 0, 1]	[0.090, 0.245, 0.665]

3)　https://huggingface.co/rinna/japanese-gpt-neox-3.6b 2023/6/1 アクセス

5　実験

本実験では Decoder モデルである GPT-NeoX の優先度推定精度を検証する．推定精度の比較対象として先行研究で使用されている BOW と BERT を用いた．

5.1　実験設定

全てのモデルにおいて，学習時の損失関数には高・中・低それぞれに対して 4・2・1 の重みを与えて学習した．また，学習では開発データの loss が 10 回更新されなかった場合に停止し，開発データに対する loss が最も小さい epoch の学習結果を利用する．この際に，開発データの loss の計算で使用する損失関数の目標出力は表 4 のようなハードラベルを与えた．重みに関しては，クロノロジーの優先度高と中の推定精度を重視するために，優先度が低の場合 0.0001 のような低い値とした．

こうして学習した各モデルが推定したクラスに対して，適合率・再現率・F 値を用いて評価する．

5.2　実験結果

全てのモデルに対する実験結果を優先度が高・中・低ごとにまとめて表 5 から表 7 に示す．

表 5　優先度：高の実験結果

モデル	Precision	Recall	F 値
BOW モデル	0.50	0.50	0.50
BERT モデル	0.43	0.50	0.46
GPT-NeoX モデル	0.75	0.50	0.60

表 6　優先度：中の実験結果

モデル	Precision	Recall	F 値
BOW モデル	0.28	0.51	0.36
BERT モデル	0.30	0.59	0.40
GPT-NeoX モデル	0.37	0.54	0.44

表 7　優先度：低の実験結果

モデル	Precision	Recall	F 値
BOW モデル	0.95	0.88	0.91
BERT モデル	0.96	0.87	0.91
GPT-NeoX モデル	0.95	0.91	0.93

実験結果より，GPT-NeoX モデルは優先度高・中・低のいずれにおいても，最も高い F 値となった．GPT-NeoX モデルは，優先度高では F 値が 2 番目に高い BOW モデルより 10%高い 0.60 となった．優先度中では F 値が 2 番目に高い BERT モデルより 4%高い 0.44 となった．優先度低では F 値が 2 番目に高い BOW モデル・BERT モデルより 2%高い 0.93 となった．

6　考察

実験結果より，BOW モデルや BERT モデルと比較して GPT-NeoX モデルが高い優先度の推定精度を示すことが明らかとなった．

特に，文書分類などのタスクでは BERT などの Encoder モデルがよく利用されているが，今回の実験では BERT モデルよりも GPT-NeoX モデルの方が高い性能を発揮している．これに関して，本実験で使用した BERT-Base のパラメータ数は 1.1 億であるのに対して，GPT-NeoX は 36 億でありパラメータ数に大きな差があ

る．また事前学習で使用されたデータは，BERT-Base が日本語 Wikipedia (2019 年) を使用しているのに対して，GPT-NeoX は日本語 Wikipedia (2023 年) に加えて CC-100，mC4 データセットの日本語データを使用しており，事前学習で使用されたデータ数にも違いがある．このような，パラメータ数と事前学習で使用するデータ数が大きいことが，クロノロジーの優先度推定精度に寄与していると考えられる．そのため，本研究で利用した GPT-NeoX の 36 億パラメータよりも大きいモデルを利用することで，さらに推定精度が向上する可能性がある．

一方で，本研究で使用した BERT と GPT-NeoX はどちらも日本語データのみで事前学習が行われている．近年の大規模言語モデルの中には，特定の言語だけではなく多言語のデータセットを利用して膨大なデータで事前学習を行ったモデルが存在する．例えば，LLaMA[18] は複数の言語の Wikipedia が学習データに含まれており，これをファインチューニングして作成されたオープンソースの Vicuna[19] などのモデルが存在している．このような多言語で学習された大規模言語モデルは言語間要約 (CLS) などのタスクで有効性を示しており [20]，日本語の知識を有するモデルであればクロノロジーの優先度推定などの文書分類タスクでも効果を発揮する可能性が考えられる．

7 まとめ

本研究では，クロノロジーの優先度を推定する機械学習モデルを構築し，その精度の比較を行なった．実験の結果として GPT-NeoX を用いたモデルが，BERT や BOW を用いたモデルよりもクロノロジーの優先度推定において高い精度を示した．

また，GPT-NeoX が BERT よりもパラメータ数，事前学習で用いられたデータ数ともに大きいことから，パラメータ数や事前学習のデータ数がより大きなモデルを用いることで，さらなる推定精度向上の可能性が考えられる．

謝辞

本研究の一部は厚生労働科学研究費補助金 (21IA2401)(数理最適化モデルによる小学校グリッドに基づく他組織連携システム (MACS) の解析)(略称 A-MACS) に基づいて実施された．

参考文献

[1] UNISYS. クロノロジー型危機管理情報共有システム 災害ネット. (https://www.unisys.co.jp/solution/biz/disaster-net/ 2022/05/21 アクセス).

[2] NICT. 対災害 SNS 情報分析システム (DISAANA), 2020. (https://www.nict.go.jp/resil/ 2023/5/21 アクセス).

[3] NEC. Twitter 上の災害に関する情報をリアルタイムで解析・可視化する「高度自然言語処理プラットフォーム」, 2020. (https://jpn.nec.com/press/202006/20200626_01.html 2023/05/21 アクセス).

[4] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. A Survey on Text Classification: From Traditional to Deep Learning. *ACM Transactions on Intelligent Systems and Technology*, Vol. 13, No. 2, pp. 1–41, 2022.

[5] Thorsten Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of European Conference on Machine Learning*, pp. 137–142, 1998.

[6] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All You Need. *arXiv preprint arXiv:1706.03762*, 2017.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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, 2020.

[10] Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136, virtual+Dublin, May 2022. Association for Computational Linguistics.

[11] 竹内孔一, 山崎瑶, 渡邉暁洋, 平山隆浩, 中尾博之. 災害医療におけるクロノロジーの分析. 電子情報通信学会 信学技報, Vol. 121 (415) NLC2021-31, pp. 19–23, 2022.

[12] 孝壽真治, 竹内孔一, 渡邉暁洋, 平山隆浩, 中尾博之. 災害医療におけるクロノロジーの優先度識別. 研究報告情報基礎とアクセス技術（IFAT）, Vol. 2023-IFAT-149 (3), pp. 1–5, 2023.

[13] Roger Bakeman and John M. Gottman. *Observing Interaction*. Cambridge, 1986.

[14] 佐藤敏紀, 橋本泰一, 奥村学. 単語分かち書き辞書 mecab-ipadic-neologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会, pp. 875–878, 2017.

[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *CoRR*, Vol. abs/2106.09685, , 2021.

[16] 岡谷貴之. 深層学習改訂第 2 版. 講談社, 2015.

[17] Raul Diaz and Amit Marathe. Soft Labels for Ordinal Regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[19] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[20] Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. Zero-shot cross-lingual summarization via large language models, 2023.