6Q-02

深層学習を用いた論文タイトル生成における未知語処理について

大部　達也† 大囿　忠親† 新谷　虎松†
†名古屋工業大学大学院情報工学専攻

1 はじめに

論文執筆において，論文のタイトルは主張点を的確に表す必要があり，タイトルの決定は重要である．初学者が論文のタイトルを作成するためには，タイトル候補の提示などの支援が好ましい．本研究では，論文のアブストラクトからタイトル候補を生成し，論文執筆の支援を行うシステムの実現を目的としている．論文のタイトルの単語はアブストラクトに含まれていると仮定すると，論文のアブストラクトからタイトルを生成するタスクは，文書要約の一種として考えられる．アブストラクトからタイトルを生成する際に辞書に登録されていない未知語の扱いが問題となる．本稿では，既存の深層学習を用いた文書要約手法と未知語の扱いの，論文タイトル生成への応用についての実験と考察を述べる．

2 文書要約と論文タイトル生成

深層学習を用いた文書要約手法として，Encoder-Decoder 翻訳モデルおよび Attention 機構を利用した手法がある [1]．Encoder-Decoder 翻訳モデルでは学習時に作成した辞書から次に生成する語を決めるため，未知語の扱いが問題となる．一般には，原文における未知語を事前に未知語用のダミートークンに置き換えることで解決する．しかし，文書要約タスクにおいては，原文における未知語が要約に重要となるキーワードである場合が多い．そのため，ダミートークンを用いて文書要約を行うと，ダミートークンを多く含む不適切な要約文となる．この問題を解決する手法として辞書から単語の生成を行うのではなく，原文中の単語を直接指す手法がある．Vinyals らは Attention 機構を応用して Encoder 側の要素を指し示す手法 (Pointer Networks) を提案している [2]．また，Gulcehre らは辞書からの単語の生成に加えて，Pointer Networks による原文中の単語を指す手法を併用した，未知語に頑健な Encoder-Decoder 翻訳モデルを提案している [3]．

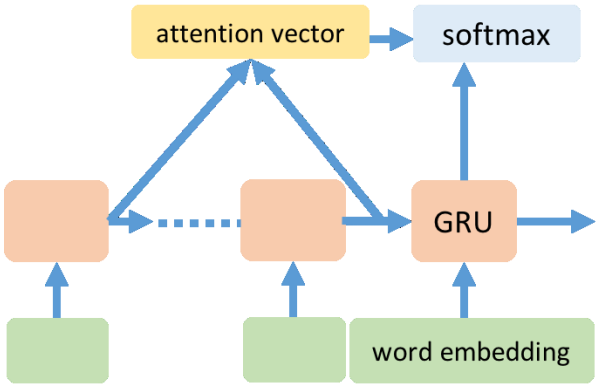

図 1: 抽出型手法のネットワーク

通常の文書要約と論文タイトル生成の違いについて述べる．文書要約は大量の文書を処理するために用いられ，要約文の可読性も重要となる．論文タイトル生成では一度に大量の論文を扱うことはなく，人手による修正も考慮することができるので，必ずしも文法的に正しい必要はない．また，論文タイトルには定型的な言い回しが存在するため，テンプレートを利用することができる．要約長の点でも文書要約と論文タイトル生成では異なる．通常の文書要約は目的ごとに要約長の制限の有無および制限する長さが異なる．しかし，論文のタイトルは一般に数ワードから 10 数ワード程度の長さであるため，常に要約長の制限は一定である．

3 論文タイトル生成における未知語の扱い

文書要約には抽出型の要約と生成型の要約がある．抽出型の要約は，原文中から文書単位 (単語，文など) レベルで抽出することで要約文を生成する．生成型の要約は，辞書から単語を選ぶことで要約文を生成する．本研究では，抽出型と生成型の両手法について論文タイトル生成を試みる．

抽出型の手法では，原文から単語の抽出を行うので，未知語の存在を考慮する必要はない．アブストラクト中の各単語をタイトルに出現するか出現しないかの二値に分類する．抽出した単語とテンプレートを併用し

Processing Unknown Words in Generating Titles Using Deep Learning
Tatsuya OHBE†, Tadachika OZONO‡and Toramatsu SHINTANI‡
†Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology.

Copyright ©2017 Information Processing Society of Japan.
All Rights Reserved.

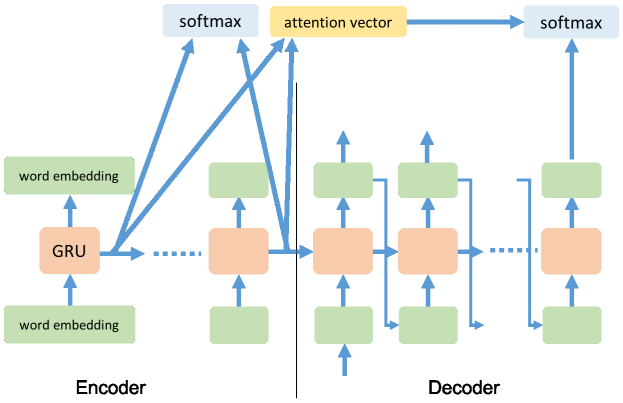

図 2: 生成型手法のネットワーク

て論文タイトルを生成する．図 1 に抽出型のネットワークを示す．アブストラクトの各単語について Encoder によりそれまでの単語をエンコードする．エンコードされたベクトルに対して Softmax 関数を適用し，誤差の計算を行う．生成型の手法では，Encoder-Decoder 翻訳モデルを用いてアブストラクトからタイトルを生成する．また，Gulcehre らと同様に Pointer を用いてアブストラクト中の単語を指す．抽出型および生成型の両手法で，Encoder および Decoder のユニットには Gated Recurrent Unit (GRU) を用いる．誤差関数としてはクロスエントロピーを用いる．

4 実験・考察

本研究では，Confer API[1] から取得可能な論文のタイトルとアブストラクト 3,208 件を用いて各手法の性能を検証した．テストデータには学習データに含まれない 337 件の論文を使用した．抽出型のキーワード抽出については精度および再現率を検証し，生成型のタイトル生成については BLEU により性能を検証した．生成されたタイトルが 3 ワード以下もしくは 21 ワード以上の場合は論文のタイトルとして不適切と考え，除外する．生成をされたタイトルが 4 ワード以上かつ 20 ワード以下であった 325 件の論文を用いた．

それぞれの手法について 50 エポック学習したモデルを用いた結果について表 1 に示す．抽出型の手法では，精度，再現率ともに低い値となったが，論文のタイトル生成を支援するためには再現率が重要である．そのため，今後再現率の向上が望まれる．生成型の手法では，通常の Decoder による生成のほうが Pointer による生成より BLEU において低い値となった．これは，Decoder では未知語を扱うことができないからと考えられる．一方，Pointer により生成したタイトルには同じ単語が連続して現れる場合がしばしばみられた．こ

表 1: 各手法の実験結果

キーワード抽出 精度	0.23
キーワード抽出 再現率	0.27
Decoder タイトル生成 BLEU	0.17
Pointer タイトル生成 BLEU	0.34

のことから，Gulcehre ら同様に Decoder と Pointer を併用することが必要である．本研究ではまた，表 1 の結果から Pointer の出力の一部を Decoder の出力に置き換えることを検討した．論文タイトルには代名詞やカンマなど，出現しにくい品詞の語があるが，Pointer によるタイトル生成はアブストラクトから単語を選択するため，出現しにくい品詞の語が現れる．このことから，Pointer の出力の品詞に応じて Decoder の出力と切り替えることを試したが，性能は向上しなかった．この問題を解決する手段として，タイトルに出現する単語や品詞の頻度を学習時に計算しておき，タイトル生成時に使用する手段が考えられる．

5 おわりに

本稿では，論文のアブストラクトからタイトルを生成する際の未知語の扱いについて述べた．本研究では，アブストラクト中の未知語を扱うために，辞書を必要としない抽出型の手法および，生成型の手法において原文の単語を指す手法についてタイトル生成の性能を検証した．実験では，既存の手法では十分にタイトル生成ができないことが示された．今後は，論文タイトル生成に対応するための既存手法の改良が必要である．

参考文献

[1] RUSH, Alexander M.; CHOPRA, Sumit; WESTON, Jason. A neural attention model for abstractive sentence summarization. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp.379–389. 2015.

[2] VINYALS, Oriol; FORTUNATO, Meire; JAITLY, Navdeep. Pointer networks. In: Advances in Neural Information Processing Systems. pp.2692–2700. 2015.

[3] GULCEHRE, Caglar, et al. Pointing the Unknown Words. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics .Vol1 . pp.140–149. 2016.

[1] http://confer.csail.mit.edu/developer

Copyright ©2017 Information Processing Society of Japan.
All Rights Reserved.