フィルタリングによるタイトル-本文ペアの要約教師データの改善

狩野 竜示 † 谷口 友紀 † 大熊 智子 †
† 富士ゼロックス株式会社
{kano.ryuji, Taniguchi.Tomoki, Ohkuma.Tomoko}@fujixerox.co.jp

1 はじめに

タイトルを要約とみなして，生成型要約モデルを学習させる試みは，Rush[1] 以降広く行われてきた．多くはニュース記事のタイトルを利用しているが，それ以外にも，ソーシャルメディアの投稿 [2]，レビューサイトの投稿 [3]，メールのタイトル [4] など，様々な媒体のテキストで応用されている．しかし，タイトルが要約の教師データとして適切かどうかは度々疑問が呈されてきた．特にソーシャルメディア，レビューサイト，メール等の，不特定多数の人物が自由に執筆できる媒体においては，その質は担保されていない．Li ら [3] はレビューサイトのデータに，Zhang ら [4] はメールデータにおいて，要約として不適切なタイトルが多く存在している事を指摘している．本研究の目的は，要約の学習データからこうした不適切なデータをフィルタリングする手法を提案することである．

要約データから一部のデータをフィルタリングする先行研究は，知る限り 2 つ存在する，松丸ら [5] は，含意関係判定器を使い，タイトルが本文を含意しないデータを学習データから除去する手法を提案した．長谷川ら [6] は，要約モデルが本文中のフレーズを多用する事に着目し，本文との重複度が低いタイトルを学習データから除去する手法を提案した．しかしながら，前者は含意判定のための教師データの構築を必要とし，後者はデータセットのタイトルが抽出性の強いものであるという前提に依拠したルールベースの手法を使用している．本研究では，教師無しで機械学習による要約データのフィルタリング手法を提案する．

今回我々は，Gregoire ら [7] の手法を要約タスクに応用する．これは，翻訳タスクにおいて，Siamese Network を用いて対応する 2 文を取得する手法である．正しい本文とタイトルのペアを正例，誤ったペアを負例として，フィルターモデルを学習させる．負例はネガティブサンプリングで取得する．学習させたフィルターモデルを使い，再び学習データの正例のみを判定させる．フィルターモデルは学習データに含まれる正例であっても，負例と判定する事がある．その確率値が閾値以下のデータを除去し，要約モデルを学習させる実験を行った．

メールデータとソーシャルメディアデータの代表として，Enron メールデータ [4][1] の subject と，Reddit[2][3] TIFU データ [2] のタイトルを実験に使用した．結果，我々の提案モデルは，RedditTIFU データにおいて，ランダムにフィルタリングするよりも高い ROUGE 値を得られる事が判明した．また，実際にフィルタリングされたデータを分析すると，それらの多くは，本文からタイトルを予測する事が困難なペアである事が判明した．

2 関連研究

タイトル-本文のペアが学習データとして不適切だという報告は多くの先行研究でされている．Li ら [3] はレビューサイトのタイトルを要約の教師データとして利用したが，その際，ある程度の質を担保させるために，ルールベースや，分類モデル等を使用して，フィルタリングを行っている．Zhang ら [4] はメールのタイトルをメール本文の要約とみなして学習を行っているが，質が担保されていないため，開発データとテストデータを人手で別途作成している．

入力と出力の適切なペアを取得するという観点での研究は，要約タスク以外では広く行われてきた．Gregoire ら [7] は，Siamese Network を使って，2 言語のテキストから対応関係のある 2 文を抽出し，得られたデータを既存学習データに加える事で翻訳性能を向上させた．Qin ら [8] は，関係抽出の性能の差を報酬として，Web スクレイピングしたデータのノイズ除去を行う強化学習器を提案した．Csaky ら [9] は，対話モデルが'I don't know' 等のどのような発言に対して

[1] https://www.cs.cmu.edu/~enron/
[2] https://www.reddit.com
[3] Reddit は、Reddit Inc. の登録商標です

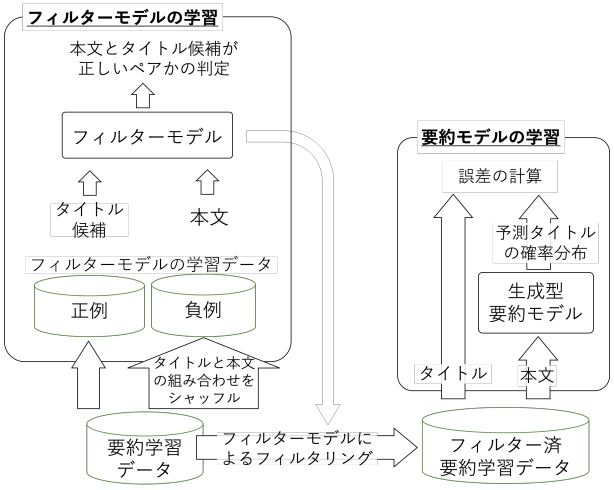

図 1: フィルターモデルと要約モデルの学習フロー

も応答となりうる Generic Response を生成する傾向にある事を問題視し，それらを学習データから除去するためにエントロピーを使用した．

要約データにおける，入力と出力ペアのフィルタリングの研究については，**1 はじめに**で述べた，松丸ら [5] と．長谷川ら [6] の研究がある．

3 実験

図 1 に，フィルターモデルと要約モデルの学習フローを載せる．本研究では，不適切なタイトル-本文ペアを除去するフィルターモデルと，要約モデルを使った実験を行う．提案手法のフィルターモデルによって，要約モデルの学習データから不適切なペアを除去した後に，要約モデルを学習させ，精度の変化を検証する．

3.1 フィルターモデル

フィルタリングの方法として，Gregoire ら [7] の手法を踏襲する．この研究では，Siamese Network を利用して，翻訳の対となる文を取得し，新たに学習データに加える事で翻訳モデルの精度を向上させた．翻訳前の言語の文と翻訳後の言語の文をモデルへの入力とする．モデルは正しい翻訳になっているペアとそうでないペアかを判定するように学習する．学習後のモデルで，文単位の対応関係がわからないペアに対して予測を行い，正例を新たに学習データに加える事で精度を向上させた．

本研究では，タイトル-本文のペアの適切さをフィルターモデルが学習する．先行研究との相違点は，先行研究では学習データを増やすために分類モデルを使用しているのに対し，本研究では学習データから不適切なものを除去するために採用している．実際のタイトル-本文ペアを正例，ランダムにサンプリングされたペアを負例として学習を行う．学習後，フィルターモデルは，学習データの内の正例のみを再判定する．予測確率の低いデータ下位 n%を要約モデルの学習データから除去する．フィルターのモデルには Decomposable Attention[10] を用いた．

パラメータ　単語 Embedding の次元は 300, 初期値を GloVe[4]の単語ベクトルと同等にした．Decomposable Attention モデル内の，Attend Feedforward ネットワーク，Aggregation Feedforward ネットワークに通した後の次元はそれぞれ 100 とした．Optimizer には Adagrad を使用し，損失関数は Cross Entropy を使用した．

3.2 要約モデル

要約モデルには CopyNet[11] を使用した．CopyNet は注意機構付き Encoder-Decoder モデルに，入力文（本文）に含まれる未知語を出力文（要約）に生成できる機構を加えたモデルである．

今回は，フィルターモデルによって予測確率の下位 5%, 10%, 15%, 20% を学習データから除去して要約モデルを学習した場合の精度と，ランダムに同数だけデータを除去して要約モデルを学習させた場合の精度を比較する．要約モデルの精度評価には，ROUGE-1-F (R1)，ROUGE-2-F (R2)，ROUGE-L-F (RL) を使用する．最適化時，パラメータの初期化時，フィルタリング時のランダム性が結果に影響する事を防ぐため，要約モデルの学習は 10 回行い，各精度の平均値を利用する．Epoch 数は 5 で，開発データにおける ROUGE-1-F 値が最大の Epoch のモデルをテストに使用する．

パラメータ　フィルターと同様に単語 Embedding の次元を 300，初期値に GloVe を採用した．隠れ層の次元は 256 とした．Beam Search のサイズを 8 とし，Optimizer には Adam を採用し，損失関数には Cross Entropy を採用した．

[4] https://nlp.stanford.edu/projects/glove/

	評価指標	除去するデータの割合				
		0%	5%	10%	15%	20%
提案手法	R1	0.168	0.167	0.167	0.170	0.171
Random	R1	0.168	0.167	0.165	0.167	0.164
提案手法	R2	0.064	0.064	0.063	0.064	0.065
Random	R2	0.064	0.064	0.063	0.064	0.063
提案手法	RL	0.084	0.082	0.083	0.084	0.085
Random	RL	0.084	0.083	0.082	0.082	0.081

表 1: TIFU タイトルでの結果

	評価指標	除去するデータの割合				
		0%	5%	10%	15%	20%
提案手法	R1	0.241	0.241	0.239	0.247	0.242
Random	R1	0.241	0.240	0.241	0.243	0.240
提案手法	R2	0.096	0.098	0.097	0.098	0.094
Random	R2	0.096	0.096	0.097	0.095	0.090
提案手法	RL	0.127	0.126	0.124	0.130	0.126
Random	RL	0.127	0.126	0.126	0.128	0.128

表 2: Enron subject での結果

3.3 データ

実験には，Enron メールデータ [4] の subject と，Reddit TIFU データ [2] のタイトルを使用する．

Enron データセット Enron メールデータは元々，2004 年に公開された Enron 社のメールデータセット [12] であるが，これらのデータセットをタイトル生成タスク用に整備したものが，Zhang ら [4] により公開された．これは 14,436 の学習データと，1,906 の開発データと 1,906 のテストデータを含む．学習データのメール subject は，2004 年に公開されたデータセットと同じものが使われているが，開発データとテストデータについては，新たに人手で作成されたものである．これは，はじめにで指摘したように，元々のメールデータに含まれる subject に内容を反映していない不適切なものが多いからである．メール本文と subject は，nltk[5]を用いて単語に tokenize した．

Reddit データセット Reddit TIFU データセット[6]は，Reddit の Subreddit の一つである TIFU（Today I fucked up）の投稿を集めたものである [2]．各投稿にはタイトルが付けられており，そのタイトルを投稿本文の要約とみなすデータセットである．投稿本文とタイトルの対，計 79,015 対を 9:0.5:0.5 の割合で，学習データ，開発データ，テストデータに分割し，各データの数は 71,113，3,951，3,951 となった．公開されているデータセットに含まれるテキスト（投稿本文とタイトル）は，予め spacy[7]を用いて単語に tokenize されているため，そちらを利用する．

[5] https://www.nltk.org
[6] https://github.com/ctr4si/MMN

4 結果と考察

4.1 フィルターモデルの学習結果

学習後フィルターモデルのタイトル-本文ペアを正しく判定する精度（F1 値）は，TIFU タイトルデータで 0.930，Enron subject データで 0.800 であった．TIFU タイトルデータにおいてより精度が高かった理由としては，TIFU タイトルの方が Enron の subject に比べ要約長が長い事，また，Reddit の投稿自体の内容がメールデータに比べ多岐に渡るので，本文との関係性を予測しやすい点が挙げられる．

各データセットのフィルタリング（全データの 5%，10%，15% ，20%）を行う際のフィルターモデルの予測確率値の閾値は，Enron subject データにおいて，0.215，0.307，0.390，0.467 であり，Reddit タイトルデータにおいて，0.246，0.424，0.584，0.717 であった．閾値の値が高めになっているのは，フィルター対象のデータがフィルターモデルの学習データにおける正例だからである．

4.2 要約モデルの学習結果

フィルター後の要約モデル学習結果を表 1 と表 2 に載せる．TIFU タイトルデータの場合，フィルターによって除去される学習データが増えるたび，Random の結果は悪化していったが，我々の提案手法では，精度が向上していった．Enron subject データにおいては，除去率が 15%の際は，提案手法の精度が Random を上回ったが，他の除去率においては同程度となった．

4.3 フィルタリングされたペアの具体例

フィルターによってフィルタリングされたデータの具体例を表 3 に載せる．フィルタリングされたデータの多くは，本文から要約を予測することが難しかった．ソーシャルメディアやメールに起こりうることは，本文とタイトルが別の内容を伝えているということである．特に TIFU データでは，表の例のように，タイト

[7] https://spacy.io

データ	タイトル	本文	予測確率
TIFU タイトル	Trimming my beard; a tale of woe	I have strong beard, it's been growing for 10 months. start trimming accidentally trim off too much compensate. Depression kicks in.	1.000
TIFU タイトル	Telling my students a PERSON PERSON joke	They just looked at me weirdly and thought I was some kind of horrible person now I guess I should just teach what is written in the textbook	0.004
Enron subject	Offline NDA form	As an fyi, from time to time I will be preparing NDAs for the networks team headed by marks. PERSON working with PERSON on. Project offline has evolved a form of NDA and added a non-solicitation clause and a residuals clause. (後略)	1.000
Enron subject	Lexis luncheon - Wed. 9/22 11:30 - 1:00 eb46c1	PERSON, although this presentation is for the legal dept. I thought maybe if you have a representative from your group there it might be helpful. Do you have someone, like PERSON PERSON, that you would like to attend? Let me know, and I'll get their name added to the list.	0.009

表 3: フィルターモデルの予測確率が高かった／低かったタイトル-本文ペア．人名は PERSON に置換されている．

ルの続きを本文で記しているため，本文にタイトルの情報が含まれていない例が多く見られた．反対に，予測確率が高かったペアのタイトルは，本文の内容を反映したものになっていた．

5 おわりに

今回我々は，教師あり要約モデルで使用されているタイトル-本文のペアに教師データとして不適切なものが多く含まれていることを問題提起し，それを機械学習モデルでフィルタリングする手法の提案と実験を行った．フィルターモデルは，正しいタイトル-本文ペアとランダムサンプリングによって得られたペアを分類することで学習される．

実際のタイトル-本文のペアの内，学習されたフィルターモデルが負例だと判定する確率に応じてフィルタリングを行い，フィルターされたデータを使い要約モデルの学習を行った．実験では，Reddit TIFU データのタイトルと，Enron メールデータの subject を使用し，検証を行った．また，検証のため，ランダムにデータをフィルターした場合と比較した．結果，Enron データセットでは精度はほぼランダムと同等であったが，TIFU データセットでは，ランダムよりやや高い ROUGE 値が得られた．

今回の実験では，要約モデルの精度がそれほど改善しなかったため，今後はモデルの改善を重ね，より多くの精度向上を目指す．具体的には，フィルターモデルと要約モデルの同時学習を実現するために，強化学習などを使用する．また，テストデータ自体にも，フィルタリングされるべき不適切なデータが含まれている可能性があるため，今後は人手でのテストデータの整備をも検討する．

参考文献

[1] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *EMNLP 2015*.

[2] Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. Abstractive summarization of Reddit posts with multi-level memory networks. In *NAACL 2019*.

[3] Junjie Li, Haoran Li, and Chengqing Zong. Towards personalized review summarization via user-aware sequence network. In *AAAI 2019*.

[4] Rui Zhang and Joel Tetreault. This email could save your life: Introducing the task of email subject line generation. In *ACL 2019*.

[5] 松丸和樹, 高瀬翔, 岡崎直観. 含意関係に基づく見出し生成タスクの見直し. 情報処理学会 自然言語処理研究会 (NL), 2019 年 6 月.

[6] 長谷川駿, 上垣外英剛, 奥村学. 生成型文要約のための抽出性に着目したデータ選択. 情報処理学会 自然言語処理研究会 (NL), 2019 年 8 月.

[7] Francis Grégoire and Philippe Langlais. Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation. In *COLING 2018*.

[8] Pengda Qin, Weiran Xu, and William Yang Wang. Robust distant supervision relation extraction via deep reinforcement learning. In *ACL 2018*.

[9] Richárd Csáky, Patrik Purgai, and Gábor Recski. Improving neural conversational models with entropy-based data filtering. In *ACL 2019*.

[10] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *EMNLP 2016*.

[11] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *ACL 2016*.

[12] Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *ECML 2004*.