大規模言語モデルを用いた生成による企業の業種体系の拡張

山岸駿秀　貞光九月
株式会社マネーフォワード
{yamagishi.hayahide,sadamitsu.kugatsu}@moneyforward.co.jp

概要

企業情報の分析のために，各企業の各事業に適切な業種を割り当てることは重要である．このとき，既存の業種体系に合致しない事業内容があるため，業種体系を適宜拡張して使う必要がある．本研究では，LLM を用いて既存の業種体系にない業種名称を生成し，業種体系を拡張する．ただし，生成された業種には不要な名称が含まれる．不要な名称を削減する手法として，LLM への生成制約，既存業種にある名称のフィルタリング，生成業種間で類似した名称のクラスタリングを提案し，効果を検証する．

1　はじめに

企業情報の分析では，各企業が行う事業を表す**業種**を事業別に付与し，業種ごとに集計を行うことが重要である．このために，各企業の各事業に対して適切な業種を推定する必要がある．

業種推定時に，事業に合致する業種が既存の業種体系にないことがある．この理由として，新規事業の発達の速度に体系の更新が追いついていないことが挙げられる．例えば主要な体系の日本標準産業分類 [1] は，1993 年版から 4 回しか改訂されていない．

事業に合致する業種がないときは，比較的関連する既存業種をしばしば強引に割り当てる．日本標準産業分類には区分の細かさが異なる 4 つの体系があり，最も細かい細分類では 1,473 種が定義されている．しかし，うち 598 種は製造業に属しており，情報通信業は 45 種しかない．例えば 45 種の中に「顧客の Web サイトの制作」を表す業種はない．45 種の中から “受託開発ソフトウェア業” などを割り当てることもできるが，事業内容を正確に反映できていない．また強引に割り当てた場合，分析者も “Web サイトの制作” が “受託開発ソフトウェア業” に紐づくことを理解する必要があり，分析業務が煩雑になる．このようなときは “Web サイト制作業” などの新規名称を作成すべきであると考える．

本研究では大規模言語モデル（LLM）を用いて既存の業種体系にない新規業種を生成し，生成結果を業種体系に追加して拡張する．ただし，生成結果には不要な名称も含まれると予期できる．不要な名称の体系への追加を防ぐため，LLM に複数の生成制約を与える方法，既存業種と近い業種を除去するフィルタリング，生成業種間で似た名称を統合するクラスタリングの 3 つを実施した．実験により，メイン事業とサブ事業を区別させる制約や，フィルタリングやクラスタリングが有効であることを示した．

2　関連研究

大規模な業種体系から適切な業種を選ぶタスクは，ラベルが数千種以上ある状況での多ラベル分類である，Extreme Multi-label Classification（XMC）とみなせる．例えば XMC の主要データセットである AmazonCat-13k[2] では，13,000 種の商品カテゴリから適切なものを推定する．従来は LSTM などによる分類的手法が提案されていた [3, 4]．しかしこのデータには “mic”, “mics”, “microphone” などラベル揺れとみなせる名称があり，分類的手法では揺れを解決できない課題があった．そこで Jung らは T5 [5] による生成的手法を検証し，生成により揺れを緩和できることを示した [6]．本研究で扱う業種体系に揺れはないが，生成した業種が互いに揺れとみなせる場合や，既存業種の別名とみなせる場合がある．これらの揺れは生成制約や後処理で低減する．

分類対象となる文書集合と既存のラベル体系を繰り返し LLM に与えてクラスタリングし，クラスタ境界を更新する手法も提案されている [7]．しかしこの手法ではクラスタ数を事前に定める必要があり，かつその数が 10 から 25 と少ない．本研究では，クラスタ数を動的に拡張する手法を提案する．

企業間の関係をグラフや埋め込み空間などで表現し，企業情報の分析を行う研究も提案されている [8, 9]．しかし，これらの研究は類似企業がない場合に事業内容を把握できない．本研究では各企業の各

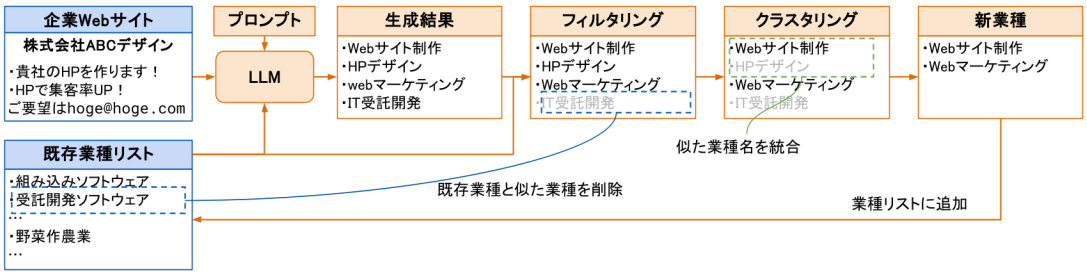

図 1　提案手法の概要

No.	Web ページのテキスト（抜粋，一部改変）	既存業種	生成業種	サブ業種
1	業務システムの開発は ABC コンサルティング	受託開発ソフトウェア業	–	–
2	介護タクシーで自宅での介護をサポート	訪問介護事業, 介護老人保健施設	介護タクシー	–
3	当農園ではぶどう狩りができます．12 月 1 日，園内カフェにて就活イベントを開催します	果樹作農業, 喫茶店	観光業, 屋外体験施設	就活支援業

表 1　データセットに含まれる事例の抜粋

事業に対して業種を生成するため，類似企業がなくとも事業内容が把握できる．

3　提案手法

3.1　業種名称の生成

図 1 に提案手法の流れを示す．企業の事業を示す文書をプロンプトとともに LLM に与え，業種名称を生成する．本研究では事業を示す文書として，企業の Web サイトのトップページから抽出したテキストを用いる．表 1 に，本研究で扱うデータセットを示す．生成された業種（生成業種）は新規業種として，既存の業種体系（既存業種）に追加する．なお，既存業種は日本標準産業分類の細分類とする．

単に「テキストの内容から業種名称を生成せよ」と命令すると生成業種に不要な業種名称が多く生成される懸念があるため，以下の生成制約を与える．

既存業種リストの照合　LLM には既存業種に含まれる業種名称を明示していないため，既存業種に似た名称を生成する可能性がある．そこで，生成時に既存業種リストを照合させ，「事業内容と適合する業種が既存業種にないときは新業種名称を生成せよ」と命令する．既存業種と似た業種の生成を抑制し，新規業種のみが得られることを期待する．

既存業種の選択　既存業種リストの照合を行うとき，より確実に LLM に照合させるために，事業内容と適合する業種が既存業種にあるときはその既存業種の名称を別の欄に出力させる．これにより，照合の効果をさらに高める効果を期待する．

サブ業種の生成　企業が行う事業はメイン事業とサブ事業に分けられるが，サブ事業の業種（サブ業種）は分析における重要度が低い．例えば表 1 の 3 例目の企業はぶどう狩り体験等がメイン事業であるが，不定期に就活イベントなどを開催する．このとき “就活支援業” がサブ業種となるが，これはイベント案内の更新のたびに変化するため，分析には適さない．そこで，「事業をメイン事業とサブ事業に分けて考え，メイン事業の業種を新規業種欄に，サブ事業の業種はサブ業種欄に生成せよ」と命令し，サブ業種が新規業種欄に含まれないようにする．

3.2　フィルタリング

生成制約を与えても既存業種と関連しない業種名称だけを確実に生成させることは難しい．そこで後処理として，既存業種を用いたフィルタリングを行う．GLuCoSE-base-ja [10] を用いて各業種名称の文埋め込みを作成し，各生成業種に対し全ての既存業種との cosine 類似度を測る．いずれかの既存業種との類似度が 0.7 以上となった生成業種を除去した．なお，フィルタリングの既存業種には日本標準産業分類にある 4 体系の全業種名称 2128 種を用いた．

3.3　クラスタリング

図 1 の “Web サイト制作” や “HP デザイン” などのように，生成業種内に同じ事業を指す業種が複数生成される場合がある．類似した生成業種をすべて業種体系に追加すると体系内で揺れが生じるため，追加前に生成業種内でクラスタリングを行い，類似名称を統合する．LLM に生成業種の集合と「同じ事業を表す業種名ごとにクラスタにせよ」というプロ

No.	既存業種の確認	既存業種の選択	サブ業種の区別	Precision	Recall	F1	のべ残存数	クラスタ種数
1				0.4681	**0.6667**	0.5500	94	35
2	✓			**0.7778**	0.2333	0.3590	18	11
3	✓	✓		0.6429	0.3103	0.4186	28	14
4			✓	0.5692	0.6379	**0.6016**	65	27
5	✓		✓	0.5455	0.3051	0.3913	33	15
6	✓	✓	✓	0.6000	0.4615	0.5217	50	18

表 2　クラスタリングの性能比較．✓ がついている制約を与えて生成した．太字は各評価指標での最大値．

				生成後				フィルタリング後			
No.	既存業種確認	既存業種選択	サブ区別	Prec.	Recall	F1	生成数	Prec.	Recall	F1	残存数
1				0.2420	**0.9620**	0.3868	314	0.4286	**0.8462**	0.5690	154
2	✓			0.4111	0.5211	0.4596	90	0.5323	0.4583	0.4925	62
3	✓	✓		0.4412	0.5921	0.5056	102	0.5132	0.5200	0.5166	76
4			✓	0.3405	0.8750	0.4903	185	**0.6064**	0.7917	**0.6867**	94
5	✓		✓	**0.4660**	0.6575	**0.5455**	103	0.5811	0.5811	0.5811	74
6	✓	✓	✓	0.3946	0.7532	0.5179	147	0.5521	0.6883	0.6127	96

表 3　生成後，フィルタリング後の段階での性能比較．

ンプトを与え，クラスタリングを実施した．得られたクラスタのうち，クラスタに含まれる業種がのべ 2 社以上に出現するものを残し，人手で統合後の名称を付与する．これを新業種として体系に追加する．

4　実験設定

4.1　データセット

本実験では，著者が用意したデータセットを用いて評価を行う．実例は表 1 に示した．このデータは，企業の公式サイトの情報をもとに業種情報を付与したデータである．事業内容がすべて既存業種にある企業を 50 社，生成業種が新たに必要な企業を 50 社用意し，前者は既存業種のみ，後者は両方の業種を付与した．既存業種は日本標準産業分類 [1] の細分類から選択し，生成業種は新たな名称を著者が作成した．この結果，既存業種は 100 社に 133 種・計 202 件（1 社あたり最大 5 種），生成業種は 50 社に 23 種・計 63 件（1 社あたり最大 3 種）付与された．

4.2　実験設定

実験では，前述の**既存業種リストの照合**，**既存業種の選択**，**サブ業種の生成**の 3 つの制約の効果を比較し，より正しく生成できる制約を調べる．業種生成やクラスタリングで用いる LLM は OpenAI の GPT-4o[1)] [11] とし，temperature を 0 にして 1 度だけ生成した．生成時には few-shot として，既存業種を持つ企業，既存業種と生成業種を持つ企業，既存業種・生成業種・サブ業種を持つ企業，の事例を 3 事例ずつ，計 9 事例を与える．生成制約を変える場合は制約に対応する情報のみを削除し，事例数は変えない．例えばサブ業種を別欄に生成させる実験では，全事例からサブ業種の情報のみを削除する．全ての制約を与えるプロンプトを付録 A に添付する．

評価では，正解と生成業種が意味的に一致している数を測定し，Precision，Recall，F 値を算出した．意味的な類似性は評価者 1 名の目視で判断した．1 企業に複数業種が生成された場合は，各業種名称に対して評価を行う．例えばある企業に “Web 制作業” と “ウェブ制作業” の 2 つが生成され，これらが “Web 制作業” クラスタとなった場合，クラスタリング前は名称が 2 つあるため 2 回評価し，クラスタリング後は 1 クラスタであるから 1 回評価する．クラスタリング前後で評価対象とする業種の数が異なるため，数値比較が行えないことに注意されたい．

5　実験結果

表 2 に，クラスタリング後の評価結果を示す．サブ業種の生成制約のみを与えた実験 4 が相対的に高い F 値を示した．この制約は新規業種欄に生成されるサブ業種を減らす．実験 2 と実験 5 を比較すると生成数は 13 増加したが，サブ業種の生成数は 26 から 17 に減少した．また，実験 1 では “企業研修”, “企業向け研修”, “農業研修” などが同一クラスタになったが，“農業研修” が生成された企業のテキストでは農業研修の情報は小さく，サブ業種とみなせる．実験 4 では “農業研修” がサブ業種欄に生成さ

1）　2024 年 12 月 2 日から 12 月 20 日の間に API 経由で利用．

No.	Web テキスト	生成業種の正解	生成例	サブ業種	フィルタリング後	クラスタリング後
1	Web 制作・ノベルティ制作は当社へ	グッズ製作業, ウェブ制作業	グッズ製作業, ウェブ制作業	DTP 制作業, Web 広告業	グッズ製作業, Web 制作業	グッズ製作業, Web 制作業
2	冷房シーズンの前にエアコン清掃	ハウスクリーニング業	清掃業	–	–	–
3	親子三代，歴史あるふぐ料理店	–	ふぐ料理店	–	ふぐ料理店	和食店

表 4　実験 4 の生成例．web ページのテキストは抜粋．クラスタ名は著者が付与．既存業種相当は下線を引いた．

れ，“企業研修業” クラスタが得られた．

その他の制約を与えた実験 5, 6 の F 値が実験 4 と比べて低い要因を調べる．表 3 にクラスタリング前の結果である生成後・フィルタリング後の評価結果を示す．既存業種を照合させる制約は生成数を減らす．実験 4 に比べて実験 5 では生成数が 82 減少したが，このうち 65 件は既存業種であった．一方で，生成すべき新規業種も 15 減少したことで，クラスタリング後の F 値に対する効果は小さくなった．

既存業種を選択させる制約では Precision の改善を期待していたが，実際は Recall が改善した．この制約を与えた実験 6 は実験 5 より生成数が 44 増加し，正解数も 10 増加した．一方で，既存業種やサブ業種の生成数も，それぞれ 35 から 55，17 から 28 に増加した．この中には，既存業種よりも細かく見える業種が複数存在した．例えば，実験 6 ではとんかつ料理店を営む企業に新規業種として “とんかつ店” を生成していた．しかし日本標準産業分類の定義文 [1] によればとんかつ料理店は “日本料理店” に含まれるため，既存業種である．LLM が既存業種より細かい業種名称を発見したとき，こうした包含関係を考慮できず既存業種にないとした可能性がある．

生成された既存業種はフィルタリングで除去できるため，全実験で F 値が改善する．改善幅は既存業種を照合させない実験 1 と 4 の方が大きい．例えば実験 1 や 4 では薬局に “薬局業” を生成するが，既存業種を照合させる実験では生成しなかった．既存業種を照合させる制約は生成時にフィルタリング相当の処理をしているとみなせるため，フィルタリング対象の既存業種が生成されにくかった．また，既存業種を選択させる制約で増加した生成業種のうち既存業種と似たものはフィルタリングで除去された．

以上から，既存業種に関する制約は有効ではなかった．既存業種の照合制約は既存業種だけでなく新規業種の生成も減らす．既存業種の選択制約は生成数を増やすが，その多くはサブ業種や既存業種とみなせる．サブ業種は不要であり，既存業種はフィルタリングで除去されるため，クラスタリングの結果は改善しない．一方サブ業種の制約はクラスタリングの妨げとなる名称を削減でき，有効であった．

クラスタリング後の性能が最大であった実験 4 の生成例をいくつか表 4 に示す．1 つ目は，新規業種が必要な企業に正しく生成できた事例である．Web テキストに小さく書かれていた事業をサブ事業と認識できた．また，“ウェブ制作業” は他の企業に生成された類似名称と集約して “Web 制作業” となった．

2 つ目は既存業種と似た名称を生成しフィルタリングで除去された事例である．LLM はテキストから「清掃」を抜き出して生成した．家庭の清掃に特化した業種は既存業種になく，新規業種とみなせるが，“清掃業” は産業廃棄物処理を示す “清掃事業所” と類似度が高いため除去された．この事例以外でも業種名称に見える文字列を抜き出す傾向があった．

3 つ目は，新規業種の生成が不要にも関わらず生成された事例である．“ふぐ料理店” は前述の “とんかつ店” と同様に “日本料理店” に含まれるが，包含関係を考慮できず既存業種とみなせなかった．一方クラスタリングでは LLM が “ふぐ料理店” や “和牛料理店” などを “和食店” にまとめており，包含関係に関する制約を与えれば改善する可能性がある．

6　おわりに

本研究では，大規模言語モデルを用いて既存の業種体系にない業種名称の生成を行い，業種体系を拡張する検証を実施した．不要な業種が体系に追加されることを防ぐため，いくつかの生成制約を与え，後処理として既存業種と近い名称を除去するフィルタリングと生成業種間で似た名称を統合するクラスタリングを実施した．検証により，フィルタリングやクラスタリングは有効であることを示した．また，生成制約は企業ごとにメイン事業とサブ事業を区別させる制約が有効であった．一方，生成された不要な業種が既存業種と包含関係にある場合，生成制約やフィルタリングで除去できず，クラスタリングの妨げとなる課題がある．今後は，包含関係を考慮させる制約やフィルタリング方法を検討したい．

参考文献

[1] 総務省政策統括官. 日本標準産業分類 (令和 5 年 7 月告示), 2023. https://www.soumu.go.jp/toukei_toukatsu/index/seido/sangyo/R05index.htm.

[2] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In **Proceedings of the 7th ACM Conference on Recommender Systems**, RecSys '13, p. 165–172, New York, NY, USA, 2013. Association for Computing Machinery.

[3] Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. AttentionXML: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 32. Curran Associates, Inc., 2019.

[4] Kunal Dahiya, Ananye Agarwal, Deepak Saini, K Gururaj, Jian Jiao, Amit Singh, Sumeet Agarwal, Purushottam Kar, and Manik Varma. Siamesexml: Siamese networks meet extreme classifiers with 100m labels. In **International conference on machine learning**, pp. 2330–2340. PMLR, 2021.

[5] 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. **J. Mach. Learn. Res.**, Vol. 21, No. 1, January 2020.

[6] Taehee Jung, Joo-kyung Kim, Sungjin Lee, and Dongyeop Kang. Cluster-guided label generation in extreme multi-label classification. In Andreas Vlachos and Isabelle Augenstein, editors, **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 1670–1685, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[7] Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W. White, Longqi Yang, Reid Andersen, Georg Buscher, Dhruv Joshi, and Nagu Rangan. Tnt-llm: Text mining at scale with large language models. In **Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining**, KDD '24, p. 5836–5847, New York, NY, USA, 2024. Association for Computing Machinery.

[8] Lele Cao, Vilhelm von Ehrenheim, Mark Granroth-Wilding, Richard Anselmo Stahl, Andrew McCornack, Armin Catovic, and Dhiana Deva Cavalcanti Rocha. CompanyKG: A large-scale heterogeneous graph for company similarity quantification. In **Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining**, KDD '24, p. 4816–4827, New York, NY, USA, 2024. Association for Computing Machinery.

[9] UZABASE. 言語モデル LUKE を経済の知識に特化させたモデル「UBKE-LUKE」の提案, 2024. https://tech.uzabase.com/entry/2024/12/24/173942.

[10] 福地成彦, 星野悠一郎, 渡邉陽太郎. 2 段階対照学習による日本語文埋め込みモデルの汎用性獲得. NLP 若手の会 (YANS) 第 18 回シンポジウム, 2023.

[11] OpenAI. GPT-4o system card, 2024. https://cdn.openai.com/gpt-4o-system-card.pdf.

A　プロンプト

生成プロンプト

指示
ある企業の web ページから抽出したテキストをもとに事業内容を理解し，この企業の業種を推定してください．

手順
推定時には，以下の手順に沿ってください．

1. テキストをよく読み，この企業の事業内容を理解する．複数の事業を展開している場合は，すべての事業内容を理解すること．テキストは “## web ページのテキスト” の下に添付する．
2. 各事業を主要事業とサブ事業に分けて整理しておく．主要事業とはテキスト中によく出てくる事業内容であり，この会社の売り上げの多くを生み出していると考えられる事業を指す．主要事業が複数ある場合もある．サブ事業は，主要事業でない事業のことである．
3. “# 業種リスト” に書かれた業種名称をよく読む．業種とは，各企業が提供する商品やサービスの違いによって定められる，事業内容の名称のことである．業種リストにある業種名称は，日本標準産業分類の業種細分類の名称を 1 行 1 業種ずつ示したものである．
4. 手順 1 で理解した主要事業やサブ事業の各事業内容を端的に表す業種名称を自分で考える．事業内容が複数ある場合は，それぞれについて業種名称を考える．業種名称相当の文字列をテキストから抜き出す必要はなく，1 で理解した事業内容をあなたの言葉で整理すること．
5. 手順 4 で考えた主要事業の業種名称について，業種リスト中の業種名称と文字列的な関連性と意味的な関連性の観点で比較し，業種リスト中の業種のいずれにも関連しない業種とどれかに関連する業種に分けておく．
6. 手順 5 で比較した結果を出力する．また，手順 4 で検討したサブ事業の業種名称を出力する．出力時は，以下のフォーマットに従うこと．
   (a) 出力時は，“## 出力” の下に JSON 形式で出力する．
   (b) 手順 5 で考えた「業種リスト中の業種のいずれにも関連しない業種」は，「新業種」フィールドに string の list 型で記載する．
   (c) 手順 5 で考えた「業種リスト中の業種のどれかに関連する業種」は，「既存業種」フィールドに string の list 型で記載する．
   (d) 手順 4 で考えたサブ事業の業種名称は，「サブ業種」フィールドに string の list 型で記載する．
   (e) 各フィールドについて，複数の業種を出力する場合は最大で 5 つまで出力してよい．
7. “# 例” の下に実例をいくつか示すので，参考にすること．

注意
以下の行為は絶対にやめてください．

- 「新業種」フィールドに，業種リストにある名称をコピーしたり意味的に同じものを出力したりすること
- 業種リストにある業種では表現できない事業内容であるにも関わらず，「新業種」フィールドに何も生成しないこと
- 主要事業ではない事業に関する業種名称を「新業種」フィールドまたは「既存業種」フィールドに出力すること
- 各フィールドに，5 個以上の業種名称を出力すること
- 「既存業種」フィールドに，新業種相当の業種名称を出力すること
- サブ事業に関する業種名称を「サブ業種」以外のフィールドに記載すること

業種リスト
米作農業
米作以外の穀作農業
（中略）
他に分類されないサービス業

例 1
（略）

分析対象の企業
web ページのテキスト
（略）

出力
{“新業種”: [], “既存業種”: [], “サブ業種”: []}