評価対象抽出における few-shot 生成手法の提案

今里 昂樹[1]　嶋田 和孝[2]

概要：評価対象抽出とは，テキスト中で評価が向けられている対象表現を抽出するタスクである．評価対象抽出は単語・フレーズ単位のラベル付けが必要であるため，十分なデータセットを確保しにくく，従来のモデルでは学習が困難である．一方，近年 LLM は少数のラベル付きデータ（few-shot）でタスクを遂行することが可能となってきた．LLM はラベル付けを数件に抑えつつタスクを遂行できるが，few-shot の内容で精度が変動する問題も報告されている．そこで本研究では，LLM を用いた評価対象抽出に対し，文法・文構造の整ったデータを生成して few-shot として与える手法を提案し，精度上昇を目指す．

キーワード：評判分析，情報抽出，大規模言語モデル，few-shot learning

A Few-Shot Generation Method for Aspect Term Extraction

Abstract: Aspect term extraction is a task that extracts expressions in a text that are being evaluated. This task requires labeling at the word or phrase level. Because it is hard to collect enough datasets, traditional models are difficult to train. Recently, large language models (LLMs) can perform tasks with only a small amount of labeled data (few-shot). This makes it possible to reduce the number of labeled examples. However, the accuracy changes depending on the few-shot examples. In this study, we propose a method that generates grammatically correct and well-structured data and uses them as few-shot examples for an LLM. Through this approach, we aim to improve the performance of aspect term extraction.

Keywords: Sentiment Analysis, Informatin Extraction, Large Language Models, Few-shot Learning

1. はじめに

自然言語処理分野において，評判分析は古くから広く研究されているタスクの一つである [1]．評判分析は，製品レビューや SNS の投稿などからユーザーの意見や感情を抽出・分析するタスクである．評判分析は企業や組織が消費者のニーズを理解し，製品やサービスの改善につなげるために重要な役割を果たす．テキストの肯定・否定を判別する極性分類を例として，評判分析の多くは，大量の学習データを用いてモデルを構築することで実現されてきた [2]．しかし，評判分析のタスクの中には，データセットが十分に存在しないタスクも存在する．その一つが評価対象抽出である．

評価対象抽出とは，テキスト中に含まれる評価や意見が向けられている対象表現を抽出するタスクである [3]．たとえば，製品レビューでは「品質」「価格」など，ユーザーが特定の製品のどの要素について評価しているかを自動的に抽出する．例文と，その評価対象を以下に示す．

- 例文：このパソコンは何よりデザインが良い．
- 評価対象：デザイン

この抽出によって，ユーザーがどの要素に対して意見を述べているかを明確に理解することができ，より詳細な評判分析が可能となる．しかし，評価対象抽出ではデータセットの作成において，単語やフレーズごとにラベル付けを行う必要があるため，アノテーション作業には多くの労力とコストを要する．また，対象となるドメインによっては専門的な知識が求められることも多く，大規模なラベル付きデータセットの構築は容易ではない．このような背景か

[1] 九州工業大学 大学院情報工学府
Department of Creative Informatics, Kyushu Institute of Technology 680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
[2] 九州工業大学 大学院情報工学研究院 知能情報工学研究系
Department of Artificial Intelligence, Kyushu Institute of Technology 680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN

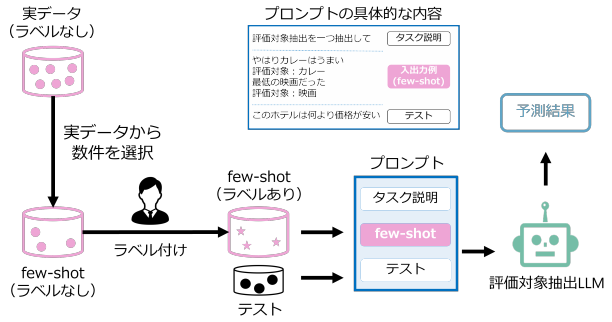

図 1: 一般的な few-shot learning の流れ

ら，学習に必要なデータ量を確保できず，従来の機械学習モデルでは十分な精度が得られにくいことが課題となる．

しかし，近年 GPT に代表される大規模言語モデル（Large Language Models: LLM）が登場し，プロンプトと呼ばれる指示をモデルに与えることによって，多様な自然言語処理タスクを遂行することが可能になってきた．大量の学習データを必要とせず，プロンプトを介してタスクを実行する手法は，In-Context Learning と呼ばれる．特に，プロンプトにタスクの入出力例（few-shot）を少数挿入した学習手法を few-shot learning と呼び，ラベル付きデータが限られているタスクにおいて有効な手法として注目されている．評価対象抽出をタスク例として，few-shot learning の流れを図 1 に示す．few-shot learning では，実環境に含まれるデータ（実データ）を few-shot 候補として利用することが一般的である．具体的には，実データから数件のデータを選択し，人手によるアノテーション後，それらを few-shot として LLM に提供する．LLM は与えられた few-shot を参照することで，タスクの形式や解決のパターンを把握し，推定精度の向上が期待される．この枠組みにより，大量のラベル付きデータを用意することなく，少数のアノテーションでタスクを遂行できる点が few-shot learning の利点である．

一般的に few-shot は，実データからランダムに選択され，各テストデータに対して同一の few-shot が LLM に与えられるが，提供される few-shot によってモデルの推定精度が大きく変動するという問題が指摘されている [4]．この問題に対し，テスト入力と意味的に近い例を検索し，それらを few-shot として与える選択手法 [5] や，類似性に加えて代表性を考慮した選択手法 [6] など，テストデータに合わせて few-shot を選択する動的な選択手法が提案されてきた．

一方で，これらのような既存の few-shot 選択手法の多くは，ラベル付きの実データが十分に存在することを前提としている．しかし，本研究で対象とする評価対象抽出では，ラベル付きの実データを十分に用意することが困難であり，既存の選択手法をそのまま適用できないという課題がある．さらに，SNS への投稿文などを対象とする場合，それらには口語的表現や文法的に不完全な文が多く含まれている．このような実データを few-shot として与えた場合，評価対象と評価表現の対応関係が明確に示されず，LLM が本来理解すべきタスク内容の把握を妨げるノイズとなる可能性がある．

そこで本研究では，実データから few-shot を直接選択するのではなく，LLM を用いて文法および文構造が整ったデータを生成し，それらを few-shot として利用する手法を提案する．このような few-shot を用いることで，LLM はタスク（与えられた指示）をより理解しやすくなり，適切なタスク理解に基づく予測が可能になると考えられる．評価対象抽出において，生成した few-shot を用いることにより，実データに基づく few-shot learning 手法よりも高い推定精度を期待する．

また，生成されたデータであっても，評価対象抽出に対する有用性はデータごとに異なると考えられる．そこで本研究では，生成されたデータからランダムに few-shot を選択するだけでなく，few-shot 選択手法を生成データに対して適用した，生成と選択の組み合わせ手法についても提案する．組み合わせ手法では，選択する生成データに既にラベルが付与されているため，テストデータごとに異なる few-shot を選択する動的選択手法を適用できる点も，利点の一つである．本研究では，まず生成手法単体の有効性を検証し，続いて生成手法と選択手法を組み合わせた手法の有効性を検討する．

2. 関連研究

2.1 評判分析

評判分析は，レビューや SNS 投稿などのテキストから，意見や感情を分析し，ユーザーの評価傾向を把握することを目的とする．評判分析では，ラベル付きデータを用いて分類器を学習し，意見や感情を推定する枠組みが多く用いられてきた [7]．例として，Pang ら [8] は文章の極性分類に，ラベル付きデータを用いた教師あり学習を採用している．

しかし，評価対象抽出のように単語ごとにラベルを付ける必要があるタスクでは，ラベル付きデータを十分に用意できない場合がある．その結果，教師あり学習を前提とした従来手法では，学習に必要なデータ量を確保できず，十分な精度が得られにくいことが課題となる．

2.2 few-shot learning

近年，大規模言語モデルは特定のタスクのみに特化することなく，適切なプロンプトをモデルに与えることで，未知のタスクに対しても対応可能であることが示されている [9, 10]．このようなプロンプトを介した学習手法を In-Context Learning（ICL）と呼ぶ．またこのとき，プロンプト内にタスクの入出力例（few-shot）を少数含め，タ

スクを解かせる学習手法は few-shot learning と呼ばれている．

few-shot learning では，プロンプトに含める few-shot（入出力例）の選び方によって性能が大きく変わることが知られており，few-shot 選択において多数の手法が提案されてきた．代表的な選択手法として，テスト入力と意味的に近い例を検索し，それらを few-shot として与える検索型の選択が有効であることが示されている [5]．さらに，LLM にとって有用な few-shot を選びやすくするため，ICL 用の few-shot 検索モデルを構築する方法も提案されている [11]．また，検索に基づく選択だけではなく，代表性の高い例の選択を組み合わせるなど，複数の基準を統合した選択も検討されている [6]．しかし，これらの手法はいずれも検索対象となる few-shot 候補データに，正解ラベルが付与されていることを前提としている．

検索型の選択手法の他に，多様性や不確実性に着目した研究も多くある．Margatina ら [12] は，能動学習の観点から，不確実性・多様性・類似性などの基準を用いた few-shot 選択を検討した．加えて，Qin ら [13] は，候補例の選択を一度で固定せず，入力と選択を繰り返すことで類似性と多様性の両立を狙う反復的な選択手法を提案している．また，few-shot が推定精度に与える影響は使用するモデルによって異なることが指摘されており，この点に着目して，タスク遂行モデルに応じて few-shot を選択する手法が提案されている [14]．

ただし，多くの few-shot 選択手法は，ラベル付き実データの確保を前提とする場合が多い．そのため，評価対象抽出のようにラベル付きデータの用意が難しい状況では，その前提が満たされないことがある．加えて，評価対象抽出で扱われる SNS 投稿文やレビュー文といった実データには，口語的表現や文法的に不完全な文が含まれることが多く，そのようなデータを few-shot として用いることは，モデルのタスク理解を妨げる可能性がある．このように，ラベル付きデータの不足に加え，実データがノイズを含みやすいという点においても，実データに依存した few-shot 選択手法は，評価対象抽出への適用が容易ではない．そこで本研究では，評価対象抽出のラベル付き実データを用いず，文法・文構造の整ったデータを生成して few-shot として利用する．また，生成データに few-shot 選択手法を適用することで，評価対象抽出における few-shot learning の性能向上を目指す．

3. 実験設定

本章では本研究で用いるデータセットや評価指標，使用するモデルやプロンプトについて説明する．

3.1 データセット

評価対象抽出タスクにおいて，栗原らによって作成された「評価対象-評価表現データセット」 [15] を用いる．栗原らは評判分析タスクのため，Twitter からデータセットの構築を行った．このデータセットは，意見や感想を含むツイートに対して，意見や感想の対象（評価対象）と，意見や感想の内容を表す部分文字列を人手でアノテーションしたものである．本研究では，評価対象-評価表現抽出コーパスの中でも，評価対象が含まれている 4762 ツイートを扱う．実験は 5 分割交差検証を行い，各分割において 953 件をテストデータとして用いる．

なお，本研究の提案手法では，実データから few-shot の選択を行わないが，few-shot 生成時や，比較対象となるベースライン手法（実データから few-shot を選択する手法）では，実データを利用する必要がある．そのため，各分割において，テストデータを除いた 3809 件を本実験における実データとして設定し，利用する．

3.2 ベースライン

本実験では，比較手法として，実データから few-shot を選択する 2 つのベースライン手法を設定する．一つ目は，few-shot をランダムに選択する手法である few-shot$_{\mathrm{RND}}$ である．二つ目は，評価対象抽出に特化した few-shot 選択手法である few-shot$_{\mathrm{IG}}$ である．以下に，各手法の詳細を述べる．

- **few-shot**$_{\mathrm{RND}}$ は，実データからランダムに引き抜いたデータを few-shot として LLM に提供する手法である．few-shot learning において，最も基本的なベースライン手法である．
- **few-shot**$_{\mathrm{IG}}$ は，今里ら [16,17] により提案された，評価対象抽出に特化した few-shot 選択手法である．類似タスクとして極性分類を用い，Information Gain（IG）に基づいて few-shot を選択する．IG は，ある事例を与えたときにモデルの予測不確実性がどれだけ低減されるかを表す指標であり，予測が明確で情報量の高い事例ほど高い値をとる．本手法では，IG が高い事例，すなわちモデルにとって判断が容易で構造が明瞭な事例を few-shot として LLM に提供する．

両手法ともに，実データから few-shot を選択し，すべてのテストデータに対して同一の few-shot を提供する．また，LLM に提示する few-shot には入出力例が必要となるため，選択後に得られた少数の few-shot に対してのみ評価対象のラベルを付与してプロンプトに挿入する（本実験では，データセットに付与済みの正解ラベルを参照して付与する）．

3.3 評価指標

本研究では評価対象抽出の精度を，完全一致，部分一致の 2 点から評価する．完全一致は，予測対象の始点と終点が，どちらも正解と一致した予測対象の数をカウントする．

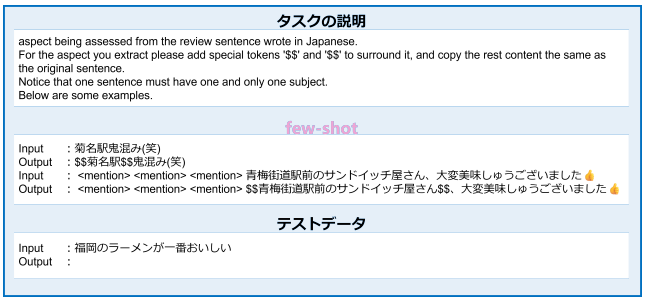

図 2: 評価対象抽出プロンプト

部分一致は予測した文字列と，正解の文字列が一致した文字数をカウントする．下に完全一致，部分一致の例を示す．

- 例文: 福岡のラーメンが一番うまい！
- 正解評価対象: 福岡のラーメン
- 予測評価対象: ラーメンが

この例では正解と予測が完全に一致していないため，完全一致の尺度ではカウントをしない．部分一致では，「福岡のラーメン」という正解（7 文字）に対して，「ラーメン」という予測（4 文字）が一致しているため，Recall は $\frac{4}{7}$ となる．Precision は，「ラーメンが」という予測（5 文字）のうち，「ラーメン」（4 文字）が正解に含まれるため $\frac{4}{5}$ となる．

3.4 評価対象抽出モデル

評価対象抽出を行うモデルには，OpenAI 社によって開発された大規模言語モデル，gpt-4o-mini-2024-07-18[*1]（GPT）を用いる．GPT は，膨大な量のテキストデータを用いて事前学習された，多数のパラメータを持つ深層学習モデルである．実験では，評価対象抽出を行う GPT の temperature の値を 0 とする．temperature は生成されるテキストの多様性を制御するパラメータである．temperature が低いほど生成する文章の多様性が低下し，一貫性のある推定が期待できる．

3.5 プロンプト

本節では，評価対象抽出モデルに対して提供したプロンプトについて説明する．今回モデルに与えたプロンプト例を図 2 に示す．プロンプトは大きく，タスクの説明，few-shot，テストデータの 3 つに分けられる．タスクの説明は，「日本語で書かれたレビュー文から，評価対象を抽出してください．抽出した対象を’$$’ と’$$’ で囲み元の文を出力してください．1 つの文に 1 つの対象が存在することに注意してください.」という内容を英語で記載している．評価対象抽出タスクでは Wang ら [18] を参考に，モデルに評価対象を特殊記号（$$）で囲んで出力するよう指示し，Input,Output を使って出力形式を示している．

なお，本研究では，プロンプトに挿入する few-shot の数を 2 件と 8 件とし，結果を比較する．

[*1] https://platform.openai.com/docs/models/gpt-4o-mini

表 1: 用語とその説明

用語	説明
G-LLM	few-shot 生成 LLM
G-few-shot	G-LLM に与える few-shot
G-プロンプト	G-LLM に与えるプロンプト

4. few-shot を生成する手法

本章では，実データを few-shot として提供するのではなく，LLM を用いて few-shot を生成する手法と，その実験について述べる．

本研究で対象とする評価対象抽出は, SNS 投稿やレビュー文を主なドメインとしている．これらのテキストには，口語的・感情的な表現が多いこと，文法的に不完全な文が頻出すること，ハッシュタグや記号が含まれることといった特徴がある．このようなデータを few-shot として LLM に与えた場合，few-shot が LLM のタスク理解に寄与しない，あるいは LLM のタスク理解を妨げる可能性がある．例として，以下のような状況を想定する．few-shot における正解評価対象を太字で示す．

- few-shot：すげえ！会場に入れない！ #**隼駅まつり**
- テスト：主演の演技がすげえ！ #週末シネマ

上記のような few-shot を与えた場合，評価対象はハッシュタグ部分であるという誤った規則を LLM が学習する可能性がある．その結果，テスト文において本来の正解評価対象である「主演の演技」ではなく，「週末シネマ」を誤って予測する場合が起こり得る．これは 1 例に過ぎないが，文法・文構造が整っていない文章では，評価対象と評価表現の対応関係が明確でない．そのため，文中において評価が向けられている対象を抽出するという，本来理解してほしいタスクの定義が LLM に十分に伝わらない可能性がある．そこで本研究では，まず LLM に評価対象抽出というタスクを正しく理解させることが重要であると考え，文法的に整っており，評価対象と評価表現の対応関係が明確な文を LLM に生成させ，それらを few-shot として用いる．文構造が明確な文では，どの表現が評価であり，どの語句が評価されている対象であるかが明瞭である．このような few-shot を用いることで，LLM は評価されている対象を抽出するというタスクの意図を理解しやすくなり，適切なタスク理解に基づく予測が可能になると期待される．

4.1 手法の流れ

生成した few-shot を用いて，few-shot learning を行う本手法を few-shot$_{\mathrm{GEN}}$ と呼ぶ．評価対象抽出を行う LLM と，few-shot を生成する LLM の識別のため，表 1 に用語とその説明を示す．

few-shot$_{\mathrm{GEN}}$ の流れを図 3 に示す．まず，実データからランダムに選択した数件のデータに対し，人手ラベル付け

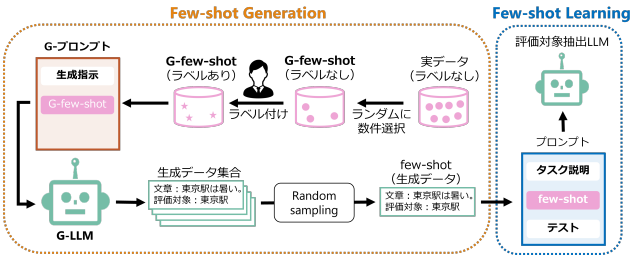

図 3: few-shot$_{\mathrm{GEN}}$ の流れ

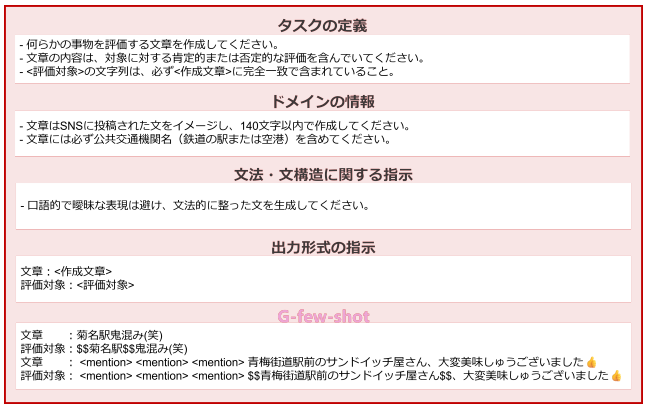

図 4: few-shot$_{\mathrm{GEN}}$ で使用する G-プロンプト

を行い，G-few-shot として G-プロンプトに挿入する．G-プロンプトには G-few-shot のほかに，生成させるタスク（評価対象抽出）の説明や，対象データセットのドメイン情報，文法・文構造が整った文を生成させるといった，生成指示を挿入する．これにより，ドメインの情報を保持しつつ，ノイズを抑えたデータ生成を目指す．最後に，生成されたデータの中から few-shot を選択し，評価対象抽出 LLM に提供するプロンプトに挿入する．本実験では，生成手法単体の有効性を検証することを目的としているため，生成データからの few-shot 選択に特別な基準を設けず，ランダムに選択したデータを few-shot として用いる．

4.2 実験

本節では，few-shot 生成手法における実験設定と，本手法の有効性を検証する．

4.2.1 実験設定

ここでは，few-shot 生成手法に関する実験設定について述べる．具体的には，使用した G-LLM，提供した G-プロンプトの内容，および G-few-shot の数と生成データ数について述べる．

few-shot を生成する G-LLM には，評価対象抽出と同一モデルである gpt-4o-mini-2024-07-18[*1] を用いる．ただし，few-shot 生成過程と評価対象抽出過程は独立したプロセスとして実行しており，互いに影響を与えることはない．また，temperature の値は 1.0 に設定している．G-LLM が生成するデータ数は，各交差ごとに 100 件生成する（計 500 件）．生成されたデータからランダムに選択したものを評価対象抽出 LLM に提供し，評価対象の推定を行う．

G-LLM に提供する G-プロンプトを図 4 に示す．G-プロンプトにはタスクの定義，対象データセットのドメインに関する情報，文法および文構造が整った文を生成するという制約，出力フォーマットに関する指示および G-few-shot を挿入している．本研究で扱うデータセットは，公共交通機関を含む Twitter 文から構成されているため，ドメイン条件として文長を 140 文字以内に制限し，かつ文中に公共交通機関名を必ず含めるよう指示している．また，G-プロンプトに挿入する G-few-shot は，実データ 3809 件からランダムに 8 件選択する．

表 2: few-shot 生成手法の実験結果とその比較

	完全一致			部分一致		
	Pre	Rec	F1	Pre	Rec	F1
2-shot$_{\mathrm{RND}}$	0.397	0.392	0.395	0.475	0.517	0.492
2-shot$_{\mathrm{IG}}$	0.506	0.500	0.503	0.560	0.596	0.577
2-shot$_{\mathrm{GEN}}$	**0.547**	**0.534**	**0.540**	**0.599**	**0.611**	**0.602**
8-shot$_{\mathrm{RND}}$	0.484	0.477	0.481	0.561	0.520	0.539
8-shot$_{\mathrm{IG}}$	0.516	**0.511**	0.513	0.559	0.576	0.567
8-shot$_{\mathrm{GEN}}$	**0.520**	0.509	**0.514**	**0.605**	**0.600**	**0.595**

4.2.2 結果

結果を表 2 に示す．実験の結果，few-shot$_{\mathrm{GEN}}$ は 2-shot，8-shot の双方において，ベースラインである few-shot$_{\mathrm{RND}}$ および few-shot$_{\mathrm{IG}}$ を上回る性能を示した．few-shot$_{\mathrm{GEN}}$ で実際に使用された few-shot の例を以下に示す．この結果から，文構造が整った few-shot を評価対象抽出 LLM に与えることで，LLM が評価対象抽出というタスクをより正確に理解できた可能性が示唆される．

- **Tokyo 駅の新しいビル**が立派で、写真映えしますね！観光名所になりそう。

文中で太字で示されている文字列が評価対象である．生成されたデータは文構造が整っており，可読性の高い文が生成されていることが分かる．

さらに，few-shot 生成手法の利点として，評価対象抽出 LLM に提供する few-shot の数を大幅に増加できる点が挙げられる．一般的に，few-shot learning は few-shot 数の増加とともに推定精度が上昇することが知られているが，提供する few-shot の数にともない，選択されたデータに対して人手によるラベル付けが必要である．しかし，本手法では，選択されたデータには既にラベルが付与されており，few-shot の提供数を増加させても，追加のアノテーションが不要である．そこで，few-shot 数の増加による精度上昇を期待し，16-shot および 32-shot による実験を行った．実験結果を表 3 に示す．結果から，few-shot 数を増加させた場合において，性能の向上を確認することはできなかった．

表 3: few-shot 数を増加させた few-shot$_{\mathrm{GEN}}$ の実験結果

	完全一致			部分一致		
	Pre	Rec	F1	Pre	Rec	F1
2-shot$_{\mathrm{GEN}}$	**0.547**	**0.534**	**0.540**	0.599	**0.611**	**0.602**
8-shot$_{\mathrm{GEN}}$	0.520	0.509	0.514	**0.605**	0.600	0.595
16-shot$_{\mathrm{GEN}}$	0.501	0.492	0.496	0.589	0.577	0.576
32-shot$_{\mathrm{GEN}}$	0.507	0.497	0.502	0.587	0.589	0.581

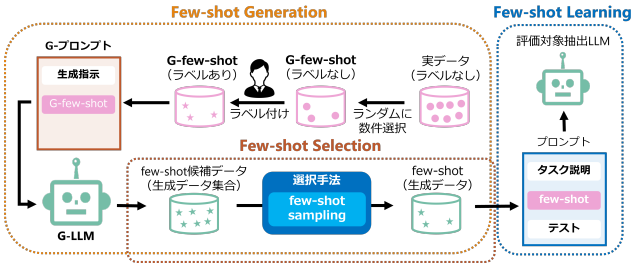

図 5: few-shot の生成と選択を組み合わせた手法の流れ

これは，生成されたデータの多様性が低く，few-shot 数を増加させても LLM に新たな情報を与えられていなかったことが原因だと考えられる．実際に，few-shot$_{\mathrm{GEN}}$ において生成したデータの一部を以下に示す．

- **福岡空港**は国内外のアクセスが良好で、常に快適です。ボード直前も余裕で過ごせます。
- **新大阪駅のトイレ**が綺麗で快適でした。清掃が行き届いていて、安心して利用できました！

太字部分は評価対象である．いずれの文も，「評価対象+評価+補足文」のような文構造を持っており，評価対象の位置や文の展開などが共通している．このように表層は異なるものの，文型や評価対象の配置が固定化されており，生成データの多様性が不足していた．そのため，16-shot や 32-shot のように few-shot 数を増加させたとしても，LLM に新たな情報を与えることができず，性能の向上につながらなかったと考えられる．

以上より，few-shot 生成手法は，生成データの多様性に課題が存在するものの，実データに基づく選択手法以上の性能を示し，有効なアプローチであることが確認できた．

5. few-shot の選択と生成を組み合わせる手法

本章では，生成したデータに対して few-shot 選択手法を適用する手法と，その実験について述べる．4 章の実験結果より，評価対象抽出における few-shot learning では，実データに基づく few-shot よりも，生成データを用いた few-shot の方が高い性能を示すことが確認された．一方で，生成データであっても，すべてのデータが同程度に有用であるとは限らず，データごとに few-shot としての有用性にはばらつきが存在すると考えられる．そこで本研究では，few-shot 生成手法に few-shot 選択手法を組み合わせることで，有用な生成データを選択し，さらなる性能向上が可能かを検討する．

5.1 手法の流れ

本節では，few-shot の生成と選択を組み合わせる本手法の流れについて説明する．手法全体の概要を図 5 に示す．本手法は，few-shot 候補データの生成と，生成した候補データからの few-shot 選択の 2 段階から構成される．まず，LLM を用いて文法および文構造が整ったデータを生成し，few-shot 選択手法を適用するための few-shot 候補データを構築する．次に，生成された few-shot 候補データに対して，何らかの基準に基づく選択手法を適用し，評価対象抽出 LLM に提供する few-shot を決定する．最後に，選択された生成データを few-shot として評価対象抽出 LLM に与え，評価対象抽出を行う．

5.2 実験

本節では，few-shot の選択と生成を組み合わせる手法における実験設定と，本手法の有効性を検証する．

5.2.1 実験設定

まず，few-shot の生成について述べる．few-shot の生成に関しては，G-LLM の種類およびパラメータ，生成に用いるプロンプトの内容，G-few-shot の数，ならびに生成データ数のいずれも，4 章と同一の実験設定を用いる．これにより，生成条件を統一し，選択手法を用いた影響のみを評価する．

次に，生成された few-shot 候補データに対して適用する選択基準について述べる．本実験では，ランダム選択（few-shot$_{\mathrm{GEN}}$）と比較して，few-shot$_{\mathrm{IG}}$ に基づく選択手法の有効性を検証する．また，生成データにはラベルが付与されているため，実データでは困難であった動的選択を適用できる点も本手法の特徴である．そこで，代表的な動的選択である，テストデータとの類似度に基づく選択手法についても実装する．前者に基づく手法を few-shot$_{\mathrm{GEN+IG}}$，後者に基づく手法を few-shot$_{\mathrm{GEN+SIM}}$ と呼ぶ．以下に，各手法の詳細を述べる．

- **few-shot**$_{\mathrm{GEN+IG}}$ は，4 章より生成されたデータ集合に対して，ベースライン手法として紹介した few-shot$_{\mathrm{IG}}$ を適用する手法である．
- **few-shot**$_{\mathrm{GEN+SIM}}$ は，4 章より生成されたデータ集合に対して，類似度に基づく few-shot 選択を行う手法である．類似度に基づく選択は，動的な few-shot 選択手法において一般的に用いられている手法の一つである．具体的には，テストデータおよび生成データをそれぞれベクトル表現に変換し，類似度に基づいてテストデータに近い生成データを few-shot として選択する．これにより，テストデータと構造的・表層的に近いデータを動的に選択することが可能となる．

表 4: few-shot の生成と選択を組み合わせた手法の実験結果とその比較

	完全一致			部分一致		
	Pre	Rec	F1	Pre	Rec	F1
2-shot$_{\mathrm{RND}}$	0.397	0.392	0.395	0.475	0.517	0.492
2-shot$_{\mathrm{IG}}$	0.506	0.500	0.503	0.560	0.596	0.577
2-shot$_{\mathrm{GEN}}$	**0.547**	**0.534**	**0.540**	0.599	**0.611**	**0.602**
2-shot$_{\mathrm{GEN+IG}}$	0.537	0.529	0.533	**0.601**	0.595	0.594
2-shot$_{\mathrm{GEN+SIM}}$	0.529	0.514	0.521	0.596	0.605	0.596
8-shot$_{\mathrm{RND}}$	0.484	0.477	0.481	0.561	0.520	0.539
8-shot$_{\mathrm{IG}}$	0.516	**0.511**	0.513	0.559	0.576	0.567
8-shot$_{\mathrm{GEN}}$	**0.520**	0.509	**0.514**	**0.605**	**0.600**	**0.595**
8-shot$_{\mathrm{GEN+IG}}$	0.513	0.504	0.509	0.592	0.592	0.584
8-shot$_{\mathrm{GEN+SIM}}$	0.515	0.505	0.510	0.588	0.596	0.586

各手法によって選択されたデータを評価対象抽出 LLM に提供し，評価対象の推定を行う．テキストのベクトル化には OpenAI 社が開発した Embedding モデルである `text-embedding-3-small`[*2]を使用し，類似度計算にはコサイン類似度を用いる．

5.2.2 結果

評価対象抽出の結果を表 4 に示す．実験の結果，2-shot および 8-shot のいずれにおいても，生成データを用いた手法（few-shot$_{\mathrm{GEN}}$，few-shot$_{\mathrm{GEN+IG}}$，few-shot$_{\mathrm{GEN+SIM}}$）は，ベースラインである実データに基づく比較手法（few-shot$_{\mathrm{RND}}$，few-shot$_{\mathrm{IG}}$）より高い性能を示した．

一方で，生成データに選択手法を組み合わせた few-shot$_{\mathrm{GEN+IG}}$ および few-shot$_{\mathrm{GEN+SIM}}$ は，few-shot$_{\mathrm{GEN}}$ と比較して顕著な性能向上を示さなかった．4 章より，本研究で生成したデータは，文構造や表現の多様性に限界があることが明らかとなっている．このように多様性の低い生成データ集合に対しては，選択手法を適用した場合であっても，選択される few-shot 間の情報差が小さくなりやすく，選択基準の違いが性能差として現れにくかったと考えられる．そのため，生成データからランダムに few-shot を選択する手法である few-shot$_{\mathrm{GEN}}$ と比較して，性能の向上が見られなかったと解釈できる．言い換えると，生成されたデータはランダムに選択した場合であっても安定した性能を示していることが分かる．これは，生成した各データの品質が一定以上であり，ランダムに選択した場合でも，評価対象抽出に悪影響を及ぼすような few-shot が含まれにくいことを示唆している．

6. おわりに

本研究では，評価対象抽出における few-shot learning の性能向上を目的として，few-shot の生成手法および生成と選択の組み合わせ手法を提案した．

従来の few-shot learning に関する研究の多くは，ラベル付きの実データが十分に存在することを前提とし，その中から有用な few-shot を選択する枠組みを採用している．しかし，評価対象抽出のように単語・フレーズ単位のアノテーションが必要なタスクでは，ラベル付き実データの確保が難しく，既存の選択手法をそのまま適用することは容易ではない．また，SNS 投稿文やレビュー文といった実データには，口語的表現や文法的に不完全な文が多く含まれ，それらを few-shot として与えることが，LLM のタスク理解を妨げるノイズとなる可能性がある．そこで本研究では，実データの選択に依存するのではなく，LLM によって文構造が整ったデータを生成し，それらを few-shot として用いるアプローチを採用した．

生成手法である few-shot$_{\mathrm{GEN}}$ の実験結果から，2-shot および 8-shot のいずれにおいても，生成データを用いた few-shot は，実データを用いた比較手法を上回る性能を示した．この結果は，評価対象と評価表現の対応関係が明確な文を few-shot として与えることで，LLM が評価対象抽出というタスクの意図をより正確に理解できた可能性を示唆している．一方で，16-shot および 32-shot に拡張した場合には性能向上は確認されず，生成文の文型や評価対象の配置が類似していることによる多様性不足が課題として明らかとなった．

さらに，生成したデータに対して few-shot 選択基準を適用することで，生成と選択を組み合わせた手法についても検討した．実験の結果，生成データを用いた手法はいずれも実データに基づく比較手法より高い性能を示したものの，生成データをランダムに用いる few-shot$_{\mathrm{GEN}}$ と比較して，顕著な性能向上は確認できなかった．これは，生成データ集合内の情報差が小さく，選択基準の違いが性能差として現れにくかったためであると考えられる．この結果は，生成されたデータが全体として一定以上の品質を有しており，必ずし複雑な選択を行わなくても安定した性能が得られることを示している．

今後の課題としては，生成段階での多様性強化（文構造や評価対象の出現位置などの分布制約）を導入し，より多様な生成データ集合を構築することが挙げられる．また，本手法の適用範囲を明確化するため，異なるドメインや，異なる LLM に対する汎化性能の調査も重要であると考えられる．

謝辞

本研究は科研費 23K11368 の一部です．

[*2] `https://platform.openai.com/docs/models/text-embedding-3-small`

参考文献

[1] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, Vol. 2, No. 1–2, pp. 1–135, 2008.

[2] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pp. 79–86, 2002.

[3] Kim Schouten and Flavius Frasincar. Survey on aspect-level sentiment analysis. *IEEE transactions on knowledge and data engineering*, Vol. 28, No. 3, pp. 813–830, 2015.

[4] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pp. 12697–12706. PMLR, 2021.

[5] Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures*, pp. 100–114, 2022.

[6] Chencheng Zhu, Kazutaka Shimada, Tomoki Taniguchi, and Tomoko Ohkuma. Staykate: Hybrid in-context example selection combining representativeness sampling and retrieval-based approach – a case study on science domains. *CoRR*, Vol. abs/2412.20043, pp. 1–11, 2024.

[7] Bing Liu. *Sentiment Analysis and Opinion Mining*. Springer International Publishing, 2012.

[8] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 271–278, Barcelona, Spain, July 2004.

[9] 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 *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[10] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[11] Liang Wang, Nan Yang, and Furu Wei. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1752–1767, 2024.

[12] Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. Active learning principles for in-context learning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5011–5034, 2023.

[13] Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. In-context learning with iterative demonstration selection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7441–7455, 2024.

[14] Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Revisiting demonstration selection strategies in in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9090–9101, 2024.

[15] 栗原理聡, 水本智也, 乾健太郎. Twitter による評判分析を目的とした評価対象-評価表現データセット作成. 言語処理学会 第 24 回年次大会発表論文集, pp. 344–347, 2018.

[16] 今里昂樹, 嶋田和孝. 評価対象抽出における関連タスクを利用した few-shot 選択手法. 言語処理学会 第 31 回年次大会発表論文集, pp. 3708–3713, 2025.

[17] Koki Imazato and Kazutaka Shimada. Is a similar task useful for few-shot selection? aspect term extraction using llm. In *International Conference on Applications of Natural Language to Information Systems*, pp. 47–57. Springer, 2025.

[18] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*, 2023.