2S-01

# 章構造による学術論文からの Structured Abstract 自動生成

橋本　快生†　　井上　潮‡

東京電機大学大学院 工学研究科†　　東京電機大学 工学部‡

## 1. はじめに

研究者にとって学術論文を読むことは必須である．しかしながら，論文の量は膨大であり，増え続けている．この問題を緩和するため，医学系分野の論文では，Structured Abstract というアブストラクトの記述方法が採用されている．本研究では自動要約技術による，論文の章構造と深層学習を用いた論文の Structured Abstract の自動生成手法を提案する．具体的には，論文における章構造を示す Method，Result 等のテキストを活用し，Structured Abstract の生成を行う．論文の本文をそのまま利用した評価値と，章構造を利用した評価値を比較し，提案手法の有効性を確かめる．

## 2. 前提

### 2.1 Structured Abstract

Structured Abstract(構造化抄録)とは，Objective，Methods 等の見出しの形式で構造化された抄録である．各見出しが短文であるため，短時間で効率的に内容を把握することができる．

**Abstract**

*OBJECTIVES:* Remodeling of the extracellular matrix (ECM) is a key event in different lung disorders, such as fibrosis and cancer. The most common cell type in the connective tissue is fibroblasts, which transdifferentiate into myofibroblasts upon activation. All myofibroblasts express α-SMA, which has been found to be upregulated in lung fibrosis and cancer. We evaluated the potential of α-SMA as a noninvasive biomarker of activated fibroblasts in lung fibrosis and cancer. *METHODS:* A monoclonal antibody was raised against the N-terminal of α-SMA, and a novel competitive enzyme-linked immunosorbent assay (ELISA) measuring α-SMA was developed and technically characterized. Levels of α-SMA were measured in the fibroblast model, "scar-in-a-jar", and in serum from patients with idiopathic pulmonary fibrosis (IPF), chronic obstructive lung disorder (COPD) and non–small cell lung cancer (NSCLC) belonging to two different cohorts. *RESULTS:* The novel α-SMA assay was developed and validated as technically robust. Based on the scar-in-a-jar results, α-SMA was only present in the fibroblasts activated by TGF-β. In cohort 1, levels of α-SMA were significantly higher in IPF, COPD and NSCLC patients compared to healthy controls ($P = 0.04$, $P = 0.001$ and $P < 0.0001$, respectively). The area under the receiver operating characteristics (AUROC) for separation of healthy controls from IPF patients was 0.965, healthy controls from COPD patients was 0.892 and healthy controls from NSCLC patients was 0.983. In cohort 2, levels of α-SMA were also significantly higher in NSCLC patients compared to healthy controls ($P = 0$) and the AUROC for separating NSCLC and healthy controls was 0.715. *CONCLUSIONS:* In this study we developed and validated a robust competitive ELISA assay targeting the N-terminal of α-SMA. The level of α-SMA was upregulated when adding TGF-β, indicating that α-SMA is increased in activated fibroblasts. The level of α-SMA in circulation was significantly higher in patients with IPF, COPD and NSCLC compared to healthy controls. This assay could potentially be used as a novel noninvasive serological biomarker for lung disorders by providing a surrogate measure of activated fibroblasts.

図 1.　Structured Abstract の例
(Nielsen ら[1]による Abstract から引用)

### 2.2 先行研究

筆者らは，科学系に特化した BERT を用いた抽出的要約による手法で Structured Abstract の自動生成を行ってきた[1]。これまでの研究結果で科学系に特化した BERT は生成に有効であることを確認した。しかしながら、これまでの手法ではデータに特別な処理を施さずにモデルの学習を行わせていたため、生成精度の向上は小さかった．データに何らかの前処理を施す必要があるということが課題となっていた。

## 3. 提案手法

本研究では，先行研究の手法を踏まえつつ，その課題に対応するアプローチをとる．手法としては，SciBERT[2]を用いた抽出型要約と生成型要約を用いて，Structured Abstract を自動生成する．これは SciBERT が Structured Abstract の生成に有効であることが，明らかになっているためである．見出しとして，Objectives(目的)，Methods(方法)，Results(結果)，Conclusions(結論)の文を生成する．モデルはそのままに，データに処理を施す．通常，論文のテキストは文が長く，Seq2Seq においては難易度が高い．そこで本研究では，学習データの文章として論文テキストを用いるのではなく，多くの論文に共通して存在する章テキストを活用する．これにより，学習で利用する文の範囲を限定し，難易度を下げる．Structured Abstract の各見出しと多くの論文に存在する章タイトルのテキストを対応付けたデータを活用する．見出しと章タイトルの対応を表 1 に示す．

表 1. 見出しと章タイトルの対応

| 見出し | 章タイトル |
|---|---|
| Objectives | Introduction |
| Methods | Method |
| Results | Result |
| Conclusions | Conclusion |

また、先行研究では行われていない生成的要約による Structured Abstract の生成も行う．生成的要約でもデータを章テキストを活用する手法が有効であるかを確かめる．

## 4. 実験

### 4.1 要約モデル

本研究では Structured Abstract の生成として，先行研究においても利用した抽出的要約モデルの BERTSUMEXT と新たに生成的要約モデルの BERTSUMABS を用いる[3]．これらは言語モデルである BERT を要約タスクに適応させた深層学習モデ

Automatic Generation of Structured Abstracts from Research Papers by Chapter Structures
Kai Hashimoto, Inoue Ushio
†Graduate School of Engineering Tokyo Denki University
‡School of Engineering Tokyo Denki University

ルである. これらのモデルには, 先行研究に基づき, 科学的な言語を含む文に効果的なSciBERTを適用する.

### 4.2 学習データ

学習データには，Semantic Scholar Open Research Corpus[4]に記載されている論文 PDF データ約 8000 件を用いる．データには PDF を XML に変換するライブラリである GROBID を適用する．XML に変換することにより，章テキストを抽出しやすくすることが可能である. XML ファイルより抽出した章テキストと Structured Abstract の見出しをペアにし，学習を行う. データの割合は学習データを 80%， テストデータ 20%とする.

## 5. 評価

評価にはテキスト要約の指標である ROUGE-1, 2, L の F 値を利用する．この値はモデルが作成した見出しと参照元の見出しの一致度を%単位で示す．ROUGE-1，2 はモデルの生成文と参照要約の一致した N-gram の個数から決定される. 式を以下に示す.

$$\mathrm{ROUGE-N} = \frac{\sum_{S\in\{ReferemceSummaries\}} gram_n\in S \sum Count_{match}(gram_n)}{\sum_{S\in\{ReferemceSummaries\}} gram_n\in S \sum Count(gram_n)} \quad (1)$$

N は n-gram の長さを表し，$Count_{match}$はモデルの生成要約と参照元の要約の n-gram の最大一致数を示す.

ROUGE-L では参照された要約とモデルによって生成された要約の最長共通部分列を利用し，その長さを用いる．それより適合率，再現率を求め，それらの重み付き F 値をスコアとしている.

比較のために，章テキストを用いて実験を行った結果だけでなく，本文をそのまま使用した結果も示す．これらの評価結果を表 2,3,4,5 に示す.

表 2. 章テキストを活用した
抽出的要約による評価結果

| 見出し | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| Objectives | 45.8 | 36.1 | 43.2 |
| Methods | 34.9 | 14.5 | 30.1 |
| Results | 37.9 | 17.4 | 35.5 |
| Conclusions | 42.3 | 25.0 | 38.1 |

表 3. 抽出的要約による評価結果

| 見出し | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| Objectives | 37.8 | 17.7 | 33.4 |
| Methods | 34.4 | 13.5 | 30.7 |
| Results | 33.6 | 12.9 | 29.8 |
| Conclusions | 37.3 | 20.2 | 32.5 |

表 4. 章テキストを活用した
生成的要約による評価結果

| 見出し | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| Objectives | 42.3 | 25.7 | 38.6 |
| Methods | 19.9 | 3.4 | 17.0 |
| Results | 22.9 | 4.9 | 20.3 |
| Conclusions | 31.1 | 12.9 | 28.6 |

表 5. 生成的要約による評価結果

| 見出し | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| Objectives | 29.4 | 9.6 | 25.8 |
| Methods | 20.7 | 4.7 | 17.8 |
| Results | 19.7 | 3.8 | 16.5 |
| Conclusions | 22.1 | 4.2 | 19.9 |

## 6. 考察

評価結果より，章テキストの効果について考察を行う．抽出的要約である表 2 と表 3 の結果を比較する．Objectives，Results，Conclusions においては，すべての評価値で数ポイント向上している．Methods においても，僅かに評価値が向上している。生成的要約である表 4,表 5 では，Objectives において，評価値がそれぞれ 10 ポイント以上向上し，Results と Conclusions においても数ポイント向上している．しかし，Methods においては評価値を下げる結果となった.

## 7. まとめと今後の課題

本稿では章構造を用いた Structured Abstract の生成と生成的要約による生成手法を提案した．評価結果としては, 生成的要約の Results 以外では章構造を用いたモデルにおいて有効性が示された．今後の課題としては, 評価値の妥当性がある. Structured Abstract は要約であるため，生成結果を人間による定性的な方法で評価を行う必要がある.

## 参考文献


[1] Signe Holm Nielsen et al.," Serological Assessment of Activated Fibroblasts by alpha-Smooth Muscle Actin (α-SMA): A Noninvasive Biomarker of Activated Fibroblasts in Lung Disorders" Transl Oncol, 368–74, 2019.

[2] Kai Hashimoto and Ushio Inoue, "Automatic Generation of Structured Abstracts from Research Papers by using Deep Learning", IIAI-AAI, 2020.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan."SciBERT: A pretrained language model for scientific text." EMNLP-IJCNLP 2019, pages 3615–3620., 2019.

[4] Yang Liu and Mirella Lapata, "Text summarization with pretrained encoders." EMNLP-IJCNLP 2019, pages 3728–3738., 2019.

[5] Waleed Ammar et al., "Construction of the Literature Graph in Semantic Scholar.", NAACL, 2018.