3A1-1

Recurrent Neural Network を用いた抽出型および生成型論文タイトル生成について

On Extractive or Abstractive Title Generation Using Recurrent Neural Networks

大部達也 *1　大圃忠親 *1　新谷虎松 *1
Tatsuya Ohbe　Tadachika Ozono　Toramatsu Shintani

*1名古屋工業大学大学院情報工学専攻
Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

Making proper titles are difficult for novices in composing papers because titles should contain key points in limited length. Text summarization can be used to generate titles; however, we need much more works to adapt existing text summarization methods to title generation tasks. We applied text summarization methods based on Recurrent Neural Networks (RNN) to title generation tasks, and we tried to develop an extractive and an abstractive title generation based on RNN. The extractive method is based on key word extraction using RNN, and the abstractive method is based on Attention-based Encoder-Decoder language model. Moreover, we introduced Pointer Networks to treat unknown words in abstracts. The paper described the designs and experiments of the methods.

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．公演発表申し込み時には論文のタイトルが必要となるが，論文の本文が執筆中である可能性があることから，アブストラクトを用いてタイトル決定の支援を行う．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．

本稿では，通常の文書要約と論文タイトル生成の違いについて説明し，Recurrent Neural Network(RNN) を用いた論文タイトル生成について述べる．文書要約同様に生成型および抽出型の各手法による論文タイトル生成について説明し，各手法の性能について実験を行った結果と考察について述べる．

2. 文書要約と論文タイトル生成

本研究では，論文タイトルの生成について検討する上で，まず，論文タイトルの生成を，特殊な形式の要約であると考えた．本章では，文書要約手法について概説し，次に文書要約と論文タイトル生成の差異について述べる．

深層学習を用いた文書要約手法として，機械翻訳に用いられている Encoder-Decoder 言語モデルの応用が成果をあげている [Rush 15]．Encoder-Decoder 言語モデルでは，入力となる文を単語に分割して順番に RNN に入力することで，文を一つのベクトルに Encode し，Encode された文のベクトルから Decoder を用いて次に出現する確率の高い単語を辞書から生成する．また，Encoder-Decoder 言語モデルでは，Encoder 側の各時刻の隠れ層を保存しておき Decode 時に使用する Attention 機構を導入することにより性能が上がることが知られている [Luong 15]．論文タイトル生成において，入力はアブストラクトであり 100〜300 ワード程度となるため，長期的な依存関係がより重要となる．また，入力となるアブストラクトと出力のタイトルの文長が大きく異なる．これらの要因から論文タイトル生成においても Attention 機構が重要であると考えられる．

Encoder-Decoder 言語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．また，低頻度による辞書のサイズの増大も問題であるため，高頻度語以外を辞書に登録しない場合もある [Jean 15]．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になる場合がある．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 [Vinyals 15]．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 言語モデルを提案している [Gulcehre 16]．論文のタイトルはキーワードが占める割合が多く，論文の著者による造語なども用いられる可能性があるため，未知語の扱いがより重要となる．

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と考えられるため，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．通常の文書要約は要約長の制限によっては原文の内容の一部のみを用いるが，論文タイトルにおいては論文の主張点を過不足なく含む必要がある．

連絡先: 大部達也，名古屋工業大学大学院工学研究科情報工学専攻，愛知県名古屋市昭和区御器所町，052(735)5584，tobe@toralab.org

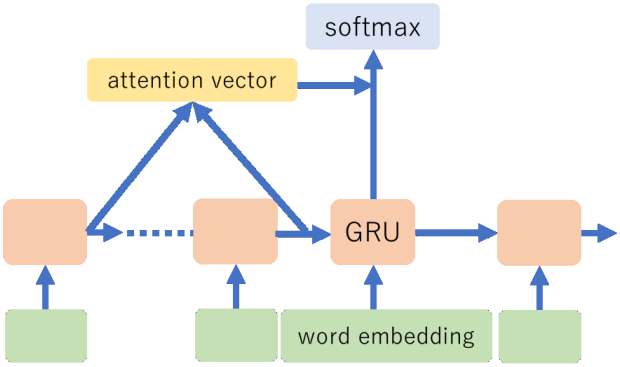

図 1: 抽出型手法のネットワーク

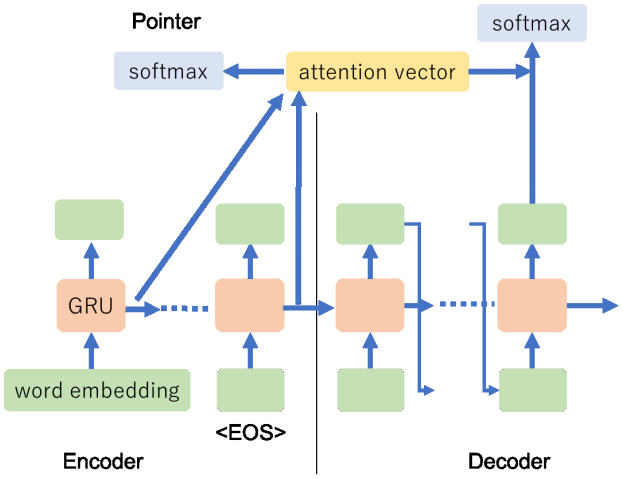

図 2: 生成型手法のネットワーク

3. 生成型および抽出型論文タイトル生成

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，RNN を用いて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．抽出型および生成型の両手法で，RNN のユニットには Gated Recurrent Unit (GRU) を用いる．学習時の誤差関数としてはクロスエントロピーを用いる．

3.1 抽出型タイトル生成

抽出型の手法では，アブストラクトから抽出した単語とテンプレートを併用して論文タイトルを生成する．本稿では特に単語の抽出について説明する．抽出型手法では，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．RNN により単語の情報に加えて文脈の情報が付与されるため，適切な分類が可能になると考えられる．また，抽出型手法では，アブストラクト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．図 1 に抽出型の RNN のネットワークを示す．まず，タイトルおよびアブストラクトを TreeTagger を用いて形態素解析し，辞書を作成する．この際，文頭に出現する単語などを同一の単語として扱うために，すべての単語を小文字にする．作成した辞書を用いて，単語の ID から one-hot-vector を作成し，単語ベクトルとする．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 Encode する．Encode されたベクトルに対して Softmax 関数を適用することでタイトルへの出現の有無を分類する．時刻 t における抽出手法の出力 $E(t)$ は，時刻 t における隠れ層 $\boldsymbol{h}_t$ および Attention ベクトル $\boldsymbol{c}_t$ を用いて以下の式で示される．

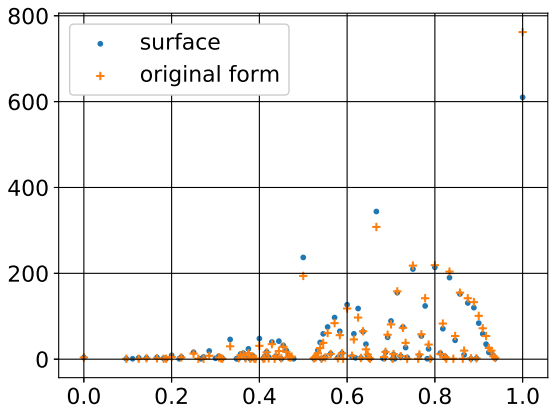

図 3: タイトル中の語の出現頻度

$$E(t)\{0,1\} = argmax(Softmax(\boldsymbol{W}_s \boldsymbol{h'}_t))$$
$$\boldsymbol{h'}_t = tanh(\boldsymbol{W}_{c1}\boldsymbol{h}_t + \boldsymbol{W}_{c2}\boldsymbol{c}_t)$$

ここで，$\boldsymbol{W}_s$， $\boldsymbol{W}_{c1}$ および $\boldsymbol{W}_{c2}$ は重み行列をあらわす．抽出型手法の時刻 t における Attention ベクトル $\boldsymbol{c}_t$ は，それまでの各時刻の隠れ層 $\boldsymbol{h}_i$ を用いて以下の式で示される．

$$\boldsymbol{c}_t = \frac{\sum_{i=1}^{t} \boldsymbol{h}_i * exp(dot(\boldsymbol{h}_i, \boldsymbol{h}_t))}{\sum_{i=1}^{t} exp(dot(\boldsymbol{h}_i, \boldsymbol{h}_t))}$$

3.2 生成型タイトル生成

生成型の手法では，Encoder-Decoder 言語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．図 2 に生成型の RNN のネットワークを示す．アブストラクトの各単語を Encoder により Encode し，終端記号が入力されると Decode を開始する．Decoder では入力された単語と一つ前の時刻の隠れ層から次に生成する単語を決定する．時刻 t における Decoder の出力 $D(t)$ は，時刻 t における隠れ層 $\boldsymbol{h}_t$ および Attention ベクトル $\boldsymbol{c}_t$ を用いて抽出型手法と同様に示される．生成型手法の時刻 t における Attention ベクトル $\boldsymbol{c}_t$ は，アブストラクトの単語数を n とすると Encode 時の各時刻の隠れ層 $\boldsymbol{h}_i$ を用いて以下の式で示される．

$$\boldsymbol{c}_t = \frac{\sum_{i=1}^{n} \boldsymbol{h}_i * exp(dot(\boldsymbol{h}_i, \boldsymbol{h}_t))}{\sum_{i=1}^{n} exp(dot(\boldsymbol{h}_i, \boldsymbol{h}_t))}$$

また，アブストラクト中の未知語に対応するために，Pointer を用いてアブストラクト中の単語を指す．Pointer においては，Attention ベクトルに対して直接 Softmax 関数を適用することで Encode 時のある一時刻，すなわちアブストラクトの一単語を指す．時刻 t における Pointer の出力 $P(t)$ は以下の式で示される．

$$P(t) = argmax(Softmax(\boldsymbol{c_t}))$$

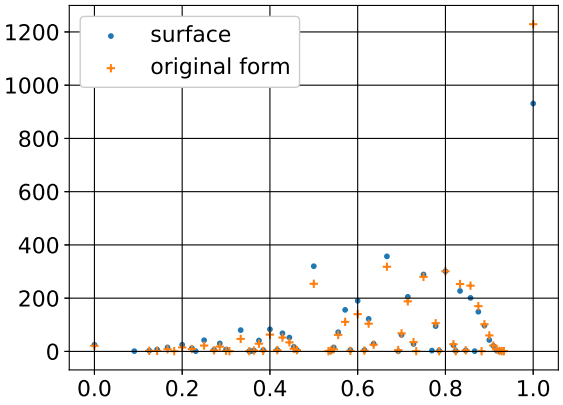

図 4: タイトル中のキーワードの出現頻度

表 1: キーワードとして扱う品詞

品詞	POS タグ
名詞	“NN”, “NNS”, “NP”, “NPS”
一般動詞	“VV”, “VVD”, “VVG”, “VVN”, “VVP”, “VVZ”
形容詞	“JJ”, “JJR”, “JJS”
副詞	“”RB”, “RBR”, “RBS”
その他	“FW”, “POS”, “RP”

4. 実験・考察

4.1 データセット

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を用いて各手法の性能を検証した．cc2015，chi2014，chi2015，cscw2014，cscw2015，cscw2017，dis2014，ijcai2016，kdd2014，kdd2016，nordichi2016，recsys2016，sigmod2016，wsdm2014 から論文のタイトルとアブストラクト 3,208 件を学習データに用いたテストデータとしては学習データに含まれない coling2016 から 337 件の論文を使用した．Encoder 側では論文のアブストラクトから得られた辞書，Decoder 側ではタイトルから得られた辞書を用いる．学習データのアブストラクトから得られた辞書の語彙数は 22,822 語で，タイトルから得られた辞書の語彙数は 6,640 語であった．

図 3 にタイトル中の語のアブストラクト中の出現頻度を示す．横軸が出現頻度，縦軸が論文の数をあらわす．丸の点は単語の表層をそのまま用いた場合の出現頻度をあらわし，十字の点は単語の原形を用いた場合の出現頻度をあらわす．本研究では論文の主張点を表す単語を論文のキーワードと呼び，表 1 に示す品詞の単語をキーワードとして扱う．論文タイトルにおいてはキーワードが重要であることから，品詞情報を元に，キーワードに着目した出現頻度を調べた．図 4 にタイトル中のキーワードのアブストラクト中の出現頻度を示す．原形が同じ語であってもタイトル中とアブストラクト中では異なる活用形で現れる語が存在する．そのため，原形を用いた場合のほうがタイトルの単語がアブストラクトに出現する論文の数が増加している．タイトルの単語がすべてアブストラクトに出現する論文は約 600 件で全体の 1/5 程度となっている．また，タイトルのキーワードがすべてアブストラクトに出現する論文は，原形についてみた場合は約 1,200 件で全体の 1/3 以上となっている．

*1 http://confer.csail.mit.edu/developer

表 2: 各手法の実験結果

単語抽出 精度	0.23
単語抽出 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

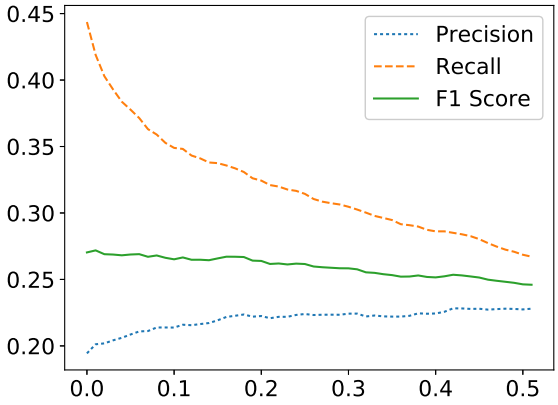

図 5: 単語抽出の閾値と精度および再現率

4.2 各手法の性能評価

抽出型の手法については，抽出された単語と正しいタイトルに含まれる単語から精度および再現率を検証した．生成型の手法については，生成されたタイトルと正しいタイトルから，機械翻訳および文書要約に用いられる評価指標である 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．また，Pointer によるタイトル生成は，各時刻ごとにアブストラクト中から単語が生成されるため，同じ単語が続いて出力される場合がある．本研究では，同じ単語が出力された場合は最初の単語のみ使用し，残りの単語は省略する．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 2 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．

表 3 に抽出型による単語抽出の成功例を示す．この例では，“non-native”，“speakers”や “simplification” といったタイトル中のキーワードを抽出できている．一方で，キーワードの一つである “lexical” は抽出されておらず，また，タイトルにはない “semantic” 等の語も抽出されている．このことから，単語抽出は論文タイトル生成における支援になるが，十分ではないといえる．表 4 に生成型によるタイトル生成の成功例を示す．Pointer によるタイトル生成では，タイトルに含まれにくいピリオドなども出力されるため，不必要な品詞の単語は除外すべきである．Decoder によるタイトル生成では，タイトルらしい文は生成されるが，入力のアブストラクトとは関係のな

表 3: 単語抽出例

タイトル	抽出された単語
Understanding the Lexical Simplification needs of Non-native Speakers of English	'text', 'needs', 'non-native', 'on', 'simplification', 'speakers', 'and', 'for', 'of', 'semantic', ',', 'the'

表 4: タイトル生成例

タイトル	生成されたタイトル
Two-view Label Propagation to Semi-supervised Reader Emotion Classification	'the', 'a', 'semi-supervised', 'propagation', 'reader', 'to', 'propagation', '.', 'in', 'classification', 'approaches'

い単語が出力されやすかった．

表 2 の結果から各手法の改良を検討する．単語抽出では，人手による修正を含むため，精度よりも再現率が重要である．通常は Softmax 関数の出力の内，最大となる要素からカテゴリを決定する．再現率を上げるために，タイトルに含まれる語である確率が一定の値以上の場合に抽出することを考える．図 5 に閾値を 0.01 から 0.5 まで 0.01 間隔で変化させた場合の精度および再現率の変化を示す．図の点線が精度，破線が再現率，実線が F1 スコアをそれぞれあらわし，縦軸がそれらの値を，横軸が閾値をあらわす．閾値を下げた場合，精度の低下は少ないが再現率は向上し，F1 スコアが向上していることがわかる．このことから，Softmax 関数の出力の最大値から分類するよりも，閾値を設けてタイトルに含まれる確率が一定の値以上の単語を抽出することが有効であるといえる．生成型の手法では Pointer による生成の性能が高かったことから，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．アブストラクトを一度に一つのベクトルにエンコードすることは，一般的な Encoder-Decoder 言語モデルよりも長い系列を入力することになる．このことから，アブストラクトを一文ずつ処理したり，Recursive Neural Network などの別の手法も検討すべきである．

論文タイトルは分野によって特徴が異なる．例えば，疑問形のタイトルやサブタイトルを含むタイトルが使用されやすい分野がある．実験では，テストデータの論文は学習データの論文とは異なる会議の論文を用いた．学習時に使用する論文を同じ分野の論文に絞りテストデータも同じ分野の論文を使うことで，分野ごとのタイトルの特徴に対応しやすくなるため性能は向上すると考えられる．しかしこの場合，複数の分野についてそれぞれ学習したモデルが必要になる．また，タイトル生成を行う際には，事前に論文の分野を特定するための分類が必要になる．実験では，品詞が変わる語や同じ意味を置き換えた語については別々の語として扱っているため，これらを同一の語として扱うことができれば性能が向上すると考えられる．また，単語ベクトルを one-hot vector としているが，事前に学習した単語ベクトルの使用により性能が向上すると考えられる．

5. おわりに

本稿では，通常の文書要約と論文タイトル生成の違いについて説明し，RNN を用いて論文のアブストラクトからタイトルを生成する手法について述べた．RNN を用いた抽出型の手法および Encoder-Decoder 言語モデルを応用した生成型の手法を提案し，それぞれの手法の性能について実験を行い考察した．生成型の手法では通常の Decoder に加えて Pointer Network を用いた未知語に対応可能なタイトル生成について述べた．実験では，生成型の手法では Pointer による生成が Decoder による生成よりも優れていることが示された．これにより，今後は Pointer を主軸として Decoder を併用することで性能の向上が見込まれる．また，抽出型の手法では単語の抽出の性能が十分ではないが，タイトルに含まれる確率が一定以上の単語を抽出することで再現率が向上した．今後は両手法において，品詞情報や事前学習した単語ベクトルを用いるなどして性能の向上が望まれる．

参考文献

[Gulcehre 16] Gulcehre, C., Ahn, S., Nallapati, R., Zhou, B., and Bengio, Y.: Pointing the Unknown Words, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 140–149 (2016)

[Jean 15] Jean, S., Cho, K., Memisevic, R., and Bengio, Y.: On Using Very Large Target Vocabulary for Neural Machine Translation, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 1–10 (2015)

[Luong 15] Luong, T., Pham, H., and Manning, C. D.: Effective Approaches to Attention-based Neural Machine Translation, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421 (2015)

[Rush 15] Rush, A. M., Chopra, S., and Weston, J.: A Neural Attention Model for Abstractive Sentence Summarization, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389 (2015)

[Vinyals 15] Vinyals, O., Fortunato, M., and Jaitly, N.: Pointer Networks, in Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. eds., *Advances in Neural Information Processing Systems 28*, pp. 2692–2700 (2015)