空調分野におけるドメイン特化コーパス構築手法の検討

森本 康太 [1]　森田 隆紘 [1]　冨士本 直起 [1]　比戸 将平 [1]　小田 悠介 [2]
[1] ダイキン工業株式会社 [2] 国立情報学研究所 大規模言語モデル研究開発センター
{kota.morimoto, takahiro2.morita, naoki1.fujimoto, shohei.hido}@daikin.co.jp
odashi@nii.ac.jp

概要

ドメイン特化型大規模言語モデル（LLM）の開発には，高品質で十分な規模のドメインコーパスを用いた継続事前学習が不可欠である．しかし，空調分野のように専門性が高く公開データが乏しい産業領域において，Web コーパスから高品質なドメイン文書を抽出し，ダウンストリームタスクにおける性能まで検証した実証研究は限られている．本研究では，日本語 Web コーパスからキーワードフィルタにより取得したシードを起点とし，分類器による候補抽出と LLM による品質評価を反復するコーパス構築手順を提案する．提案手法により高品質文書 167,401 件（約 0.16B tokens）を収集し，FineWeb2-ja を用いた日本語空調コーパスの構築と継続事前学習の結果，空調分野の質問応答ベンチマークにおいてベースモデルより正答率が 4 ポイント向上した．

1　はじめに

大規模言語モデル（LLM）の発展により，製造業においても問い合わせ対応や設計支援など，専門知識を要する業務への応用が進みつつある．一般的に用いられる汎用 LLM では専門用語や暗黙の前提を扱うことが難しい場合がある．実際にダイキンの社内業務における汎用 LLM の適用可能性を検証したところ，空調負荷の計算や機種名称の扱いなどで誤りが多く生じており，業務で求められる回答精度を安定して満たすことが困難であった．このような課題に対して，継続事前学習によりドメイン固有の知識・表現を獲得させるドメイン特化 LLM の開発が有効なアプローチとして注目されている [1, 2]．

ドメイン特化 LLM 実現のためには，学習データの整備，継続事前学習の学習設定，評価方法の設計など複数の工程が存在する．中でも学習データは LLM の性能に与える影響が大きく，追加学習に用いるコーパスの品質と規模に性能向上が強く依存することが報告されている [3]．一方で空調分野には大規模な公開コーパスが存在せず，一企業が保有するデータだけで十分な規模のテキスト情報を確保することも容易ではない．そこで本研究では，ノイズが多い一方で大量のテキストを確保しやすい Web コーパスから空調ドメインに有用な高品質文書の抽出方法を提案する．

大規模なコーパスから学習データを抽出する研究として，Shao ら [4] は，高品質な初期シードから学習した分類器を反復的に更新しつつ，Web コーパスからドメイン文書を段階的に拡張する手法を示した．Zhou ら [5] は，粗抽出した候補文書に対して LLM で数学関連度をスコア付与し，そのラベルで分類器を再学習することで，より高品質な文書集合を得る手法を提案した．Penedo ら [6] は，LLM で文書の教育的品質をスコア化し，そのスコアに基づくモデルで FineWeb 全体をスコアリングして教育的品質の高い文書集合（FineWeb-Edu）を抽出した．

これらの先行研究は，分類器と LLM を組み合わせた反復的なデータ精製が有効であることを示している一方，日本語を対象とした産業応用において要求されるデータ品質やドメイン適合性については十分に検討されていない．本研究ではこれらの先行研究を参考に，日本語 Web コーパスを対象にキーワードフィルタで作成した初期シードを起点とし，分類器による候補抽出と LLM によるスコアリングを反復的に適用することで，空調分野に特化した高品質コーパスを構築する．

2　空調コーパスの構築手法

空調特化の高品質コーパスの構築手順を図 1 に示す．本手法は以下の 4 段階から構成される．

1. FineWeb2 日本語（FineWeb2-ja）[1)] に対してキーワードフィルタリングを行い，分類器の学習に

1） https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v4/-/tree/main/ja/ja_fineweb-2

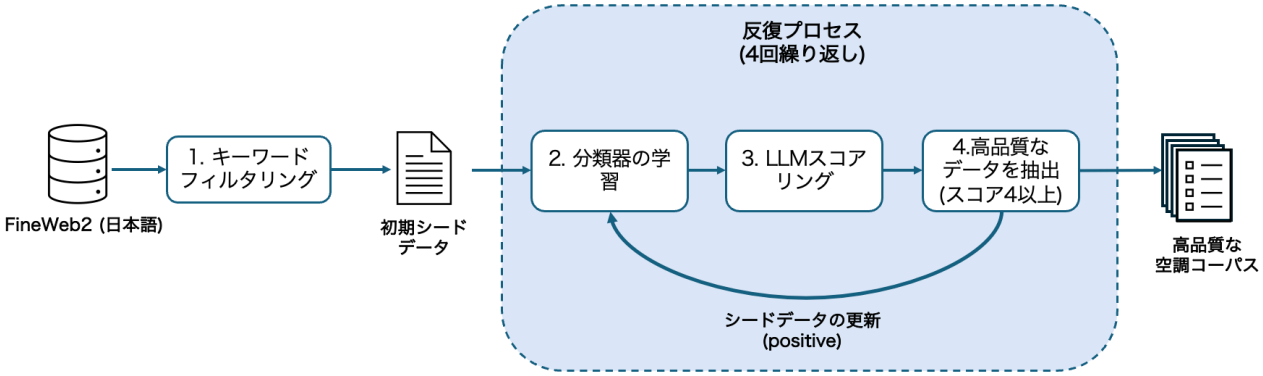

図 1　空調特化の高品質コーパスの構築手順

用いる初期シードを構築する（2.1 節）
2. 初期シードを用いてドメイン判定分類器を学習する（2.2 節）
3. 学習した分類器を Web コーパスへ適用してドメイン候補文書を抽出し，LLM により各候補文書をスコアリングする（2.3 節）
4. LLM スコアに基づき高品質文書を選別し，選別結果を正例としてシードを更新する（2.4 節）

2〜4 を 1 サイクルとして計 4 回反復する．最終サイクルで得られた候補集合中の全文書に対して LLM を用いてスコアリングを実施し，その結果に基づいて高品質文書を選定する．

2.1　キーワードフィルタリング

空調分野においては公開データが乏しく，分類器学習に用いる初期シードの収集が課題である．前節で示したように，本研究では反復的なコーパス構築手法を採用するため，初期シードの品質が低くても反復処理により緩和されると考えられる．このため，計算コストが低く容易に適用可能なキーワードフィルタリングにより初期シードとなる空調関連文書を収集した．具体的には，まず事前に定義した必須キーワードを 1 つ以上含み，かつ除外キーワードを含まない文書のみを候補とする．次に候補文書 d に対して，(1) 文脈キーワードの出現，(2) 必須キーワードとの共起，(3) URL 情報によるドメイン示唆，の 3 条件をそれぞれ判定し，1 つ以上の条件を満たすものをシード文書として抽出する．

2.2　空調ドメイン判定分類器の学習

テキストの空調ドメイン判定には fastText[2] による教師あり二値分類器を採用した．前処理として MeCab[3] により文書を分かち書きし，2.1 節で得たシード文書を正例，FineWeb2-ja からランダムにサンプリングした文書を負例として学習データを構築した．学習ハイパーパラメータは，学習率を 0.2，エポック数を 5，埋め込み次元を 256，word n-gram 長を 2，最小出現回数を 2 とした．こうして学習した分類器を FineWeb2-ja に適用することで，空調ドメイン候補文書を抽出する．

2.3　LLM スコアリングの適用

分類器で抽出した空調ドメイン候補に残るノイズを除去するため，追加手段として記述内容に基づく文書スコアリングを適用する．人手による品質評価が精度面では最も望ましいが，数十万件規模の候補文書を人手で評価するのは時間・コストの面で現実的ではないため，LLM を用いて大規模な文書評価を行うこととした．具体的には，空調分野の知識獲得に有用かを基準として，文書の関連性と技術的深さを 1〜5 点で評価する．LLM には文書テキストと評価基準を入力し，スコア（1〜5）と判定根拠（100 文字程度）を出力させる．

2.4　LLM スコアリング結果によるシード更新と分類器の再学習

キーワードフィルタリングで構築した初期シードを用いて学習した分類器は，キーワードへの依存が強く，文脈を十分に捉えられないため分類結果として得られる候補文書の品質が低い．そこで，LLM

2）https://fasttext.cc/
3）https://taku910.github.io/mecab/

によるスコアリングで 4 点以上と評価された高品質文書を正例として再度ラベル付けした学習データを用い，分類器を再学習することで高品質データを段階的に増量する．なお，シード更新では誤ラベル混入を避け高精度化を優先するため LLM スコア 4 点以上の文書を用い，最終的なコーパス抽出ではカバレッジ確保のため 3 点以上の文書を採用した．

3 実験

本節では，2 節で述べた反復的データ選別手順に基づき，空調特化コーパスの構築過程を評価する．具体的には，(1) 各回における分類器が抽出した文書の空調ドメインへの適合率（precision）の推移，(2) LLM スコア（品質指標）の推移，の 2 観点から分析する．

実験では，シードデータから fastText 分類器を学習し，抽出，LLM スコアリング，シード更新からなる手順を 4 回反復した．初回の抽出・評価を含め，抽出と評価は計 5 回実施した．LLM スコアリングには Qwen3-235B-A22B-Instruct-2507[4] を用いた．

さらに，構築した空調特化コーパスのダウンストリームタスクにおける有用性を確認するため，本コーパスを用いて Qwen3-4B（Base）[5] に対して継続事前学習を行った．学習は NeMo 2.0[6] で実施し，評価には llm-jp-eval（v2.1） [7] を用い，評価仕様の詳細は同文献に従った．また，評価 llm-jp-eval に含まれない独自評価データセットとして，空調分野に特化した 121 問からなる QA 評価セット Daikin QA を用いた．学習は 2 エポックとし，1 エポックあたり 170 ステップ（計 340 ステップ）とした．グローバルバッチサイズは 256，シーケンス長は 4096 とし，学習率は 1.0e-5 で一定とした．ドメイン特化データへの過度な適合を避けるため，学習データに日本語

表 1 各イテレーションにおける分類器の抽出結果（件数・文字数・適合率）

イテレーション	件数	文字数	適合率
1 回目	8,184,027	21,316,790,666	4.0%
2 回目	1,255,709	1,708,127,227	17.0%
3 回目	790,289	970,505,001	24.0%
4 回目	749,060	902,194,920	28.0%
5 回目	777,875	933,322,281	25.0%

4） https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507
5） https://huggingface.co/Qwen/Qwen3-4B-Base
6） https://github.com/NVIDIA-NeMo/NeMo

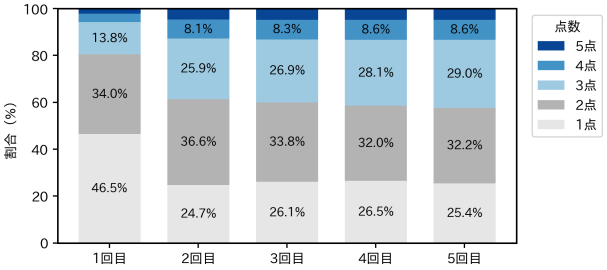

図 2 各サイクルにおける，fastText の予測確率上位 200,000 件に対する LLM 品質評価スコア（1〜5 点）の割合

Wikipedia（ja-wiki）[7] を学習データの 10%となるよう混合した．チェックポイントごとの値のばらつきを軽減するため，性能評価には最終 3 チェックポイントの相加平均モデルを使用した．

3.1 分類器の評価

表 1 に，分類器により各イテレーションで「空調」と判定されたデータの件数，総文字数，および適合率を示す．適合率は，各イテレーションで抽出されたデータからランダムに 100 件を抽出し，人手でアノテーションした結果に基づいて算出した．イテレーションを繰り返すにつれて抽出件数は 1 回目から大幅に減少し，3 回目以降は約 75〜79 万件で推移した．一方で適合率は向上し，特に 1 回目から 3 回目にかけての改善が顕著であった．

3.2 LLM スコアリング

fastText 分類器により正例と判定された候補文書のうち，確信度が高い順に上位 20 万件を対象として，LLM スコア 1〜5 点の品質評価を行った結果を図 2 に示す．1 回目（キーワードフィルタのみでシードを作成）では，低品質（1〜2 点）の割合が高く，文書品質が低いことが確認された．一方，2 回

表 2 最終コーパス（約 78 万件）に対する LLM スコアリング結果

スコアリング	件数	トークン数
1 点	376,172 (48.36%)	0.290B
2 点	234,302 (30.12%)	0.196B
3 点	131,147 (16.86%)	0.123B
4 点	23,892 (3.07%)	0.021B
5 点	12,362 (1.59%)	0.015B
合計	777,875 (100.00%)	0.645B
3 点以上	167,401 (21.52%)	0.159B

7） https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v4/-/tree/main/ja/ja_wiki

表 3　各評価ベンチマークにおけるベースモデルと本研究モデルのスコア比較

評価項目	Daikin QA	EL	FA	HE-EN	HE-JA	MR	MT	NLI	QA	RC	Ave.
ベースモデル	0.45	**0.33**	**0.15**	0.38	**0.27**	**0.73**	**0.72**	**0.62**	**0.21**	**0.65**	**0.45**
本研究	**0.49**	0.32	0.13	**0.43**	0.24	0.52	0.54	0.55	0.17	0.52	0.39

目以降は LLM スコアリングで 4 点以上の文書のみをシードとして利用した結果，3 点以上の割合が増加し，文書品質が改善した．具体的には，LLM スコア 3 点以上の文書の割合は 1 回目の 19.53%から 5 回目の 42.42%へ増加し，1 回目と比べて約 23 ポイント改善した．

3.3　最終コーパスの構築

5 回目のイテレーションを実施した後，得られた約 78 万件の文書全件を対象に LLM によるスコアリングを行った結果を表 2 に示す．トークン数は，継続事前学習で用いたモデル（Qwen3-4B）のトークナイザにより算出した．前述したように最終コーパスのうちスコア 3 以上の文書を学習に有用として選別し，結果として 167,401 件（約 0.16B tokens）を LLM の継続事前学習データとして取得した．

3.4　継続事前学習による性能変化

表 3 に継続事前学習を実施した評価結果を示す．ベースモデルは Qwen3-4B（Base）であり，本研究モデルはベースモデルに対して前節までに構築したコーパスによる継続事前学習を実施したものである．本研究モデルは，Daikin QA を用いた評価で正答率が 0.45 から 0.49 へ向上した．一方で，汎用評価では指標ごとに改善と低下が混在し，平均正答率は 0.45 から 0.39 へ低下した．つまり，本設定における継続事前学習は空調分野の性能を向上させる一方で，汎用評価全体では性能が低下するというトレードオフが生じている．なお，汎用評価のうち HE-EN では性能が向上し，特に GPQA [8] における正答数が増加した．これは，空調関連文書に冷媒など化学・物理分野に関する記述が一定量含まれており，HE-EN に関連する知識の補強につながったためと考えられる．

Daikin QA に含まれる問題のうち，特に改善が目立ったのは空気調和や給排水などのカテゴリであり，関連する設問で正答率が上昇した．アノテーションの際に住宅の空気環境やビルの空調管理に関する記述が一定量含まれていることを確認しており，これらの記述が継続事前学習を通じて関連知識の想起を後押ししたと思われる．一方で低下が目立ったのは電気基礎，冷凍の原理，関連法規のカテゴリであり，これらの領域では，学習データ中の記述範囲や表現の偏りにより必要知識が十分に補強されなかった，あるいは継続事前学習によってベースモデルが持つ汎用的な知識・推論の一部が相対的に弱まった可能性がある．

以上より，Daikin QA では正答率の改善が確認できた一方で，一部カテゴリでは性能低下が生じることが示された．今後は，学習に利用するコーパスの種類と混合比を検討し，学習条件も調整しながら比較検証を行い，汎用的な知識と推論能力を維持しつつ，空調分野に特化した性能を改善したモデルの構築を目指す．

4　おわりに

本論文では，FineWeb2-ja から空調ドメインのテキストを抽出し，約 0.16B tokens からなる高品質な空調 Web コーパスを構築した．構築にあたっては，軽量な分類器の適用と LLM による文書スコアリングを反復することで，高品質データを効率的に収集できることを確認した．Qwen3-4B（Base）に対して継続事前学習を実施した結果，空調分野に特化した QA 評価における正答率が 4 ポイント向上した．

今後の課題として，まず評価の拡充が挙げられる．本研究では QA 形式のみで評価を行ったが，実運用を想定したダウンストリームタスクでの検証も必要である．また，本手法の汎用性を確認するため，FineWeb2-ja 以外の Web コーパスにも分類器を適用し，空調分野のデータを同様に収集できるかを検証する予定である．最終的には，本コーパスに社内で収集した空調ドメインデータも加えた継続事前学習により空調分野に特化した高精度な LLM の開発を進めていく．

参考文献

[1] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance. **arXiv [cs.LG]**, March 2023.

[2] Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. Efficient

continual pre-training for building domain specific large language models. **arXiv [cs.CL]**, November 2023.

[3] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need. **arXiv [cs.CL]**, June 2023.

[4] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y K Li, Y Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. **arXiv [cs.CL]**, February 2024.

[5] Fan Zhou, Zengzhi Wang, Nikhil Ranjan, Zhoujun Cheng, Liping Tang, Guowei He, Zhengzhong Liu, and Eric P Xing. MegaMath: Pushing the limits of open math corpora. **arXiv [cs.CL]**, April 2025.

[6] Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. **arXiv [cs.CL]**, June 2024.

[7] Namgi Han, Nobutaka Ueda, Masatoshi Otake, Satoru Katsumata, Keisuke Kamata, Hiroshi Kiyomaru, Takashi Kodama, Yusuke Murawaki, Hiroshi Matsuda, and Bowen Chen. Llm-jp-eval: 日本語大規模言語モデルの自動評価ツール, 2024.

[8] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A graduate-level google-proof Q&A benchmark. **arXiv [cs.AI]**, November 2023.