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

今里昂樹 [1]　嶋田和孝 [1]
[1] 九州工業大学大学院
imazato.kouki927@mail.kyutech.jp　shimada@ai.kyutech.ac.jp

概要

本研究では，評価対象抽出を対象とし，その対象タスクと関連したタスクを利用した効果的な few-shot 選択手法を提案する．評価対象抽出は，テキストから評価の対象となる特定の要素を抽出するタスクであり，データセットの不足が課題となっている．提案手法では，関連タスクとして極性分類のデータを活用し，Information Gain（IG）を用いて few-shot を選択することで，評価対象抽出の訓練データ不足の問題を補いながらモデルの性能向上を目指す．実験では 5 分割交差検証を実施し，ベースラインを上回る精度を達成した．

1　はじめに

自然言語処理分野においては，評判分析が広く研究されている．評判分析は，製品レビューや SNS の投稿，カスタマーサービスにおけるユーザーフィードバックなどから，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．評判分析の代表的なタスク例として，極性分類が挙げられる．極性分類は，テキストがポジティブ（肯定的）であるか，ネガティブ（否定的）であるかを判別するタスクである．このタスクは，製品やサービスに対するユーザーの意見を迅速に把握するために広く活用されている．極性分類は評判分析の中でも代表的なタスクであり，研究が進んでいる分野であるため，公開されたデータセットが豊富に存在する．一方，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．その一つとして評価対象抽出が挙げられる．

評価対象抽出では，テキスト中から評価の対象となる特定の要素や項目を抽出することを目的とする．たとえば，製品レビューでは「品質」「デザイン」「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：このホテルは何より価格が安い
- 評価対象：価格

この抽出によって，ユーザーの意見がどの要素に関連しているかを明確に理解でき，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，極性分類などと比較してより多くの労力とコストを要する．そのため，極性分類のように学習のための豊富なデータが利用できない場合もある．

しかし，近年 GPT のような LLM は多量のデータを必要とせず，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，タスクを遂行することが可能になってきた．追加学習を行わずにプロンプトを介してタスクを実行する手法は In-Context Learning（ICL）と呼ばれる．ICL の中でも，プロンプトにタスクの入出力例（few-shot）を挿入したタスク遂行手法を few-shot learning という．few-shot を LLM に提供することにより，LLM はタスクのパターンや解決方法を学び，より精度の高い結果を導くことができる [1]．これにより，ラベル付け作業を数件に抑えながら，評価対象抽出の精度向上が可能となる．しかし，few-shot learning は，提供される few-shot によってタスクのパフォーマンスが大きく変動するという問題が報告されている [2]．LLM は few-shot の内容に敏感であるため，ラベルを付けるデータの選択は ICL において重要なプロセスである．

そこで本論文では，対象タスクに関連するタスクからの情報を用いた効果的な few-shot 選択手法を提案する．ここでは，対象タスクを前述のように評価対象抽出とし，関連タスクを極性分類とする．データセットが豊富に存在する極性分類のデータを利用して，few-shot 選択モデルを構築し，対象タスクである評価対象抽出の精度向上を試みる．few-shot の

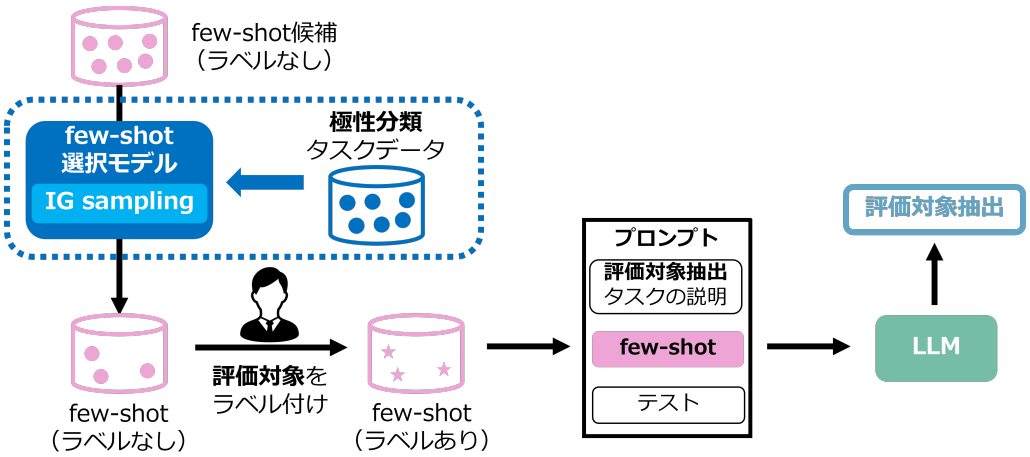

図 1　提案手法の概要

選択には Liu ら [3] の Information Gain（IG）を参考にする．IG とは，あるデータを観察することで予測の不確実性がどれだけ減少するかを定量化した指標であり，データの情報の価値を測るために用いられる．提案手法の概要を図 1 に示す．提案手法ではまず，極性分類データセットを利用して極性分類モデルを作成する．この極性分類モデルを用いて，few-shot 候補の極性（ポジティブまたはネガティブ）を判別する．本論文では，以下の仮定をおく．

- 極性分類が容易であると判断された事例は，対象タスクである評価対象抽出での評価対象が明示的であり，few-shot 候補として適切である

この容易さを「予測の不確実性が減少する」と見なし，IG の値が大きいものを few-shot 候補として選択する．得られた数個の事例に対して人間が事例中に存在する評価対象のラベル付けをし，それらを実際の few-shot として LLM に提供することで，評価対象抽出を実現する．

2　先行研究

Liu ら [3] は，IG を用いてデータの持つ情報能力を測り，few-shot を選別した．IG とは，あるデータを観察することで予測の不確実性がどれだけ減少するかを定量化した指標であり，データの情報価値を測るために用いられる．IG の計算式を式 1 に示す．

$$IG\,(Y|x_i) = H\,(Y) - H\,(Y|x_i) \tag{1}$$

ここで，$H(Y)$ は LLM の予測に対するエントロピーを示す．$H(Y|x_i)$ は few-shot 候補データ x_i を LLM に入力したときの，LLM の予測に対するエントロピーを示す．エントロピーは情報の不確実性，すなわち予測の難しさを数値化している．$H(Y)$ は，few-shot 候補データ x_i に依存しないため一定の値である．したがって，IG が最大となる場合は，条件付きエントロピー $H(Y|x_i)$ が最小の時である．IG が最大となるデータ x_i を選択することで，予測の不確実性を低減させるような情報量を持つ few-shot の獲得が可能となる．

IG の計算にはモデルの出力に対する予測確率が必要となるが，GPT-4 などの LLM では，モデルから直接予測確率分布を取得できない場合も多い．さらに，先行研究はテキスト分類タスクのみに焦点を絞っている．通常のテキスト分類では，1 つのデータ（テキスト全体）に対してラベルを割り当てるため，IG を明確に定義することが可能である．しかし，評価対象抽出のような系列ラベリングタスクでは，ラベリングの対象が複数の要素から構成される場合（複合名詞など）も多い．そのため，評価対象ラベルの割り当てが部分的かつ可変的となり，1 つのデータに対して IG を明確に定義することが困難になる．提案手法では次節で述べるように，対象となるタスクとは別の IG を計算しやすい関連タスクを設定し，BERT などを利用し，確率分布をモデルから直接得ることで，これらの問題を解決する．

3　提案手法

提案手法は，目的である対象タスクを few-shot 選択モデルに直接使わず，関連タスクとその結果の IG 値によって，few-shot を選択する．関連タスクの条件は，対象タスクに関連しており，IG を計算しやすい分類問題であり，学習のための十分なデータがあることである．本論文では，対象タスクを評価対象抽出とし，関連タスクを極性分類とする．

極性分類の例を以下に示す．

- 例文：このホテルは何より価格が安い
- 極性：Positive

極性分類は自然言語処理分野において基礎的な研究分野の一つであり，豊富なデータセットが利用可能である．また，極性分類と評価対象抽出は同じ評判分析のタスクであり，類似性がある．

図 1 における few-shot 選択モデル（IG sampling）の流れを以下に示す．まず，極性分類のデータセットを使用し，事前学習モデルをファインチューニングする．次に，ファインチューニングした事前学習モデルを用いて，few-shot の各候補（ラベルの付いていないデータ）の極性分類を行う．このモデルは，各候補に対してその事例のポジティブおよびネガティブ度合いの予測確率を，softmax 関数などを経て出力する．この予測確率を用いて各候補の IG 値を計算し，すべての few-shot 候補を IG 値でソート後，上位 n 事例を選出し，LLM へ提供する．few-shot 選択モデルは極性分類の予測にどれだけ確信があるかを測定しているだけだが，これを 1 節で定義した仮定を基に few-shot の選択手段として用いる．以降，本研究で提案する few-shot 選択手法を few-shot$_{IG}$，IG を計算するモデルを few-shot 選択モデルと呼ぶ．

4 実験設定

4.1 データセットとモデル

対象タスク：評価対象抽出　評価対象抽出のデータセットとして，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」[4] を用いる．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評価対象を人手でアノテーションしたものである．本研究ではデータセットの中でも評価対象を含む 4762 ツイートを扱う．実験は 5 分割交差検定で行い，テストデータとならない $\frac{4}{5}$ を few-shot 候補として用いる．

評価対象抽出を行う LLM には gpt-4o-mini[1)]（GPT）を用い，Temperature の値は 0 としている．LLM の実行に利用するプロンプトは付録 A を参照のこと．

関連タスク：極性分類　極性分類のデータセットとして，文単位の極性が付与された ACP Corpus [5] を使用する．今回はコーパスからランダムに抽出した 1 万件を，few-shot 選択モデルの学習データとする．

few-shot 選択モデル本体には BERT[2)] [6] を使用する．ファインチューニングの学習率は 5e-5，エポック数は 8，バッチサイズは 16 とする．

4.2 ベースライン

本実験では比較手法として，2 つのベースライン手法を設定する．1 つは，ランダムに few-shot を選択する few-shot$_{RAN}$ であり，もう一つは提案手法と同様に，モデルの予測確率を用いて few-shot を選択する few-shot$_{LCS}$ である．以下に詳細を述べる．

- **few-shot**$_{RAN}$ は，few-shot 候補データからランダムに選んだデータを few-shot として GPT に提供する手法である．
- **few-shot**$_{LCS}$ は，我々の別の論文 [7] で実装された手法である．能動学習 [8] で用いられる Least Confidence Score（LCS）という概念に基づき，few-shot を選択する．LCS の概念を踏まえて，手法の位置づけを端的に説明すると，モデルが予測に対して自信の持てない事例（すなわち予測するのが難しいと判断した事例）を few-shot として GPT に提供することになる．

なお，ベースライン・提案手法ともに 2 件と 8 件の few-shot で実験を実施し，手法の有効性を検証する．

4.3 評価指標

評価対象抽出は，完全一致と部分一致で評価する[3)]．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．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: 長崎のちゃんぽんが一番おいしい．
- 評価対象 $_{\text{正解}}$: 長崎のちゃんぽん
- 評価対象 $_{\text{予測}}$: ちゃんぽんが

この例では正解と予測結果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，「長崎のちゃんぽん」という正解（8 文字）に対して，予測対象の 5 文字が一致しているた

1) https://platform.openai.com/docs/models

2) https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

3) 関連タスクは IG の算出のためだけであるため，本論文では評価はしない．

め，Recall は $\frac{5}{8}$ となる．Precision は「ちゃんぽんが」の 6 文字のうち，5 文字が正解に含まれるため $\frac{5}{6}$ となる．

5 結果・考察

表 1 に GPT による評価対象抽出の実験結果を示す．各手法の先頭の数値は，各手法に対して選択した few-shot の数を表す．2-shot，8-shot においてそれぞれ各指標で最も値が高いものを太字で示す．

実験結果から，shot 数に関係なく，ほとんどの指標で few-shot$_{IG}$ が最も高い値を得ていることがわかる．このことより，提案手法の有効性が確認された．また，一般に shot 数は多い方が精度向上に貢献することが多いが，本実験でも few-shot$_{IG}$ や few-shot$_{RAND}$ ではその傾向が確認できる．一方で，few-shot$_{LCS}$ は，2-shot 設定の完全一致を除き，完全一致・部分一致のいずれの設定においても，few-shot$_{IG}$ や few-shot$_{RAND}$ よりも低い値となった．

few-shot$_{IG}$ と few-shot$_{LCS}$ で選択された few-shot の一例を以下に示す．太字部分は各 few-shot の正解の評価対象である．

- few-shot$_{IG}$: **東急東横線渋谷駅のドラクエ発車BGM** は、どんなに疲れていても元気になれる。
- few-shot$_{LCS}$: なかなか会場に入れない隼達がたくさん。そのくらい集まってます!すげぇ。 **#隼駅まつり** URL

few-shot$_{IG}$ では，IG の定義より，不確実性が低い事例を選択する．これは言い換えると極性分類が容易な事例を few-shot として選出し，それを GPT は評価対象抽出のために利用していることになる．極性分類が容易な事例では，この few-shot$_{IG}$ の例のように，文中でどの対象に評価が行われているかが具体的であったり，詳細でかつ明確に表現されているものが含まれやすい．別の言い方をすると，評価対象が長めの事例が含まれやすく，その長い評価対象が GPT に対する良い事例として機能している可能性がある．一方 few-shot$_{LCS}$ では，直接的な評価対象が文中に明示されない場合が多い傾向があった．評価対象が明示されにくい事例は，明示されているものよりも極性分類が難しい事例だと考えられ，そのような曖昧な文脈を持つ事例では，few-shot 事例として上手く機能しない場合がある．Gonen ら [9] の研究では，モデルにとって困惑度（不確実性）が低いプロンプトほど，タスクをより適切に実行出来ることが報告されている．文脈が曖昧になりがちな few-shot$_{LCS}$ より，評価対象が明確な few-shot$_{IG}$ の方が，モデルにとってタスクの理解が容易であるプロンプトと考えられ，この違いが本研究における few-shotIG と few-shot$_{LCS}$ の精度差を生む要因となったと考えられる．また，Peng ら [10] の研究においても，テストデータとの類似度を基にサンプリングを行った後，不確実性の小さい事例を再サンプリングする手法が提案されている．IG における計算式および Peng らの手法から，不確実性が低いデータの選定が ICL における few-shot 選択の重要な要素であることが示唆される．

表 1 評価対象の抽出精度

手法	完全一致			部分一致		
	Pre	Rec	F1	Pre	Rec	F1
2-shot$_{RAN}$	0.397	0.392	0.395	0.475	0.517	0.492
2-shot$_{LCS}$	0.433	0.417	0.425	0.515	0.437	0.472
2-shot$_{IG}$	**0.506**	**0.500**	**0.503**	**0.560**	**0.596**	**0.577**
8-shot$_{RAN}$	0.484	0.477	0.481	**0.561**	0.520	0.539
8-shot$_{LCS}$	0.428	0.420	0.424	0.520	0.437	0.474
8-shot$_{IG}$	**0.516**	**0.511**	**0.513**	0.559	**0.576**	**0.567**

6 まとめ

本研究では，評価対象抽出を対象とし，対象タスクである評価対象抽出とは異なる関連タスクを解くことで，LLM への良い few-shot を選出する手法を提案した．関連タスクは極性分類とし，Information Gain（IG）という考えに基づいて few-shot 候補の選出を行った．対象タスクを few-shot 選択に直接用いるのではなく，関連タスクによって代替するという考え方は，IG を用いた Liu ら [3] の先行研究では直接扱うことが困難であった系列ラベリングのタスクで，IG を有効利用するための工夫の一つである．

本論文では，対象タスクを評価対象抽出としたが，関連タスクを適切に設定できれば，評価対象抽出以外のタスクでも提案手法は利用可能である．したがって，他のタスクにおいても提案手法が有効に機能するかを検証することが今後の課題の一つである．また，今回の手法はデータセットから静的（static）に few-shot 候補を選択する手法だが，我々は静的な選択手法と動的（dynamic）な選択手法を組み合わせた few-shot 選択手法 [11] や複数の動的手法を組み合わせる手法 [12] についても研究している．これらの手法との統合的な利用も重要な課題の一つである．

謝辞

本研究は科研費 23K11368 の一部です.

参考文献

[1] 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.

[2] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8086–8098. Association for Computational Linguistics, 2022.

[3] Hongfu Liu and Ye Wang. Towards informative few-shot prompt with maximum information gain for in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 15825–15838, Singapore, December 2023. Association for Computational Linguistics.

[4] 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.

[5] Nobuhiro Kaji and Masaru Kitsuregawa. Automatic construction of polarity-tagged corpus from HTML documents. In **Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions**, pp. 452–459. Association for Computational Linguistics, 2006.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, 2019.

[7] Koki Imazato and Kazutaka Shimada. Automatic few-shot selection on in-context learning for aspect term extraction. In **2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)**, pp. 15–20, 2024.

[8] Robert (Munro) Monarch 著・ 上田 隼也 訳・ 角野 為耶 訳・ 伊藤 寛祥訳. Human-in-the-Loop 機械学習. 共立出版, 2023.

[9] Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. Demystifying prompts in language models via perplexity estimation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 10136–10148, Singapore, December 2023. Association for Computational Linguistics.

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

[11] 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.

[12] 朱晨成, 谷口友紀, 大熊智子, 嶋田和孝. Hybrid-set: 意味的類似性とセットカバレッジを考慮した few-shot 例選出手法. 言語処理学会 第 31 回年次大会発表論文集, 2025.

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

A　プロンプト

GPT に提供する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．プロンプトを図 2 に示す．

タスクの説明
Extract the aspect being assessed from the review sentence wrote in Japanese.
For the aspect you extract please add special tokens '$$' and '$$' to surround it,
and copy the rest content the same as the original sentence.
Notice that one sentence must have one and only one subject.
Below are some examples.

few-shot
Input　：菊名駅鬼混み(笑)
Output　：$$菊名駅$$鬼混み(笑)

Input　：<mention> <mention> <mention> 青梅街道駅前のサンドイッチ屋さん、大変美味しゅうございました👍
Output　：<mention> <mention> <mention> $$青梅街道駅前のサンドイッチ屋さん$$、大変美味しゅうございました👍

テスト
Input　：博多のラーメンが一番おいしい
Output　：

図 2　プロンプト

タスクの説明部分には，「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（$$）で囲み元の文を 出力してください．1 つの文に 1 つの対象が存在することに注意してください.」という内容を英語で記載している．評価対象の抽出では Wang ら [13] を参考に，評価対象を特殊記号（$$）で囲んで出力するよう指示している．