ISS-A-053

# 文書構造と深層学習を利用した科学論文の自動要約生成

橋本　快生†　　井上　潮†
† 東京電機大学　工学部情報通信工学科

## 1. はじめに

自動要約は科学論文のような長大なテキストから短いテキストを生成することにより，人間が短時間で内容を把握できるようにする技術である．自動要約には元のテキストから重要な文を抽出する抽出型，テキストの内容を元に自由に作成する生成型が存在する．本研究では，図，表，式等，人間が理解するのに時間がかかる項目を利用せずに，多くの論文に共通して存在する構成内容を利用し，自由度の高い生成型要約方法を提案する．生成モデルとしては，アテンション構造に特化したネットワークモデル(Transformer)とリカレントニューラルネットワーク(LSTM+Attention)を利用したモデルを比較し，科学論文の要約分野に向いたモデルの検討を行う．

## 2. 関連研究

安部ら[1]は 575 の音韻を利用して読むべき論文であるかを判断するための要約生成方法を提案した．これは TF-TDF を利用し，単語の重要度から要約を生成する手法である．

## 3. データセット

利用するデータセットには機械学習のカンファレンスである NIPS で発表された英文の科学論文約 7000 件を利用する．本研究では本文テキストをデータとして用いる．

## 4. 提案手法

科学論文の読者の視点で見ると，要約は簡潔であり，内容が分かりやすいことが重要である．通常，論文には式，図等が記載されているが，これらは瞬時に理解することが難しく，要約には不向きである．そのため，式，図等が多く含まれる章は省き，多くの科学論文に共通して存在する章である Introduction(序論), Discussion(考察), Conclusion(まとめ)のテキストを抽出し，データとして用いる．これらの章には，考え方，研究結果等が述べられており，重要な文章が多く含まれていると考えられる．生成された要約の比較対象として Abstract(要旨)を用いる．これらのペアを学習させてモデルを作成する．要約生成を行う深層学習のネットワークモデルには文書タスクに用いられる Transformer と LSTM(Long short-term memory) +Attention を利用し，比較を行う．

## 5. 実験，評価

前処理を行ったデータを一つのテキストに統合し，Abstract とペアで学習を行う．利用するデータの割合は学習データ，テストデータを 8:2 とする．評価にはテキスト要約の指標である ROUGE-L の $F_{lcs}$ を利用する．この値は，モデルが作成した Abstract と参照元の Abstract の一致度を示す．

各モデルの評価結果を表 1 に示す．

表．1 評価結果

| モデル | $F_{lcs}$ (%) |
|---|---|
| Transformer | 34.3 |
| LSTM+Attention | 30.1 |

また、Transformer により生成された要約の一部を以下に示す．要約の対象とした論文は Thrun による『Learning To Play the Game of Chess』[2]である．

> last decade ago game major oldest research artificial intelligence computer science ches programme ches rely intensive search generate move board evaluate fast evaluation function employ usually carefully design hand sometime augment automatic parameter tune automatic method

## 6. まとめ

本稿では，科学論文の章項目を用いた要約の自動生成方法を述べた．実際に生成された要約は不適切な順序で単語が並んでいることから内容を完全に把握することが出来る状態ではなく，キーワードを抽出したような文章となった．今後は図題，表題等，重要なキーワードを持つ題目の利用，データ数の増加や前処理の改善を検討していく．

## 参考文献

[1] 安部文紀, 寺田実 "575 の音韻的読みやすさを付与した学術論文の要約文自動生成手法", DEIM Forum 2018 E3-1

[2] S. Thrun, "Learning to play the game of chess", Advances in Neural Information Processing Systems 7, pp.1069- 1076, 1995.