

柔軟な言い換えのためのクオリティコントロール機構の開発

Quality control for flexible sentiment style transfer



福本 文代 (Fumiyo FUKUMOTO, Ph.D)

明治大学 総合数理学部 特任教授

(Professor, School of Interdisciplinary Mathematical Sciences, Meiji University)

言語処理学会 Association for Computational Linguistics (ACL)

受賞: Best Award FLP (2020 年), KEOD Best Paper Award 賞 (2012 年), 他
研究専門分野: 計算言語学, 自然言語処理

あらまし

言い換え生成とは、入力文と意味的に等価な文を生成する処理であり、文書の読解支援や要約、あるいは文のスタイル変換など、文生成における様々なタスクの基盤技術として盛んに研究が行われている。従来の言い換え生成に関する研究の多くは、タスクのクオリティ、例えば文のスタイル変換では、流暢さや原文の忠実性、また個々のデータ、例えば Amazon 商品サイトのレビューなどを設定し、その精度を向上させる生成を目指していた。しかし、クオリティは各タスクに依存すること、また個々のデータに依存する手法では汎用性の面で課題が残る。本研究は、文のスタイル変換に注目し、動的なプロンプト手法を提案することにより、クオリティをコントロールし言い換え生成の精度向上を目指す。

1. 研究の背景及び目的

生成 AI の進展により、言い換え生成技術、すなわち入力文と意味的に等価な文を生成する技術は、自然言語処理の応用分野の一つとしてはやくから注目を集めている[6,9,13]。従来の言い換え生成に関する研究の多くは、タスクのクオリティ、例えば文のスタイル変換では、流暢さや原文の内容を忠実に反映できているかなどのクオリティ、あるいは個々のデータ、例えば Amazon 商品サイトのレビューなどに限定し、そ

の精度を向上させる生成を目指していた。しかし、クオリティは各タスクに依存すること、また個々のデータに依存する手法では、汎用性の面で課題が残る[1,3,4]。近年、事前学習、すなわち超大規模なテキストデータから言語の構造を統計的に学習する手法が提案されたことにより、十分な量の学習データが存在しない場合にも事前学習済モデルの汎化性能を活用することが可能となった。しかし、クロスドメイン、すなわち事前学習を異なる分野に適用する場合には、依然として十分な精度が得られていないのが現状である[14,16]。

本稿では言い換え生成のうち、文のスタイル変換に注目し、動的なプロンプトによりクオリティをコントロールし、言い換え生成の精度向上を目指す。具体的には、アスペクトベースの感情分析 (Aspect-Based Sentiment Analysis: ABSA) モデルを用い、感情スタイル変換 (Sentiment Style Transfer) タスクのための動的プロンプトを構築する。さらに、大規模言語モデル (LLM) におけるハルシネーション問題、すなわち LLM が事実と異なることを生成する問題を軽減するため、Self-Refine フレームワークを活用した 2 種類の動的プロンプト手法を提案する。

2. 研究の方法

本研究における動的プロンプトの概要を図 1 に示す。図 1 は動的プロンプトを静的プロンプトとの比較により示した図である。図 1 において、静的プロンプト用のテンプレート P_{static} は入力文について、直接回答を指示するプロンプトであり、“入力文について文の意味を保持したまま、肯定的な感情を表すような文を生成してください”を LLM への入力としている。一方、動的プロンプト $P_{dynamic} = f_{prompt}(ABSA(input))$ は、入力文について感情分析により感情を示す単語 (Aspect term) とその感情 (肯定的、否定的、中立的) を解析し、この結果、すなわち“入力文には否定的な感情を示す単語である location と staff が含まれている”を加え、LLM の入力としている。

我々はさらに、動的プロンプトに Self-Refine フレームワークを組み入れることにより、出力される生成文の精度向上を図った。Self-Refine は、LLM を用い

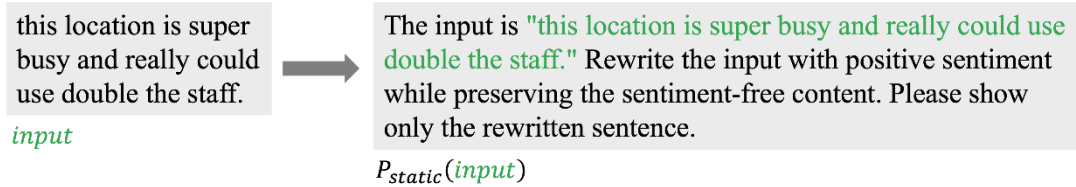
柔軟な言い換えのためのクオリティコントロール機構の開発

Quality control for flexible sentiment style transfer

出力した生成文を評価し、改良を促すプロンプトを与えるというステップを複数回繰り返すことにより、目

的とする生成文に近づける手法を学習する枠組みである。本手法の流れを図2に示す。

♣ Static Prompt



♣ Dynamic Prompt

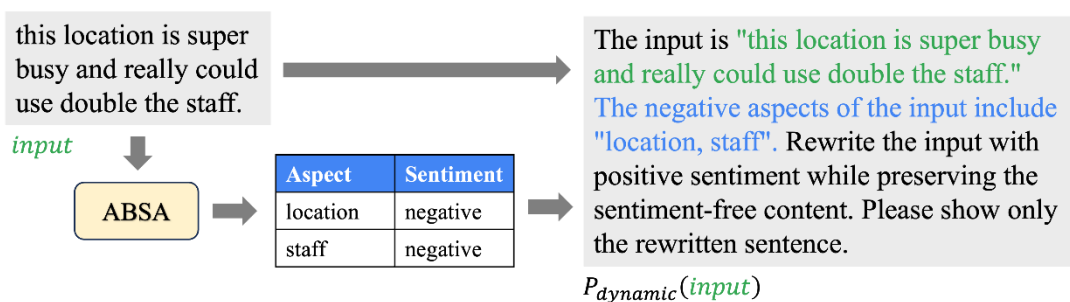


図1 静的プロンプト (Static Prompt) と動的プロンプト (Dynamic Prompt) の流れ

Algorithm Dynamic prompting by using Self-Refine framework

Require: Input x , iteration limitation η , Generation Model $\mathcal{G}_{ilm}$, ABSA model $\mathcal{M}_{ABSA}$, sentiment classifier $\mathcal{M}_{senti}$, target sentiment s_{target} , dynamic refine prompt constructor $\mathcal{C}_{drp}(\cdot)$.

- 1: $stop(text, t) = \mathcal{M}_{senti}(text) \text{ is } s_{target} \text{ or } t > \eta$
- 2: $prompt_{all} = P_{static}(x)$
- 3: $y = \mathcal{G}_{ilm}(prompt_{all})$
- 4: $i = 1$
- 5: **while** true **do**
- 6: **if** $stop(y, i)$ **then**
- 7: **break**
- 8: **end if**
- 9: $P_{refine} = \mathcal{C}_{drp}(\mathcal{M}_{ABSA}(y))$
- 10: $prompt_{all} = prompt_{all} \parallel P_{refine}(y)$
- 11: $y = \mathcal{G}_{ilm}(prompt_{all})$
- 12: $i = i + 1$
- 13: **end while**
- 14: **return** y

図2 Self-Refine を用いた動的プロンプト手法のアルゴリズム

柔軟な言い換えのためのクオリティコントロール機構の開発

Quality control for flexible sentiment style transfer

図 2 において x , η , G_{llm} , M_{ABSA} , M_{senti} , S_{target} , $C_{drp}(\cdot)$ はそれぞれ、入力テキスト、終了の上限回数、LLM モデル、感情分析モデル、感情分析分類器、ターゲット(目的)の感情、動的プロンプト生成器を示す。まず初期設定として、静的プロンプト静的プロンプト用のテンプレート P_{static} に入力文 x を入力し、直接回答を指示する (図 2、2 行目)。次に得られた $prompt_{all}$ を LLM に入力し結果であるテキスト y を得る (3 行目)。停止条件 $stop(text, t)$ を満たすまで y の感情を動的プロンプト生成器の入力とし (9 行目)、感情スタイルが目標スタイルと一致するまで処理を繰り返す (5~13 行目)。なお停止条件は、(1) 最終生成文の感情スタイルが、感情スタイル分類器により目標とするスタイルであると判定された場合、あるいは (2) 反復回数が上限 η 回を超えた場合である。ここで η は終了の上限回数であり、ハイパーパラメータである。

3. 実験

実験では、感情スタイル変換 (SST) タスクにおいて、複数の LLM モデルを用い、静的プロンプト (Static Prompt: SP) と 3 種類の動的プロンプト手法、すなわち動的プロンプト (Dynamic Prompt: DP)、Self-Refine (SR)、および動的プロンプトと Self-Refine を組み合わせた手法 (DP&SR) (図 2) との比較を行った。

実験では 2 つのベンチマークデータセットである、Yelp レビュー [15] 及び Amazon レビュー [7] を用いた。Amazon は Yelp データと比較すると商品に関するレビューであり感情表現が多く含まれていること、また Yelp よりも長文であるという違いがある。各データセットは 1,000 文から構成されており、そのうち 500 文は否定から肯定的なテキスト文への変換 (neg $\rightarrow$ pos)、残りの 500 文は肯定から否定への変換 (pos $\rightarrow$ neg) 用のテストデータとして用いた。関連研究との公平な比較を行うために、Suzgun ら [12] らの Yelp および Amazon データセットと同一のバージョンを用いた。4 種類すべてのプロンプトモードについて、LLaMA2 [8], LLaMA3 [9], Mistral [5], Gemma [13], Gemma2 [2] の 5 種類の主要なオープンソース LLM におけるゼロショット性能を評価した。本稿ではこれ

らのうち、LLaMA2 [8] と Gemma [13] に関して報告する。

すべての実験はゼロショットプロンプトを設定し実施した。ABSA モデル、及び感情分類器として、それぞれ [PyABSA](#) [16] と [pysentimiento](#) [11] を利用した。反復回数の上限 η は、SR および DP&SR について 2 に設定した。また各試行において temperature は 0、乱数シードは 1,234 に設定した。評価尺度は以下の 3 つの観点での尺度を用いた。1 点目は、内容保持 (content preservation)、すなわち原テキストの内容をどの程度保持できているかであり、reference-SacreBLEU (r-sB)、及び self-SacreBLEU (s-sB) スコア [12] を用いた。ここで、r-sB は生成文と正解参照文との距離を測定し、重複した単語が多いほど、類似していると判断される。s-sB はモデルが入力文をどの程度直接コピーしているかを測定する。2 点目は変換強度 (transfer strength) であり、生成文が目標スタイルにどの程度変換されているかを、Accuracy (Acc) を用いて評価した。3 点目は生成文の流暢性 (fluency) であり、平均トークンレベルパープレキシティ (token-level perplexity; t-P) および平均文レベルパープレキシティ (sentence-level perplexity; s-P) を用いた。

実験結果を表 1 に示す。表 1 は、7B サイズの LLM である Gemma、及び LLaMA 2 を用いた 4 種類のプロンプト設定の結果を示す。Static Prompt(SP) はベースラインである。一方、動的プロンプト手法として、Dynamic Prompt (DP)、Self-Refine (SR)、及び両者を組み合わせたプロンプトである (DP&SR) を用いた。表 1 より、LLaMA2 は、r-sB 及び s-sB において優れた精度を示す一方で、DP を用いた場合には s-P 及び t-P の双方において、安定したスコアが得られていない。Gemma は neg $\rightarrow$ pos において精度は一貫しないものの、pos $\rightarrow$ neg では動的プロンプトがすべての評価指標においてベースラインの精度よりも上回っている。Yelp と Amazon データセットそれぞれにおける SP と DP&SR を比較するといずれの LLM を用いた場合にも各評価尺度の総平均の差が DP&SR のほうが SP よりも小さい (Gemma における SP は 4.3、DP&SR は 3.99、LLaMA2 における SP は 8.89、

柔軟な言い換えのためのクオリティコントロール機構の開発

Quality control for flexible sentiment style transfer

DP&SR は 8.08)ことから、DP&SR は、SP と比較して異なるドメインにおいても安定した精度が得られていることがわかる。

表 1 実験結果

LLM	Prompt	<i>neg</i> → <i>pos</i>					<i>pos</i> → <i>neg</i>				
		Acc ↑	r-sB ↑	s-sB ↑	s-P ↓	t-P ↓	Acc ↑	r-sB ↑	s-sB ↑	s-P ↓	t-P ↓
Dataset = Yelp											
Gemma	SP	77.8	5.8	8.5	108	58	90.8	9.1	15.3	114	74
	DP	79.8	5.0	7.6	93	62	90.0	10.2	18.3	139	73
	SR	88.2	5.4	7.7	113	59	97.2	8.9	14.5	113	75
	DP&SR	91.0	4.5	6.6	95	60	97.6	10.2	17.1	113	72
LLaMA2	SP	64.0	9.6	16.2	80	38	83.8	13.2	23.1	119	53
	DP	53.6	9.5	18.1	82	40	78.0	12.8	25.7	115	53
	SR	79.0	9.0	14.3	82	37	96.2	12.9	21.9	119	53
	DP&SR	67.8	8.6	14.4	91	39	93.8	12.7	24.1	121	55
Dataset = Amazon											
Gemma	SP	78.0	6.7	9.5	107	71	76.4	12.8	16.9	117	78
	DP	79.0	6.0	9.1	102	71	75.4	13.1	17.9	152	76
	SR	88.6	6.3	8.9	114	72	88.0	12.1	15.8	113	77
	DP&SR	90.8	5.6	8.3	100	71	86.6	12.3	16.3	113	74
LLaMA2	SP	61.8	11.7	19.0	83	40	64.6	19.2	26.7	81	43
	DP	49.0	13.2	22.5	84	44	59.2	20.8	30.3	188	47
	SR	77.2	10.0	16.0	80	39	86.8	17.6	23.8	82	44
	DP&SR	64.2	11.3	18.5	90	42	81.0	19.4	27.5	95	47

4. 将来展望

本研究で開発したモデルは、複数文書要約にも利用可能であることが確認できている。本研究で提案する言い換え生成モデルは、各ドメインにおける教師付きデータが少量である問題に対処できるため、深層学習を利用する工学や経済分野、例として交通量予測や地震発生予測などにも適用可能である。今後は、本研究の他分野への適用を検討する予定である。

おわりに

本研究は言い換え生成のうち、文のスタイル変換に注目し、動的なプロンプトによりクオリティをコントロールする手法を提案し、言い換え生成の精度向上を目指した。具体的には、アスペクトベースの感情分析モデルを用い、感情スタイル変換タスクのための動的プロンプトを開発した。さらに、大規模言語モデル (LLM)におけるハルシネーション問題、すなわち、LLM が事実と異なることを生成する問題を軽減するため、Self-Refine フレームワークを活用した 2 種類の

柔軟な言い換えのためのクオリティコントロール機構の開発

Quality control for flexible sentiment style transfer

動的プロンプト手法を提案した。Amazon と Yelp、及び LLM として 5 種類のモデルを用いた実験の結果、本手法は異なるドメインに対してもベースラインの手法を上回ることが確認できた。特に Gemma では、neg → pos において精度は一貫しないものの、pos → neg では動的プロンプトがすべての評価指標においてベースラインの精度よりも高い精度が得られることが確認できた。

用語解説

1. 文のスタイル変換：文の意味を保持したまま、文のスタイル、例えば肯定的から否定的、論調的から口語的などに変換する技術。
2. プロンプト手法：ユーザが大規模言語モデル (LLM) に与える指示文。
3. アスペクトベースの感情分析：テキストに含まれるサービスや価格などのアスペクト(観点)について、肯定/否定/中立かを判断する技術
4. ゼロショットプロンプト：LLM に指示文とその回答の組を与えず、直接指示文を与え、回答を得る技術。

参考文献

- [1] Dai, N., Liang, J., Qiu, X., Huang, X.: Style transformer: Unpaired text style transfer without disentangled latent representation. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5997–6007. Association for Computational Linguistics, Florence, Italy (Jul 2019).
- [2] DeepMind, G.: Gemma 2 is now available to researchers and developers (2024), <https://blog.google/technology/developers/google-gemma-2/>
- [3] Han, J., Wang, Q., Zhang, L., Chen, W., Song, Y., Mao, Z.: Text style transfer with contrastive transfer pattern mining. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational

Linguistics (Volume 1: Long Papers). pp. 7914–7927. Association for Computational Linguistics, Toronto, Canada (Jul 2023).

- [4] He, J., Wang, X., Neubig, G., Berg-Kirkpatrick, T.: A probabilistic formulation of unsupervised text style transfer (2020),

<https://arxiv.org/abs/2002.03912> 8

- [5] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023), <https://arxiv.org/abs/2310.06825> 9.

- [6] Jin, D., Jin, Z., Hu, Z., Vechtomova, O., Mihalcea, R.: Deep learning for text style transfer: A survey. Computational Linguistics 48(1), 155–205 (Mar 2022)

- [7] Li, J., Jia, R., He, H., Liang, P.: Delete, retrieve, generate: a simple approach to sentiment and style transfer. In: Walker, M., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1865–1874. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018).

- [8] Meta: Llama 2: Open foundation and fine-tuned chat models (2023), <https://arxiv.org/abs/2307.09288>

- [9] Meta: Introducing meta llama 3: The most capable openly available llm to date (2024), <https://ai.meta.com/blog/meta-llama-3/>

- [10] Mikolov, T., Karafiát, M., Burget, L., Černocký, J., Khudanpur, S.: Recurrent neural network based language model. In: Proc. Interspeech 2010. pp. 1045–1048 (2010).

- [11] Pérez, J.M., Giudici, J.C., Luque, F.: pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks (2021), <https://arxiv.org/abs/2106.09462>

- [12] Suzgun, M., Melas-Kyriazi, L., Jurafsky, D.:

柔軟な言い換えのためのクオリティコントロール機構の開発

Quality control for flexible sentiment style transfer

Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 2195–2222. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022).

[13] Team, G., et al., T.M.: Gemma: Open models based on gemini research and technology (2024), <https://arxiv.org/abs/2403.08295>

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30 (2017).

[15] Wang, Y., Wu, Y., Mou, L., Li, Z., Chao, W.: Harnessing pre-trained neural networks with rules for formality style transfer. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3573–3578. Association for Computational Linguistics, Hong Kong, China (Nov 2019).

[16] Yang, H., Zhang, C., Li, K.: Pyabsa: A modularized framework for reproducible aspect-based sentiment analysis. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023. pp. 5117–5122 (2023).

[17] Zhang, Y., Ge, T., Sun, X.: Parallel data augmentation for formality style transfer. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3221–3228. Association for Computational Linguistics, Online (Jul 2020).

この研究は、令和4年度SCAT研究助成の対象として採用され、令和5～7年度に実施されたものです。