専門用語の活用による学術論文の生成的要約手法

梁 燦彬† 前田 亮‡

† 立命館大学情報理工学研究科 〒525-8577 滋賀県草津市野路東 1-1-1
‡ 立命館大学情報理工学部 〒525-8577 滋賀県草津市野路東 1-1-1
E-mail: † gr0319ef@ed.ritsumei.ac.jp, ‡ amaeda@is.ritsumei.ac.jp

あらまし 本研究では，深層学習を用いて学術論文の自動要約システムを構築することを目指す．インターネットの情報の増加に伴い，情報収集の労力も増加し続けている．研究者にとって，研究を行いながら論文をサーベイすることが重要であり，論文のサーベイは大量の論文を読む必要がある．論文のアブストラクトだけを読んでも，多くの時間がかかり，効率が良いとは言えない．論文を研究対象とする自動要約の既存研究は少なくないが，論文のサーベイという点から考えた要約システムはほとんど存在しない．本研究では，英語で書かれた学術論文を研究対象として，系列変換モデルとアテンションメカニズムをベースして専門用語を活用した要約手法を提案する．

キーワード 生成的要約，深層学習，再帰型ニューラルネットワーク（RNN），アテンションメカニズム

1. はじめに

インターネットの急速な成長に伴い，情報の過負荷の問題が深刻化し，データの量が増加するにつれて，情報収集の労力が増加し続けており，自動要約も重要になり，人間が膨大なテキスト文書を手作業で要約することは非常に困難となっている．インターネット上のあらゆるトピックについて情報が豊富に入手できるが，重要な情報を要約文の形で提示すれば、多くの使用者に利益がもたらされる．

ニュース記事や小説など様々なものを研究対象とする自動要約の研究が多くあるが，学術論文がその一つとして挙げられる．論文にはそれにアブストラクトやあらましが中に含まれており，それらは一般に簡潔であり，本文に対する参照要約と生成された要約文を比較することで容易に結果の評価が行うことができる．しかし，サーベイのために論文を読む場合は，論文の内容をより深く理解するために，論文そのものの要約が必要とされることも多い．また，論文データベースから大量に用意することが可能であるので機械学習に適していることが理由である．さらに，学術論文の場合は各セクションごとに役割が異なり，セクションごとに解析することで,要約システムの負担も軽くなる．

近年，ニューラルネットワーク技術が急速に発展してきており，自動要約の研究でもモデルを組み合わせた生成的要約モデルが活発に研究されている．論文のサーベイでは，ユーザは論文のトピックに関心を持つと考えられる．特にコンピュータ科学系の論文をサーベイするときには，研究でどのような技術や手法が利用されたかに強い関心を持つ．要約文に専門用語多く含まれていると，論文のサーベイに良い効果を与えると考えられる．

そこで本研究では，系列変換モデルとアテンションメカニズムをベースして専門用語を活用した学術論文の自動要約手法を提案する．

2. 関連研究

近年自動要約の分野で，深層学習を用いた生成的要約モデルが提案されている．Rush ら[1]はニュース記事を研究対象として系列変換モデルおよびアテンションメカニズムを利用した手法を提案し，自動要約について系列変換モデルのようなモデルに基づいた研究も増えいる．Abigail ら[2] は系列変換モデルの不正確に繰り返しが多いという問題を考慮し，未知語を解析した生成的要約モデルを提案した．

2.1. 系列変換モデル

系列変換モデルモデルは Encoder と Decoder の 2 つの再帰型ニューラルネットワーク(RNN)で構成されている．Encoder のニューラルネットワークで入力系列をベクトルに圧縮し，そのベクトルを Decoder に渡し出力系列を生成する。したがって，系列変換モデルは Encoder-Decoder モデルという名前で呼ばれる場合もある．

再帰型ニューラルネットワークは，一般には隠れ層は全ての入力を考慮し，以前の情報を現在のタスクに結合できるが，実際には長期的な記憶は困難である．そこで，Sutskever ら[3]は RNN の変形の一つである LSTM（Long Short Term Memory ネットワーク）を用いた系列変換モデルを提案した．RNN よりも LSTM を用いたモデルの方が長期的な記憶が可能となり，ニュース記事のような長いテキスト文書を解析しても，精度がより下がりにくい．

入力側，つまり Decoder 側で入力テキストの各単語は，embedding されてベクトル表現に変換され，各変換行列が再帰型ニューラルネットワークに入力されて，入力テキスト単語集合$w = \{v_1, v_2, v_3, v_4, v_5, \dots, v_e\}$で対して，t 番目の文ベクトル$v_t$におけるで，隠れ要素$h_t$は以

下の式（1）ように積算で計算される．

$$h_t = \tanh(v_t + h_{t-1}) \quad (1)$$

ここで，h_{t-1}は t 番目の直前隠れ要素であり，これがv_tと足し合わされている．これを活性化関数に入力することで新たな隠れ要素h_tが得られる．

出力側，つまり Decoder 側では，隠れ要素の引き渡しは Encoder 側と同じであるが，Decoder 側においては，単語を出力するメカニズムが追加されている．図 1 にアテンションメカニズム付き系列変換モデル示す．

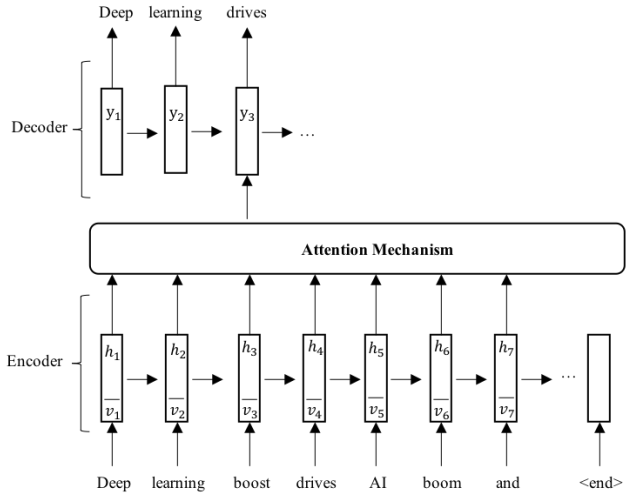

図 1 アテンションメカニズム付き系列変換モデル

2.2. アテンションメカニズム

アテンションメカニズムは Bahdanau ら[4]によって提案された手法である．単純な系列変換モデルは固定サイズのベクトルを使用するため，複雑な文書に対して表現できないことがある．そのため，入力テキスト文書の各単語の状態の情報を利用することで，より良い出力テキストを生成できるという考えに基づいている．これはアテンションメカニズムと呼ばれている．

アテンションメカニズムでは t 番目における Decoder の隠れ層のベクトル b_tと Encoder の全ての隠れ層のベクトル $\overline{b_s}$のスコアを計算し，t 番目の時にどの入力単語に注視するかのスコア α_tを決定するこのスコアをもとに Encoder の隠れ層のベクトルの加重平均 c_tを求めそれをもとに時刻 t の隠れ層のベクトルb'_tを計算する．以下の式（2）を用いて計算する

$$\alpha_t = \frac{\exp(score(b_t, \overline{b_s}))}{\sum_{s'} \exp(score(b_t, \overline{b_{s'}}))}$$

$$c_t = \sum_s \alpha_t(s)\, \overline{b_s}$$

$$b'_t = \tanh(W_c[c_t; b_t]) \quad (2)$$

アテンションメカニズムを用いた系列変換モデルでは，Encoder の隠れ層のうち，特定の入力単語やその周辺の単語に注視したベクトルを Decoder で用いる．これにより，Decoder のある時点で必要な情報に注視して使用することができ，入力文の長さに関係なく Decoder の出力を効率よく行うことができる．

3. 提案手法

系列変換モデルにアテンションメカニズムを追加して学習させると大幅に要約の精度を改善できるが，元々ユーザがコントールできずに，ベクトルを注視する仕組みである．また，専門用語には，複数の単語から構成される場合が多いため，専門用語が要約文に含まれにくい．本研究では，系列変換モデルとアテンションメカニズムをベースとして，アテンションメカニズムに専門用語を注視する機構を加えた要約手法を提案する．

コンピュータ科学系の学術論文を研究の実験対象にするため，コンピュータ科学系の辞書を利用し，embedding される前に前処理を行い，入力論文テキストの単語集合$w = \{v_1, v_2, v_3, v_4, v_5, \ldots, v_e\}$で最大 4-gram を用いることで，入力論文テキスト内の専門用語を得ることができる．さらに，専門用語である単語をタグする．図 2 に示した例のように，前処理で単語 Deep と単語 learning が Deep learning ような専門用語になる．また，同一のベクトル表現に変換される．

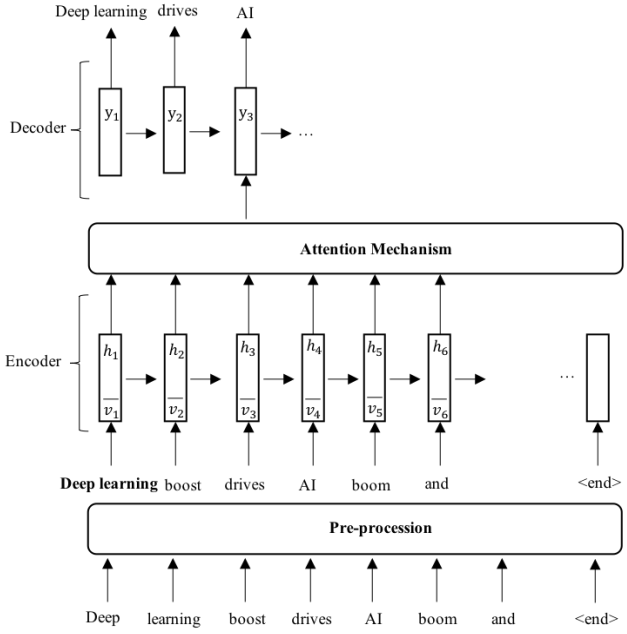

図 2 専門用語を活用したモデル

タグが付与されたベクトルの重みが高くなり，アテンションメカニズムによる重み付き和処理において計

算された中間表現の集合に対し，t 番目前の Decoder の隠れ状態を元にニューラルネットで算出した重み係数をかけ，重み付き和を導出する．

4. 実験

学術論文の自動要約タスクにおいて，提案手法によるモデルの性能を評価する．機械学習に関する 2016 年以降のコンピュータ科学系論文 40 件の実験データを手作業で用意した．

前処理で，Stanford Corenlp を用いて，入力テキストをトークン化する．2016 年に発行された Oxford Dictionary of Computer Science というコンピュータ科学の辞書を利用し，最大 4-gram を用いることで専門用語かどうか見分ける．辞書の専門用語数は約 6500 語である．辞書の一部の専門用語および n-gram の割合を表 1 に示す．

表 1 辞書の一部の専門用語および n-gram の割合

n-gram	専門用語	割合
1-gram	netiquette	42%
	octet	
	scheduler	
	…	
2-gram	egoless programming	53%
	Turing machine	
	zero function	
	…	
3-gram 以上	natural language processing	5%
	graphics processing unit	
	hazard and operability study	
	…	

深層学習ツール Tensorflow のライブラリを用いて，系列変換モデルとアテンションメカニズムのベースライン手法を実装する．提案手法によってベースライン手法のアテンションメカニズムを変形する．要約モデルに教師あり学習をさせる．本研究では，40 件の実験データを 8 対 2 の比率で訓練データとテストデータに分割し，3 回の交差検定を行って生成モデルの性能を検証する．

学術論文を研究対象とする場合に，ベースライン手法による要約モデルと提案手法による要約モデルを実験し，要約評価手法には ROUGE-N[5]用いて 2 つのモデルを比較する．ROUGE-N は要約モデルにおいて要約文書を生成した際に，参照要約と自動生成した要約の間で一致する n-gram 単位での割合を以下の式（3）を用いて計算する．

$$ROUGE(C,R) = \frac{\sum_{e \in n-gram(C)} Count_{match}(e)}{\sum_{e \in n-gram(R)} Count(e)} \quad (3)$$

*n-gram(C)*は自動生成した要約文書の n-gram，*n-gram(R)*は参照要約文書の n-gram を示す．*Count(e)*は n-gram の出現頻度を数える関数であり，$Count_{match}(e)$は，自動生成した要約における出現頻度 *Count(e ∈ n-gram(C))* と参照要約における出現頻度 *Count(e ∈ n-gram(R))*の小さいほうを採用する．

本研究では，本文のアブストラクトを参照要約として扱い，*n-gram* を 1 から 3 まで変化させ，生成モデルの検証を行う．

5. 実験結果と考察

ベースライン手法と提案手法の実験結果を図 3 に示す．提案手法はベースライン手法より ROUGE-N スコアが 1-gram で 4%，2-gram で 28%，3-gram で 21%上回った．

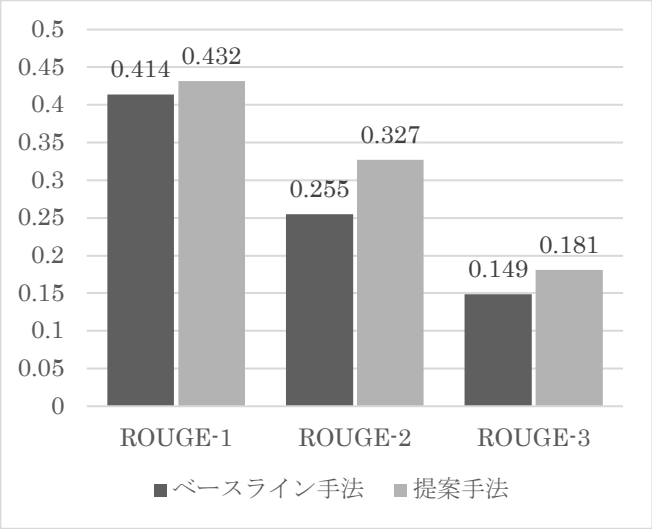

図 3　ベースライン手法と提案手法の実験結果

全体的に ROUGE-N のスコアが少し低いが，データ数が足りないと ROUGE-N という評価手法が生成的要約モデルに対して不利であるという理由がある．ベースライン手法に比べて，提案手法の場合にスコアが上がった．特に ROUGE-N の 2-gram と 3-gram で，スコアが大幅に向上した．そのため，要約文を生成する時に，提案手法のモデルが連続した複数の単語から構成された専門用語をより抽出しやすくなったと考えられる．

6. おわりに

本研究では，辞書を利用し，専門用語かどうか見分けることで，系列変換モデルとアテンションメカニズムをベースとして，アテンションメカニズムを変形し，専門用語を活用した要約手法を提案した．

研究対象は現時点ではコンピュータ科学系論文だ

けだが，今後の研究の方向性として，研究対象の幅を広ることを考えている．さらに，専門用語の活用だけではなくて，より論文のサーベイという点から考えた要約モデルを検討したい．

参 考 文 献

[1] A.M.Rush, S.Chopra and J.Weston, “A neural attention model for abstractive sentence summarization”, In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 379–389, 2015.

[2] Abigail S., P.J.Liu and Christopher D Manning, “Get To Tee Point: Summarization with Pointer-Generator Networks”, In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL’17), 2017.

[3] Sutskever I., Vinyals O. V. and Le Q., “Sequence to Sequence Learning with Neural Networks”, Proc. NIPS, 2014.

[4] Bahdanau, D., Cho, K., and Bengio Y, “Neural machine translation by jointly learning to align and translate”, arXiv preprint arXiv:1409.0473, 2014

[5] C.Y. Lin and E.H. Hovy, “Automatic evaluation of summaries using n-gram co-occurrence statistics”, In Proceedings of Human Language Technology Conference (HLT-NAACL 2003), 2003