拡散モデルを用いた自然言語文内の有害表現除去

理学専攻・情報科学コース
2340682
Zheng Ying

1　はじめに

Web などから収集されたコーパスを用いて事前に訓練された言語モデルは，有害表現を含んだ攻撃的なテキストを生成する可能性がある．言語モデルを安全に利用するため，生成された有害表現を除去することは自然言語処理においてとても重要な課題である．有害表現除去の目的は，元のテキストの意味を変えることなく，有害な表現をより穏和な表現に変換することである．本研究では，テキスト生成における有害表現を除去することに焦点を当て，有害表現を除去しつつもテキストの生成品質を維持できる拡散言語モデルの構築を目指す．複数のデータセットを用いた転移学習や，有害表現を含むテキストへの認識能力を強化するファインチューニングなどの実験を行い，有害表現を除去するために必要な共通知識を追加したモデルを提案する．

2　転移学習に基づく有害表現除去

文中の言語におけ有害表現除る去タスクにおいて言語モデルの訓練に必要となるパラレルの構築には多大な人的コストがかかるため，一般にコーパスのデータ量は十分とは言えない．訓練データが不足している状況では，言語モデルが有害表現の除去を学習する際に生成テキストの意味と文脈が大きな影響を受ける可能性がある．

一方，学習データの不足を補い知識を追加していく学習方法に転移学習がある．転移学習は複数のタスクを順次モデルに学習させ，新しいタスクを学習する際に以前に学習したタスクを忘れないようにする手法である．この方法は人間の学習プロセスから着想を得ており，人間は学習や経験を通じて得た情報を蓄積し，新しいスキルを効果的に開発できる．継続的に学習を行うことで，モデルはこれまでの学習から得られた知識を活用し，新しいタスクにおいても優れたパフォーマンスを得るように考えられている．新しい知識をモデルに追加することが可能と言われている．また，モデルを完全に再学習する必要がなく，既存の知識に新しい知識を統合することを目的として実施される．本研究では，系列の変換・生成能力が高い拡散モデルを用いて有害表現を除去しつつ生成品質に焦点を当て，転移学習に基いて有害表現の検出に役立つ方法を考案する．

図 1 に提案手法の概要を示す．モデルは転移学習用のタスクに応じて 3 つのモジュールから構成されており，換言タスク，有害表現識別タスクと有害表現除去タスク行うものである．換言タスクでは，転移学習の第一段階として，学習を通じて獲得した語彙，意味，文脈について情報を活用することで，換言における生成品質を効果的に向上させる．有害表現識別タスクでは，有害表現を含むテキストを識別するためにエンコーダが学習され，そのものを識別する能力を備える．有害

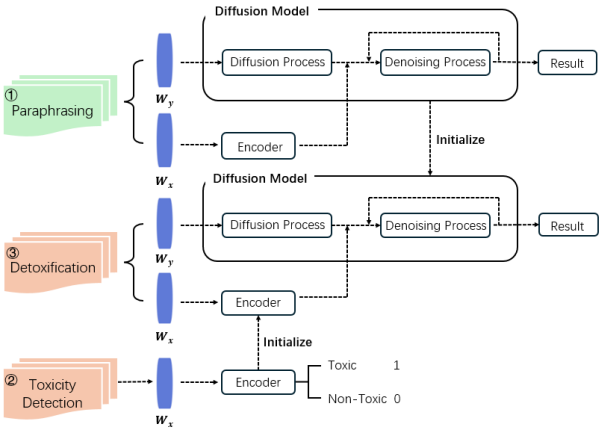


図 1: 転移学習を導入した有害表現除去モデル

表現除去タスクは，転移学習の第二段階として換言タスクで学習済みのモデルを基に再学習する．元のテキストの意味を変えることなく，有害な表現を除去する．同時にデノイジング過程において，有害表現識別タスクで訓練されたエンコーダを用いて有害表現を含むテキストを処理する．本研究では，系列を変換する拡散言語モデルとして，DiffuSeq [1] と SeqDiffuSeq [2] を使用する．

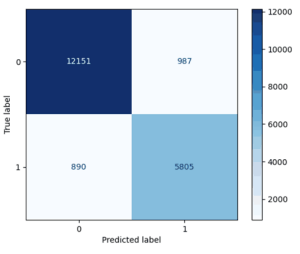
3　実験

3.1　有害表現識別の実験

3.1.1　実験設定

有害表現識別のモデルの学習には，Real Toxicity Prompts [3] というデータセットを使用した．データセットの中には，プロンプト，生成文，および Toxicity スコアの 3 つの部分で構成されている．完全なテキストを得るために，プロンプトと生成文を 1 つの文に組み合わせた．また，Toxicity スコアが 50 未満のテキストにはラベル 0 を付け，50 を超えるテキストにはラベル 1 を付けた．拡散モデルへの有害表現を含む自然言語文は事前学習済み Bert Model のエンコーダを使用して埋込ベクトルへと変換される．事前学習によって得られた知識を活用するため，12 層レイヤーの内 1-10 層をフリーズして有害表現識別に対するエンコーダの学習を行った．

表 1: 有害表現識別に対する結果



評価指標	精度
Acc	0.91
F1-Score	0.86

3.1.2 実験結果と考察

実験結果を表 1 に示す．Accuracy と F1 スコアはそれぞれ 0.91 と 0.86 となっており，エンコーダにおける識別性能が高いことを表す．混同行列から見ると，0 のラベルとして 0 に予測されたサンプル数がおよそ 12,000 であり，1 よりデータ数が多いことがわかった．データセット内でラベルが 1 とされたサンプル数が十分でないため，エンコーダがそれらのサンプルを識別する際に影響を及ぼす可能性がある．

3.2 転移学習の実験

3.2.1 実験設定

転移学習には，換言タスクとして Quora Question Pairs*，有害表現除去タスクとして Paradetox [4] のデータセットを使用した．拡散言語モデルとして DiffuSeq, SeqDiffuSeq を使用した．また，ベースラインとなる言語モデルとして自己回帰型の言語モデルとして T5-base Model†を用いた．実験 i の設定を有害表現識別エンコーダーと転移学習の組み合わせとし，実験 ii は転移学習のみとする．また，実験 iii は直接的に有害表現除去を行うものとする．

3.2.2 実験結果と考察

実験結果を表 2 に示す．評価指標には，BLEU, BERT スコア, Toxicity スコアを用いた．それら 3 つの評価指標により生成文と正解文間に文字列類似性, 意味的な類似性と有害表現を含む程度を測定することができる．また，BLEU や BERT スコアによって測られる生成品質と Toxicity スコアによって測られる有害表現の含有率においては，トレードオフの関係およびそれぞれの重要性に偏りがあることから，それらを総合して考慮するための指標として Joint を定義する．生成品質，有害表現の含有率に対する重みは経験的に設定している．

$$Joint = 0.3 * (Bleu + BertScore) + 0.7 * (1 - Toxicity)$$

有害表現識別エンコーダを加えた後，DiffuSeq は生成品質についての指標はほぼ変化がなく，Toxicity スコアが 0.03 ほど下がる．一方，Seqdiffuseq の生成品質のスコアは大幅に下がり，Toxicity スコアがほぼ変化がないことから DiffuSeq より生成不安定なことがわかった．表 3 に示す DiffuSeq の生成失敗例からみると，提案手法は有害表現を含むテキストを上手く除去できないことがある．直接，有害表現除去する場合は，Toxicity スコアが一番低くなるが，入力文に対してと意味的に関係のない文が生成されており，この実験設定下ではモデルが上手く機能していないことを示している．また，T 5 モデルと比べて DiffuSeq の BLEU は一番高いだけではなく，Toxicity スコアも一番低い．BERT スコアは T5 モデルより少し低い結果となった．実験の結果，拡散言語モデルは非自己回帰モデルとして，有害表現除去タスクに対して大規模事前学習済みの自己回帰型言語モデルに匹敵，または上回る性能があることがわかった．

*https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

†https://huggingface.co/google-t5/t5-base

表 2: 有害表現除去に対する結果

Model	Exp	Bleu	BertScore	Toxicity	Joint
DiffuSeq	i.	0.43	0.82	0.16	**0.96**
	ii.	**0.45**	0.84	0.19	0.95
	iii.	0.25	0.70	**0.09**	0.92
SeqDiffuSeq	i.	0.25	0.73	0.17	0.87
	ii.	0.33	0.80	0.18	0.91
	iii.	0.21	0.70	0.14	0.88
T5-base	iii.	0.37	**0.85**	0.18	0.94

表 3: DiffuSeq の生成失敗例

Exp	Text
Input	I'll kick you out and replace you!
Ground Turth	I'll replace you!
DiffuSeq i.	I'll kick you and replace you
DiffuSeq ii.	I'll kick you and replace you
DiffuSeq iii.	I will with you out

4 おわりに

本研究では，テキスト生成における有害表現を除去することに焦点を当て，高性能な系列変換を実現する拡散言語モデルを対象に転移学習を導入することにより，学習データの不足を補いつつ，事前学習済み自己回帰型言語モデルである T5 よりも高精度な有害表現除去を実現できることを示した．

今後の課題としては，マルチタスク学習で換言タスクと有害表現除去タスクを同時に訓練するようにしたい．また，使用するデータセットを増やし，汎用性を高めるためのさらなる実験を行なっていきたい．

参考文献

[1] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023.

[2] Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. Text diffusion model with encoder-decoder transformers for sequence-to-sequence generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 22–39, Mexico City, Mexico, June 2024.

[3] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models, September 2020. arXiv:2009.11462 [cs].

[4] Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. ParaDetox: Detoxification with Parallel Data. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6804–6818, Dublin, Ireland, May 2022. Association for Computational Linguistics.