2T-07

アテンション付きエンコーダデコーダによる複数言い換えツイートからの要約文生成

永塚　光一‡　　渥美　雅保‡

創価大学理工学部情報システム工学科‡

1. はじめに

自然言語処理の分野において，生成的要約は重要なタスクの一つであり，アテンションを取り入れたエンコーダデコーダモデル[1]による研究が盛んに行われている．特に，Rush[2]による単文要約や，Paulus[3]による複文要約の研究などにおいて，大きな性能の向上が報告されている．本論文では，ネット上のニュース記事の URL を，ツイッター上で共有する複数のツイートから，その要約を生成するような複文要約モデルを提案する．学習に用いたデータセット[4]には，各ツイートを教師信号の文に対する言い換え表現とみなした場合の，意味的類似度がラベル付けされている．本研究では，アテンションを取り入れることにより，意味的類似度の低いツイートを除去した場合と，事前にツイートを除去しない場合において，性能にどのような影響があるのかを検証する．

2. ツイート要約システム

図 1 に，本論で提案するツイート要約システムの構成を示す．ツイッターにおける要約生成を想定する時，あるトピックの元ツイートを教師ツイートとして，そのトピックを共有する複数の派生ツイートから，教師ツイートを予測するような複文要約タスクが考えられる．その際，同一の URL を含むツイートを，あるトピックを共有する言い換えツイートと仮定すれば，URL を含む教師ツイートを基に，twitter API を用いて，自動的にデータセットを構築することができる．しかし，この手法は簡易である一方，データセットに教師ツイートとの関連度の低いツイートが多く含まれることが予想される．そのため，要約学習モデルの構築においては，学習モデルが人手を介さずに，ツイート文のトピックに関連する部分にのみに着目し，要約を行えるかどうかを検証することが課題となる． 本システムでは，この問題を既存手法であるアテンションで解決できるという仮説のもと，各ツイートに割り当てられた教師ツイートに対する意味的類似度ラベルに応じて，人手によるツイート除去処理を介在させる場合とそうでない場合の比較実験を行う．

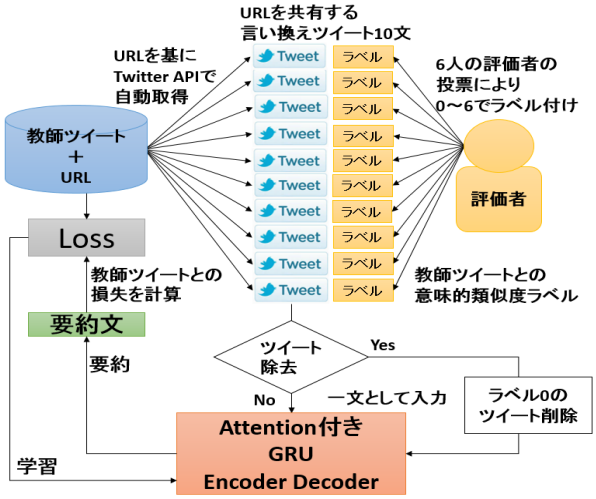

図 1．ツイート要約システムの構成

3. データセット

モデルの学習には，Lan ら[4]によって作成された Twitter URL Paraphrase Corpus を使用する．このコーパスは，教師ツイート 1 文とそれに対する複数の言い換えツイートを 1 ペアとして，5801 ペアから構成される．本論では，この内公開されている 4667 ペアの中から，言い換えツイートの数が 10 文に正規化されている 4441 ペアを使用する．教師ツイートは，CNN や New York Times, BBC などのニュース機関の公式アカウントから投稿されたもので，ニュース記事サイトへの URL が貼られている．また，10 文の言い換えツイートには，教師ツイートとのパラフレーズとしての意味的類似度がラベル付けされている．この意味的類似度の決定においては，6 人の評価者により，各言い換えツイートが教師ツイートに対するパラフレーズと言えるかどうかの投票が行われ，最小で 0，最大で 6 の 7 段階でラベル付けがされている．表 1 は本論で用いるデータセットにおける言い換えツイートのラベルの内訳を示したものである．表 1 より，データセットの言い換えツイートの 7 割以上が類似度 3 以下，すなわち評価者の過半数が言い換え表現とは認めないツイートで構成されていることが分かる．これら類似度 3 以下の言い換

Summary Generation from Multiple Paraphrase Tweets based on an Attentional Recurrent Encoder-decoder Model -

†Koichi Nagatsuka, Masayasu Atsumi

Department of Information Systems Science, Faculty of Science and Engineering, Soka University

えツイートは,文単位では,重要度が低いものの,アテンションにより,単語やフレーズレベルでは有用な情報を,学習モデルに与えるものとみなされる.

表 1. 意味的類似度ラベルに基づく言い換えツイートの構成

ラベル	ツイート数	%
0	16495	37.14
1	7670	17.27
2	5278	11.88
3	4264	9.60
4	3850	8.67
5	3593	8.09
6	3260	7.34
計	44410	-

4. 実験

実験では,意味的類似度 0 のラベルを削除した場合のデータセットと,ツイートの削除をしない場合のデータセットを用意し,同じ条件下で学習をさせた時の性能を比較する.尚,データセットには,言い換えツイートのラベルが全て 0 のものも含まれるため,10 文のうち,少なくとも 5 文はラベルに関係なく入力データとして残すように処理している.結果的に,ラベルの削除を行ったデータセットでは,言い換えツイート 44410 文からラベル 0 の言い換えツイート 13883 文が削除された.トレーニングデータセットには,4041 ペア,テストデータセットには 400 ペアを用いる.入力は複数の言い換えツイートをまとめて一文とした.文の最大の長さは 213 語である.評価指標としては,要約タスクに用いられる ROUGE を採用し,二つの 4041 ペアのトレーニングデータセットに対してランダムに 50000 回の学習を行った.

5. 結果・考察

表 2 に二つのデータセットに対する ROUGE スコアを示す.ROUGE スコアは,学習中の 500 イテレーション毎にテストデータセットに対して算出しており,最も性能の良かったスコアを採用している.尚,どちらのデータセットも 40000 回から 50000 回の学習において収束した.表 2 からも分かるように,全ての指標においてラベル 0 の言い換えツイートを削除した場合よりも,削除しなかった場合の方が,およそ 4〜5 ポイント高い性能を示した.この実験結果の原因として,類似度ラベル 0 の言い換えツイートからも,モデルが何らかの情報を学習していた可能性が考えられる.本実験において,アテンションは単語レベルで取られており,類似度ラベルが 0 のツイートであっても,文中にキーワードやキーフレーズを含むような場合がこれに当たる.また,アテンションメカニズムはノイズ除去を学習するため,あえて関連性の低い文を含ませた方が,学習モデルとしてはロバストになることが可能性として挙げられる.結果的に,今回のデータセットにおいては,人手によるラベル付けと,事前のツイート削除は,省略することができるだけではなく,むしろ人間の評価者が介在しない方が,学習にとっては良い影響を与えると言える結果となった.しかしながら,既に述べたように,今回用いたデータセットでは結論に達するには十分とは言えないため,今後データセットを増加させて同様の実験を行い,仮説の検証を試みる必要がある.

表 2. 各データセットに対する ROUGE スコア

データセット	ROUGE-1	ROUGE-2	ROUGE-L
事前削除あり	20.58	9.28	20.10
削除なし	25.28	14.60	24.41

6. むすび

本研究では,アテンション機能を取り入れた要約モデルに対して,人間の評価者により意味的類似度が低いと判定されたツイートを事前に除去して学習させた場合よりも,除去しないで学習させた場合の方が,各 ROUGE スコアにおいて高い性能を出すことを示した.但し,現状のデータセットのサイズは大きくなく,また,出力された要約文には文法の間違いや,意味を為さないものが多いので,今後モデルの改良や,データセットを増加させて検証を試みる必要がある.

参考文献

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio: Neural machine translation by jointly learning to align and translate, *arXiv preprint arXiv:1409.0473*, 2014.

[2] Alexander M. Rush, Sumit Chopra, Jason Weston: A Neural Attention Model for Abstractive Sentence Summarization, *arXiv preprint arXiv:1509.00685*, 2015.

[3] Romain Paulus, Caiming Xiong, Richard Socher: A Deep Reinforced Model for Abstractive Summarization, *arXiv preprint arXiv:1705.04304*, 2017.

[4] Wuwei Lan, Siyu Qiu, Hua He, Wei Xu: A Continuously Growing Dataset of Sentential Paraphrases, *arXiv preprint arXiv:1708.00391*, 2017.