

動的な Few-shot の事例選択を用いた商品属性情報の正規化

岡崎真治 伊東賢二 浅野孝平 稲田和明 張信鵬

株式会社 MonotaRO

{shinji.okazaki, kenji.ito, kohei.asano, kazuaki.inada, xinpeng_zhang}@monotaro.com

概要

E コマースにおいて商品の特性を表す属性情報は、ユーザーの購買判断における重要な情報であり、これらのデータの整備はユーザー体験向上のための重要な課題である。しかし未整備の属性情報には、実際には同一とみなせる表現が異なる表記で扱われていることが多く、これらを統一するためには高い精度での表現の正規化が必要となる。そこで本研究では、蓄積された人手による目視検証と修正を加えた整備済みの属性情報から、動的に Few-shot の事例選択した大規模言語モデルによる属性値正規化システムを提案し、実際の E コマースの商品情報から抽出した属性値を用いて評価と分析を行う。

1 はじめに

E コマースにおいて、商品の特徴を示す属性情報は商品購入の意思決定に重要な役割を担っている。特に専門的な商品購入には、特定の産業基準や用途への属性値の適合が求められる。たとえば建設業における「ねじ」の購入では、太さや長さなどの基本的な条件に加え、特定の環境条件に耐える材質、高い耐久性を示す強度等級などが参照される。

整備された属性情報は、ユーザーの商品選定や、商品検索・推薦への応用に有効で、E コマースでのユーザー体験の向上につながる [1]。しかし、実際の E コマースでは扱う商品点数が膨大なため、それらの商品の属性情報の整備には多くの人手の作業が必要である。たとえば、ねじカテゴリの商品がもつ「材質」の1つに「ステンレス鋼」が存在するが、属性値の表現のバリエーションとして「ステンレス」や「ステンレス製」などの省略や表記揺れの他に、「SUS304」などの専門用語を元にした短縮系が存在する。さらに、「SUS304」と「SUS430」は成分の含有量に違いがあるため厳密には異なる材質であるが、一定の基準を満たす耐食性の高い金属という観点における使用用途としてはより上位の概念でまと

めることが許容される場合がある。このような異なる表記の同一とみなせる属性を、統一された表記に変換することを属性値の正規化と呼ぶ。属性値の正規化が不十分な場合、同一とみなされる属性情報が異なる表現として扱われていることとなり、商品の比較や選定の妨げが起き、商品購入の機会損失につながる。そのため、E コマース上でのユーザー体験の向上のためには、高い精度での属性値の正規化が必要といえる。

本研究では、大規模言語モデル (Large Language Model: LLM) を活用し、Few-shot 事例を人手で作成したデータから動的に選択することで、効率的かつ効果的な属性値の正規化手法を提案する。はじめに、正規化対象の入力表現と類似した事例を動的に選択し、それらを Few-shot の一部としてプロンプトに組み込むことで LLM による正規化を実現する。また、LLM が出力した正規化結果に対して人手によるチェックと修正を加え、修正結果を Few-shot 事例の候補として利用する。評価では、提案手法における Few-shot 事例選択時の整備済み正規化データの使用方法を2通り設定し、ベースラインである静的な Few-shot と比較して、定量・定性的に分析する。

2 関連研究

属性値の正規化のように、大量のデータを分析やエンドユーザーに使いやすい形に整備・構造化することを総称して、Data wrangling と呼ぶ [2]。近年の LLM の発達に伴い、Data wrangling への LLM の活用が行われている [3, 4]。たとえば、Brinkmann ら [5, 6] は、LLM を活用して商品名と説明文から属性値の抽出と正規化を行う手法を提案している。[5] では、属性値抽出において抽出対象の属性値と意味的に類似する事例を動的に変更することで、高い精度での属性値抽出を実現した。[6] では、事例を用いて文脈内学習を行う Few-shot プロンプティング [7, 8] を用いて、WDC LSPM [9] データセット内の5つのカテゴリにおける属性値の正規化で、LLM 自

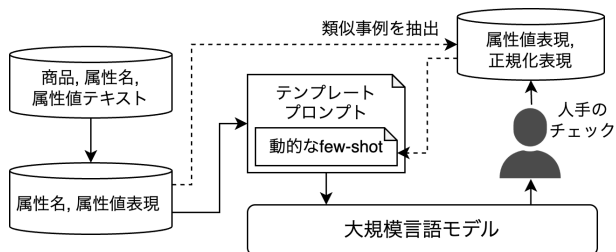


図 1 提案手法の概要

体を Fine-tuning したものと同等の性能を達成している。しかし、オープンソースのデータを用いた評価にとどまっており、E コマースなどの実データを用いた効果検証はなされていない。また、Data wrangling のアプローチの一つに人間がデータ修正プロセスに加わる Human-in-the-loop の研究もなされている [10, 11]。人手のチェックを入れて半自動化することで、完全自動では難しいデータ整備を効率的に実行できることが示唆されている [11]。

3 提案手法

提案手法の概要を図 1 に示す。まず、前提となる商品と属性情報について説明する。一般に商品は複数の属性情報を持ち、属性情報は属性名と属性値表現の組で定義される。属性名はその属性情報が指し示す商品の特徴の項目名を表すもので、属性値表現は各属性名に対してその商品が実際に持つ特徴を表す。たとえば、ステンレスで作られたねじであれば、属性名の 1 つとして「材質」があり、「材質」の属性値表現に「ステンレス鋼」や「SUS304」などがある。このような属性名と属性値表現の組を、正規表現によるルールや LLM を用いて、事前に商品情報から取得しておく。

次に、本研究で提案する LLM を用いた属性値表現の正規化手法について述べる。表 1 に示すテンプレートプロンプトを用いて、LLM に入力された属性値表現の正規化された表現 formal と、人手での判断の手がかりとなる追加の情報である同義表現 synonym、上位概念 taxonomy の三組を LLM に出力させる。表 1 内の *attr_name*, *few_shot_examples*, *input* は、入力によって値が変わる変数である。*attr_name* には入力属性値表現に対応する属性名、*few_shot_examples* には表 2 に示す例のような Few-shot 用の複数の事例、*input* には入力となる正規化対象の属性値表現、その属性値表現の抽出元の商品名をランダムに 1 つ選択したものと商品のカテゴリ名が与えられる。このようなプロンプトと Few-shot

表 1 属性値正規化のテンプレートプロンプト

あなたは、E コマースドメインでの商品の属性値の専門家です。
属性名: *attr_name*
{ "category": "", "product": "", "attr_value": "" } のフォーマットで入力されます。
入力と例を多角的に分析して、主体となる *attr_value* の以下の内容を出力してください。
主体となる属性値とは、“耐熱塩ビ”は“塩ビ”が主体で、“耐熱”は特性です。

出力の内容

- formal: 正規化名称

- synonym: 同義語・別名・略称など (多様に 3 つまで)

- taxonomy: 属性値に対して適切な一貫性のある階層的分類 (1 ~ 4 階層) を提供してください。出力形式: "*attr_name* > 分類 1 > 分類 2 > 属性値"

taxonomy の注意

- *attr_name* の特性や用途を考慮した分類階層を提供してください。

- 左から右へ一般的な分類から具体的な分類へととなるような分類階層「>」で区切り、全体的に一貫性のある重複のない分類名で階層を表現してください。

- E コマースのユーザにわかりやすいシンプルな分類名にしてください。

出力のフォーマット

- JSON 形式 { "formal": "", "taxonomy": "", "synonym": [] }

例

few_shot_examples

入力

input

出力

表 2 Few-shot プロンプトの例

属性名: 材質
input: { "attr_value": "PF 樹脂" }
output: {
 "formal": "フェノール樹脂",
 "taxonomy": "材質>樹脂>熱硬化性樹脂>耐熱性 PF",
 "synonym": ["フェノール", "ベークライト"] }

事例を用いることで、LLM から formal, synonym, taxonomy の三組が出力されることが期待される。得られた出力のうち formal が正規化済みの属性値表現として適切かを人手で確認した後、出力の三組を保持しておく。

LLM を用いた正規化では、Few-shot の事例を適切に選択することで、高精度化が期待できる [5]。本研究では静的な事例に加え、人手によるチェックが済んだ正規化事例の中から、入力の属性値表現と意味的に類似する事例を動的に選択して採用する。類似度は、正規化対象の属性情報に対応する、属性名、属性値表現と商品のカテゴリ名を連結した文字列と、正解事例として保存している出力の三組 (formal, taxonomy, synonym) を連結した文字列を、そ

れぞれ Multilingual-E5-large [12] でベクトル化し、それらの間のコサイン類似度で計算する。事例選択は、類似度が高い事例から順に貪欲に選択する。

4 評価実験

4.1 評価データ

まず、本研究で提案した正規化手法の評価のための入力となる属性値表現を準備する。評価データの抽出元として、実際の E コマースサービスであるモノタロウ¹⁾ から、専門用語が多く人手による正規化のコストが高いと想定される「科学研究・開発用品/クリーンルーム用品」のカテゴリに含まれる商品を採用した。ただし、モノタロウの商品情報に含まれる生の属性値テキストには、1つの属性名に対して複数の属性値表現が混在している場合がある。たとえば{属性名: “材質”, 属性値テキスト: “本体/硼珪酸ガラス, キャップ/メラミン (オートクレーブ可)”}のような事例が存在する。そこで、属性値表現の正規化を行いやすくするために、セミコロン、読点、括弧などのテキストの構造に着目したルールと、Few-shot プロンプティングを活用した大規模言語モデル [5] を組み合わせることで、{属性名: “材質”, 属性値表現: “硼珪酸ガラス”}のように属性値テキストから属性名と属性値表現が一对一の対応が取れるように整理した。

4,056 件の生の属性値テキストに対して抽出を試みたところ、合計で 6,486 件の属性名と属性値表現の組を取得できた。評価のためにランダムに選択した 100 件の生の属性値テキストから、合計で 162 件の属性値表現が得られた。以降の属性値正規化では、この 162 件の属性名と属性値表現の組を評価対象として使用した。

4.2 実験設定

本研究の評価では、動的な Few-shot の候補選択に制約をかけた 2 つの実験設定を設けた。設定 1 は、動的な Few-shot の事例を検索する際に、入力の属性値表現と一致するエントリを除外した上で類似事例を問い合わせる設定であり、明らかな正解データが Few-shot の事例に含まれないようにした。設定 2 では、入力の属性値表現の正規化先と一致するすべてのエントリを除外したうえで類似事例を問い合わせる設定である。この設定は、人手でのチェックで

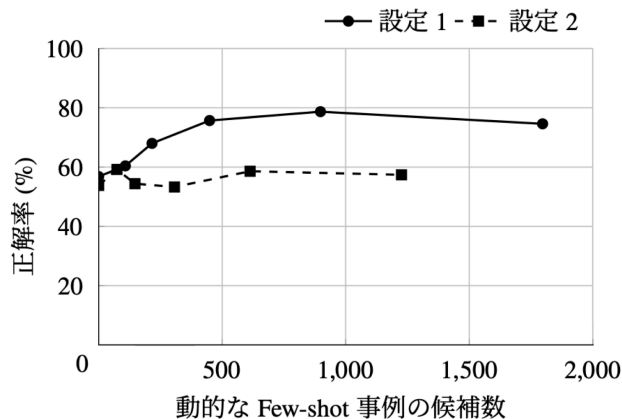


図2 属性値正規化の評価結果

扱ったことのない正規化先を持つ事例に対して、意図した正規化が行えるかを評価するためであり、難しい問題設定であると言える。設定 2 で効果が確認できれば、人手のチェックを通して作成したデータが、今回採用したカテゴリ以外に対する応用が期待できる。

さらにそれぞれの設定で、動的な Few-shot の作成時の問い合わせ先である人手でのチェック済みの事例数のデータ量を変化させることで、人手の介入がどの程度必要かを確認した。なお、設定によって候補事例の除外の仕方が異なるため、データ変化量を一致させていない点に注意されたい。

また LLM が出力した正規化された属性値表現のチェックは、2 人の作業員で実施した。LLM から出力された正規化先の表現である formal に対して、それぞれの作業員がチェックを行い、LLM の出力が正規化先として不適切と判断した場合、LLM の formal を適切な表現に変更した上で動的な Few-shot の選択候補となる正規化済みの事例として保存した。

LLM として Gemini 1.5 Pro [13] を、温度パラメータ $t = 0.0$ 、その他のパラメータは既定値として使用した。表 1 の *few_shot_examples* として、動的に選択した事例 3 件、全ての評価対象で共通する静的な事例を 2 件の合計 5 件を Few-shot の事例として用いた。なおベースラインとして、Few-shot の事例を動的に選択せず、5 件ともすべて静的に作成したものをそれぞれの設定で用意した²⁾。

4.3 評価結果

属性値正規化の実験結果を図 2 に示す。図 2 の設定 1 の結果より、入力の属性値表現の正規化先が動

1) <https://www.monotaro.com/>

2) 図 2 の動的な Few-shot 事例の候補数 = 0 がベースラインを示す

表 3 正規化結果の出力と動的な Few-shot 事例の例

	実験設定	入力	正解の正規化	LLM の出力	動的な Few-shot の例
例 A	ベースライン (設定 1)	安全装置	温度過昇防止	温度ヒューズ	-
例 B	提案手法 (設定 1)	安全装置	温度過昇防止	温度過昇防止	属性名: 安全装置 input: {"attr_value": "内部温度ヒューズ"} output: { "formal": "温度過昇防止", "taxonomy": "安全装置 > 温度制御 > 過昇温防止>ヒューズ > 内部温度ヒューズ", "synonyms": ["内部温度ヒューズ"] }
例 C	提案手法 (設定 2)	滅菌	未滅菌	滅菌無し	属性名: シール方式 input: {"attr_value": "シールなし"} output: { "formal": "なし", "taxonomy": "シール方式 > シールなし", "synonyms": ["シールなし"] }
例 D	提案手法 (設定 2)	黄色	イエロー	イエロー	属性名: 色 input: {"attr_value": "青"} output: { "formal": "ブルー", "taxonomy": "色 (本体)", "synonyms": [] }

動的な Few-shot の選択先に存在していれば、静的な Few-shot のみを用いる場合と比較して高い正解率が得られているとわかる。たとえば表 3 の例 A と例 B の事例では、共に「安全装置」という属性値表現を入力として与えているが、例 A のベースラインで誤り、例 B の提案手法（設定 1）で正解を得られていることが確認できる。与えられた動的な Few-shot の事例を見ると、似たような事例を選択されており、この類似事例を参照して目的の正規化表現が得られていると考えられる。

ただし、動的な Few-shot の選択先を多数用意しても、提案手法の設定 1 では約 500 件ほどで約 75% の正解率に飽和した。この要因の一つとして類似事例の獲得に問題が挙げられる。たとえば表 3 の例 C では、入力である「滅菌無」のデータに対して、そもそも類似する事例が存在せず、有用な動的な Few-shot の事例を獲得することができなかったため、目的の正規化表現が得られなかったと考えられる。

また図 2 の設定 2 では、動的な Few-shot のための類似事例の候補を増やしても、正解率は改善しなかった。この結果より、LLM を活用する場合であっても、開発者側がどのように正規化したいかについて、明示的に LLM に正解の事例を与える必要性が高いと考えられる。たとえば図 2 の例 D は、設定 2 で意図した正規化結果を得られた事例である。「青」から「ブルー」という正規化表現を類似事例として選択したことで、英語カタカナ表記が正規化後の表現であるというパターンを LLM が認識できたのではないかと考えられる。よって、入出力のテキストの意味的な近さだけでなく、正規化の変換のパターンが似た事例を Few-shot として選択したり、規則的

な正規化先の表現を設定したりすることが必要であると考えられる。

5 おわりに

本研究では、LLM の Few-shot プロンプティングにおいて、入力と類似する事例を動的に選択する属性値の正規化手法を提案し、実際の E コマースのデータで提案手法の評価と分析を実施した。評価と分析の結果、動的な Few-shot として入力した属性値表現の正規化先となる事例を取得して LLM のプロンプトに活用することで、指定した正規化表現を高い精度で獲得していることが確認できた。また、具体的な正規化先となる事例が動的な選択の候補にならない場合は、意図した正規化表現を獲得できず、静的に Few-shot 事例を選択したベースラインからの正解率の改善はみられなかった。一方で人手の分析から、属性値表現から正規化表現への変換として類似するパターンを持つものを Few-shot の事例として活用することで、意図した正規化が実現できる可能性があることがわかった。

今後の課題として、具体的な正規化先となる事例を事前に与えない状態での正規化の正解率の改善と、得られた正規化表現を実際の E コマースのサービスで活用を実現していきたい。

参考文献

- [1] Damir Vandić, Jan-Willem Van Dam, and Flavius Frasin. Faceted product search powered by the semantic web. **Decision Support Systems**, Vol. 53, No. 3, pp. 425–437, 2012.
- [2] Ignacio G Terrizzano, Peter M Schwarz, Mary Roth, and John E Colino. Data wrangling: The challenging journey from the wild to the lake. In **CIDR**. Asilomar, 2015.
- [3] Gonzalo Jaimovitch-López, Cèsar Ferri, José Hernández-Orallo, Fernando Martínez-Plumed, and María José Ramírez-Quintana. Can language models automate data wrangling? **Machine Learning**, Vol. 112, No. 6, pp. 2053–2082, 2023.
- [4] Xue Li and Till Döhmen. Towards efficient data wrangling with llms using code generation. In **Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning**, pp. 62–66, 2024.
- [5] Alexander Brinkmann, Roeë Shraga, and Christian Bizer. Extractgpt: Exploring the potential of large language models for product attribute value extraction, 2024.
- [6] Alexander Brinkmann, Nick Baumann, and Christian Bizer. **Using LLMs for the Extraction and Normalization of Product Attribute Values**, pp. 217–230. Springer Nature Switzerland, 2024.
- [7] 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 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.
- [8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.
- [9] Anna Primpeli, Ralph Peeters, and Christian Bizer. The wdc training dataset and gold standard for large-scale product matching. In **Companion Proceedings of The 2019 World Wide Web Conference, WWW '19**, pp. 381–386. Association for Computing Machinery, 2019.
- [10] Chengliang Chai and Guoliang Li. Human-in-the-loop techniques in machine learning. **IEEE Data Eng. Bull.**, Vol. 43, No. 3, pp. 37–52, 2020.
- [11] Tomas Petricek, Gerrit JJ van Den Burg, Alfredo Nazábal, Taha Ceritli, Ernesto Jiménez-Ruiz, and Christopher KI Williams. Ai assistants: A framework for semi-automated data wrangling. **IEEE Transactions on Knowledge and Data Engineering**, Vol. 35, No. 9, pp. 9295–9306, 2022.
- [12] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. **arXiv preprint arXiv:2402.05672**, 2024.
- [13] Google Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. **arXiv preprint arXiv:2403.05530**, 2024.