大規模言語モデルを用いたタグ付けによるデータの品質向上

草野 元紀 [1]
[1] 日本電気株式会社
g-kusano@nec.com

概要

データ分析における結果の質を向上させるためには、入力データの質を担保することが重要である。本研究では、データの品質を向上させる目的で、各データ項目にタグを付けることを考える。タグ付けを行うことにより、複数のデータを横断的に扱うデータ分析が容易になり、情報検索の高速化に貢献する。従来手法では、何らかの外部ソースから関連データを抽出し、入力データと結合することで情報拡張が行われていたが、すべてのデータで望ましい外部データを得られるとは限らない。一方で、昨今の大規模言語モデル (LLM) の進展により、関連データが存在しなくても、LLM が含む知識を基に教師なしでデータの特性を予測することが可能になっている。

本論文では、LLM を用いたタグ付けシステム **GA-Tag** (Generated and Aggregated Tag) を紹介する。プロンプトエンジニアリングにより LLM を活用して一つのデータ項目にタグを付与することは可能であるが、大量のデータ項目にタグを付ける際には、後段のデータ処理を考慮し、管理しやすい形式に整理する必要がある。GA-Tag では、LLM によってタグを生成 (generated) するが、工夫をしないとタグの総数が爆発的に増えるためデータ管理を目的としてタグを集約 (aggregated) する。実データを用いた検証では、教師なしにタグ付けが行えることと、生成されるタグを用いて比較分析ができるようになること、タグの総数を抑えらていることを紹介する。

1 Introduction

“Garbage in, Garbage out” という表現が知られているように、入力データの品質が不十分であれば、いかに高度なデータ分析方法を用いても、結果として得られる出力は不十分になる。この現象を避けるため、外部データソースから関連データを検索し統合することで、入力データの品質を上げる方法 [1] が

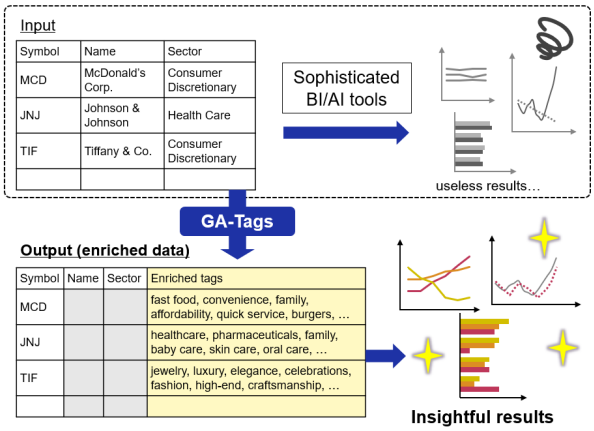

図 1 GA-Tag の入出力例。

提案されている。しかし、常に望ましい情報が存在するとは限らない。例えば、企業に関するデータ分析をする際に、各企業の利用者の印象に関するデータを統合させようにも、外部ソースがすべての会社に対して十分な情報を含んでいるわけではない。データソースに答えとなるような情報が含まれていなくても、大規模言語モデル (Large Language Model; LLM) を活用することで、その中に含まれる知識から様々な関係性を推測することが可能になっている [2, 3, 4]。

本研究では、データ高品質化の手法としてのタグ付けに注目する。タグは、各データ項目を人目で内容を把握するのに役立ち、関連情報検索の高速化に貢献する方法であり、ソーシャルメディアやニュースウェブサイトで一般的に使用されている。図 1 は、各会社にタグが割り当てられ、会社のテーブルにタグの列が新たに追加されるシナリオを示している。タグ付きのテーブルが得られると、タグなしの入力データでは実現できなかった分析が可能になる。

本研究ではタグ付けを行う方法として、**GA-Tag** (Generated and Aggregated Tag) を提案する。GA-Tag はまず、LLM を活用してて各データ項目にタグを生成する。GA-Tag は LLM の知見を活用するため、

追加のデータ収集や学習が不要であるため、ゼロショットに任意のデータに対してタグを生成することが可能である。しかし、何も工夫をしないと多種多様なタグが各データごとに大量に生成される。これはデータ管理の点や、分析の際にスパースなデータを扱うことになるなどの課題がある。そこで、意味が似ているタグは一つの代表的なタグに置き換える集約処理を施すことで、タグの品質は保ったままタグのユニーク数を減らす。これにより、後段のデータ分析に貢献するタグを付与し、データの品質向上に繋がげられる。本研究は [5] にて発表予定である。

2 GA-Tag

この章では、GA-Tag の構成について説明する。このシステムは、生成と集約のプロセスから成る。

2.1 生成タグ

$R = \{r_i\}_{i=1}^{n}$ を n 行 m 列のテーブルとし、各行 $r_i = \{(a_j, v_{ij})\}_{j=1}^{m}$ において、a_j は R の j 番目の列名を表し、v_{ij} は i 行目と j 列目の要素を示すとする。本研究の文脈では、テーブルの各行はタグ付け対象のデータ項目として扱う。図 1 の例の場合、行 $r = \{(\text{Name}, \text{Nestlé.}), (\text{Sector}, \text{Consumer Staples})\}$ が GA-Tag の入力になる。データ r へのタグ付けに関して ChatGPT[1] を利用する。プロンプトは図 2 になる。

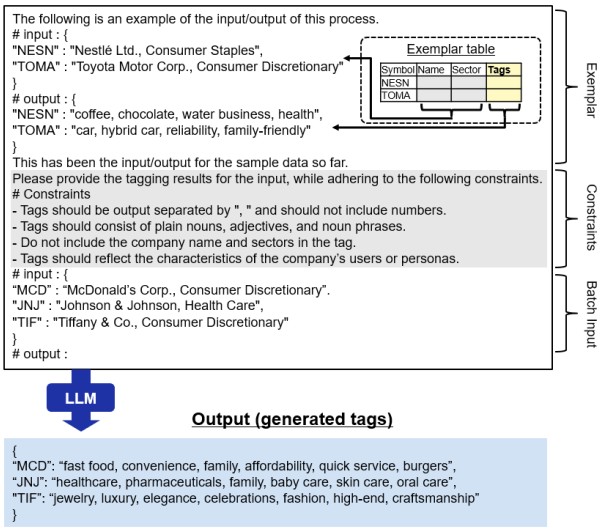

図 2 タグ付けのためのプロンプト。

タグ生成のためのプロンプトは、以下に挙げる三つの工夫を取り入れている。[Exemplar] 少数のサンプルを与えることで、意図した方向にタグが生成されるように導く。図 2 では、望ましいタグを “Tags” 列に追加した事例テーブルを準備し、プロンプトに読み込む様子を表している。[Constraints] 出力形式やタグの要望を含む制約を LLM に課す。図 2 の例では、上二つの制約は汎用的な条件で、下二つは企業分析という取り扱うテーマに特化したものになっている。[Batch Input] タグ付けを一つのデータ項目だけでなく、複数の行を一気に入れることで、同時に処理できるようにする。

図 2 の出力に見られるように、プロンプトに追加した制約の効果もあり生成されたタグは簡潔な単語で表現されており、たとえその企業のことを知らなくても企業に付与されたタグを見ることで企業の特徴を一目で理解するのが容易になっている。しかし、LLM によって生成されたタグは多種多様な表現をしており、異なるデータが共通のタグを共有することはほとんどなく、例えば SQL による集計や同じタグを持つデータの検索などの分析が困難になる。次の章では、類似するタグを一つの代表的な単語に置き換えることでこの課題の対処に取り組む。

2.2 集約タグ

この章では、類似したタグを代表単語に置き換えるタグ集約の方法を紹介する。例えば “apple”, “orange”, “grape” といったタグは “fruit” という代表的な単語に置き換え可能である。本研究では、LLM が生成したタグは**生成タグ**と呼び、その生成タグを代表単語に置き換えたものを**集約タグ**と呼ぶことにする。与えられた生成タグの集合から集約タを得るために、以下の二段階の処理を実行する：

1. タグをクラスタリングし、意味的に類似したタグをグループ化する。
2. 各クラスターを代表するような単語を割り当てる。

2.2.1 クラスタリング

意味的に類似した単語をグループ化する一般的な方法の一つは、それらの埋め込みベクトルを生成し、それらに対しクラスタリングアルゴリズムを適用することである。ここでは、タグを RoBERTa[6] を用いて埋め込みベクトルに変換し、これらのベクトルに対して凝集型階層的クラスタリング[2] を使用してグループ化する、[7] で紹介されたアプローチ

1) 本研究では、LLM として gpt-3.5-turbo-0613 を使用した。

2) https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

を採用する。実際には、BERT のような埋め込みベクトルは高次元 (d = 100 ~ 10000) であり、高次元データには球面集中現象などの “次元の呪い” が影響を及ぼす可能性がある [8]。この現象を軽減するために、[9] で提案されている埋め込みベクトルから平均ベクトルを差し引く方法を採用する。

タグの埋め込みに関しては、タグのジャンル g を指定し、タグ t を “This $\{g\}$ has a tag of $\{t\}$” というテキストに変換して埋め込む。例えば、t =“bank” の場合、その単語自体は金融機関の意味での銀行に加えて、川岸の土手という意味を持つなど、様々な解釈が持たれてしまう可能性がある。g =“company” と設定することで、タグの埋め込みベクトルは銀行の意味を優先することが期待される。

2.2.2 代表単語生成

各クラスターを代表的な単語で端的に表現するために、図 3 に示されるプロンプトを LLM に入力する。生成タグと同様に、サンプルを提示し、制約を課し、代表単語生成をバッチで処理する。

```
The following is an example of the input/output of this process.
# input : {
"sample_0" : "apple, mandarin orange, grape",
"sample_1" : "Toyota, Suzuki, BMW "
}
# output : {
"sample_0" : "fruit",
"sample_1" : "car"
}
This has been the input/output for the sample data so far.
Please generate words that represent the cluster while adhering to the following constraints.
# Constraints
- The output should primarily consist of 1 token, with a maximum of 2 tokens.
- The output should describe one of plain nouns, adjectives, and noun phrases.
- Please generate representative words that will be useful for classifying the company.
# input : {
"0" : "express transportation, medical transportation, package delivery, rail transportation",
"1" : "retail, retail branding, retail credit, retail properties, retail space, retail spaces",
"2" : "long-term care, long-term care insurance"
}
# output :
```

Exemplar / Constraints / Batch Input

図 3 クラスターに代表的なラベルを割り当てるためのプロンプト。

2.3 システムの出力

タグ集約処理が完了すると、各生成タグ t に対応する集約タグ a が割り当てられる。図 4 に示されるように、各データに付与された生成タグとそれをまとめ上げた集約タグを用いて、GA-Tag は入力データを拡張したテーブルに変換して出力する。出力されるテーブルは、データ分析やデータ保存などの観点で、いわゆる縦持ち (unpivot) や横持ち (pivot) 形式を選ぶことが出来る。

3 Demonstration

この章では、GA-Tag が拡張したデータの中身とその性質を、S&P 500 に選ばれた 500 社の企業を含む株式データセット [3] を用いてデモンストレーションする。このデータセットは、図 1 に示されるように、“Name” と “Sector” という列を持つテーブルで構成されている。

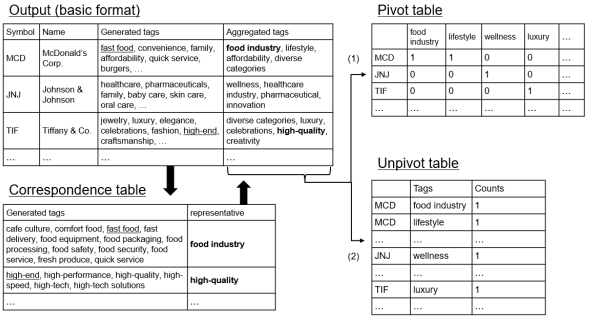

図 4 GA-Tag 出力の基本フォーマット (左上) とテーブルの各種表現 (右)。生成タグは対応表 (左下) を介して集約タグに変換される。

3.1 Scenario

ここでは、タグ付けにより拡張された企業データの活用に関するユースケースを紹介する。

あるデータサイエンティストが特定の企業、Marriott という企業を調査することになったとする。そのデータサイエンティストは、マリオットと他の企業と比較するために S&P 500 データセットの分析することにした。当初、元の入力データだけで分析しようとすると、そのテーブルには企業名とセクターの列しか含まれておらず、マリオットと同じセクターは “Consumer Discretionary” であるがこれに属する企業は 80 社以上あった。この情報量だけで、同じセクターに属する 80 社全部に対して競合分析することは対象が多すぎて有益な結果を得られなかった。これに対処するため、GA-Tag を使用して各企業にタグを割り当て、企業の詳細を分析できるようにした。その結果、マリオットには “hospitality” と “luxury” がタグ付けされており、この二つのタグを持つ企業を調査したところ、図 5 に示されるように 4 社に絞り込まれた。その中で 3 社は高級ホテルセクターで運営されており、これらはマリオットの競合企業と見なされる。このように、GA-Tag は競合分析の効率化に貢献する。

さらに、これらのタグは、マリオットと競合企業との比較分析に役立つ。企業名を行に、集約タグを列にリストしたピボットテーブル (図 6) に示されるように、すべての競争相手には “luxury”,

3) https://github.com/datasets/s-and-p-500-companies/blob/main/data/constituents.csv

Symbol	Name	Generated tags	Aggregated tags
MAR	Marriott Int'l.	hotels, travel, hospitality, luxury, business travel, family vacation	tourism industry, service industries, **hospitality**, **luxury**, business and travel, family activities
RCL	Royal Caribbean Cruises Ltd	cruise, travel, luxury, entertainment, vacation, relaxation, adventure, hospitality	tourism industry, service industries, **luxury**, vacation, lifestyle, **hospitality**
HLT	Hilton Worldwide Holdings Inc	hospitality, travel, luxury, comfort, service, vacation, business travel	service industries, **hospitality**, **luxury**, diverse industries, vacation, business and travel
MGM	MGM Resorts International	casino, resorts, entertainment, luxury, travel, hospitality	kitchen items, tourism industry, **luxury**, service industries, **hospitality**
HST	Host Hotels & Resorts	hospitality, hotels, travel, leisure, customer service, comfort, luxury, business travel, vacation, hospitality industry	**hospitality**, tourism industry, service industries, leisure activities, customer relations, diverse industries, **luxury**, business and travel, vacation

図 5　集約タグに “hospitality” と “luxury” のタグを持つ企業。

“hospitality”, “service industries” というタグが割り当てられている。マリオットに特有のタグとして “family activities” があり、それが同社にとって潜在的な競争優位性を示唆している。一方で、マリオットに “vacation” タグがないことは、休暇をターゲットにした顧客セグメントに施策改善の余地があること示している。

	Name	hospitality	luxury	service industries	tourism industry	business and travel	vacation	diverse industries	customer relations	family activities	kitchen items	leisure activities	lifestyle
MAR	Marriott Int'l.	1	1	1	1	1	0	0	0	1	0	0	0
RCL	Royal Caribbean Cruises Ltd	1	1	1	1	0	1	0	0	0	0	0	1
HLT	Hilton Worldwide Holdings Inc	1	1	1	0	1	1	1	0	0	0	0	0
MGM	MGM Resorts International	1	1	1	1	0	0	0	0	0	1	0	0
HST	Host Hotels & Resorts	1	1	1	1	1	1	1	1	0	0	1	0

図 6　図 5 の集約タグ列のピボットテーブル。

GA-Tag の適用先は企業データに限定されない。食品や日用品などの製品や、SaaS アプリケーションや保険商品などのサービスなど、さまざまなカテゴリーに拡張可能である。個人に対しても、年齢、性別、職業などの人口統計属性に基づいて行動傾向や好みに関するタグを割り当てることができ、顧客分析やマーケティングに活用可能である。

3.2 Performance

ここでは、GA-Tag に関連するいくつかの統計量を計算する。以下の各表に示される数値は、S&P 500 に対して GA-Tag によるタグ付けを独立して 3 回実行した結果の平均値と標準偏差である。

3.2.1 Tag

表 1 は、生成タグと集約タグのタグ数に関する統計を示している。ここで、記号 $\sharp\text{tags}, \sharp\text{tags} \geq 5, P(\sharp\text{tags} \geq 5)$ は、それぞれユニークなタグの総数、5 回以上出現するタグの数、そのような頻出タグの割合を表している。

表 1 に基づいて、集約タグのユニーク数は生成タグのユニーク数の約 7 分の 1 であることが観察された。一方で、5 社以上に関連する集約タグの数は、

	$\sharp$tags	$\sharp$tags > 5	$P(\sharp\text{tags} > 5)$
生成タグ	1564 ± 8.3	156 ± 3.3	10.0 ± 0.2 (%)
集約タグ	208 ± 8.0	125 ± 6.1	60.4 ± 2.2 (%)

表 1　タグの個数に関する統計量。

生成タグの数とほぼ同一である。その結果、5 社以上に関連するタグの割合は、生成タグの約 10% から集約タグの約 60% に大幅に増加している。これは、類似のタグを共有する企業を分析するには集約タグを用いる方が効果的であることを示している。

3.2.2 Costs

LLM を使用したサービスを開発するときに、その LLM の API コストも重要な関心事項である。API 使用料は、入力と出力テキストのトークン長によって主に決定される。表 2 には、入力ファイルバッチサイズが 20 の場合のプロンプト (図 2) のトークン長に関する統計を記載している。OpenAI 社の `gpt-3.5-turbo-0613`[4] の場合は、1000 入力トークンにつき 0.0015 USD(米ドル)、1000 出力トークンにつき 0.002 USD が単価である。500 社のタグ生成処理にかかる OpenAI API の総コストは 0.054 USD であり、タグ集約処理のコストは 0.016 USD であった。この数値は日本円で 10 円程度になる。

		生成タグ	集約タグ
トークン数	入力	14995	9832.0 ± 178.3
	出力	16198.0 ± 356.1	577.3 ± 23.2
費用 (USD)	入力	0.022 ± 0.001	0.015 ± 0.001
	出力	0.032 ± 0.001	0.001 ± 0.000
時間 (秒)		275.5 ± 11.3	69.0 ± 1.5

表 2　API コストと処理時間の統計量。

また、LLM の処理時間も計測した。表 2 によると、生成と集約プロセスにはそれぞれ 275 秒と 69 秒がかかった。生成と集約の両プロセスは並列処理が可能であるため、API を並列実行することで、並列数に比例して処理時間を短縮できる。

4 Conclusion

本報告では、LLM を活用して開発されたデータ拡張方法である **GA-Tag** を紹介した。GA-Tag はタグ生成とタグ集約の 2 つの機能からなり、入力データのみでは得られなかった知見を提供し、タグの総数をコントロールできることからデータ分析を効率的かつ有益なものに保つことが可能である。

4)　https://openai.com/pricing

参考文献

[1] Yuyang Dong and Masafumi Oyamada. Table enrichment system for machine learning. In **SIGIR**, pp. 3267–3271. ACM, 2022.

[2] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In **USENIX Security Symposium**, pp. 2633–2650. USENIX Association, 2021.

[3] Shotaro Ishihara. Training data extraction from pre-trained language models: A survey. In **Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)**, pp. 260–275. ACL, 2023.

[4] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. In **ICLR**. OpenReview.net, 2023.

[5] Genki Kusano. GA-Tag: Data enrichment with an automatic tagging system utilizing large language models. In **ICDE**, p. accepted, 2024.

[6] 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. **CoRR**, Vol. abs/1907.11692, , 2019.

[7] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. Classification and clustering of arguments with contextualized word embeddings. In **ACL (1)**, pp. 567–578. Association for Computational Linguistics, 2019.

[8] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Hubs in space: Popular nearest neighbors in high-dimensional data. **J. Mach. Learn. Res.**, Vol. 11, pp. 2487–2531, 2010.

[9] Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, Marco Saerens, and Kenji Fukumizu. Centering similarity measures to reduce hubs. In **EMNLP**, pp. 613–623. ACL, 2013.