Vol.2021-ICS-204 No.10
2021/9/16

データセット源としての国際会議の難易度の違いが抽出型論文タイトル生成に与える影響の分析

賀来 健人[1,a)]　菊地 真人[2,b)]　大囿 忠親[2,c)]

概要：BERT に基づく論文タイトル生成おいて，アブストラクトから抽出したキーワードからタイトルを作成する抽出型タイトル生成手法がある．本研究の先行研究では，トップカンファレンスの論文をデータセットとして利用することで，良いタイトルを生成可能であると仮定していた．本稿では，データセットに用いる論文が収録されている国際会議の難易度の違いによる抽出型論文タイトル生成への影響とその分析について述べる．

キーワード：論文タイトル生成，キーワード抽出，論文タイトル評価，BERT，文書要約

1. はじめに

論文タイトルは，その内容を簡潔かつ的確に表すものである．そのため，適切な論文タイトルを作成することが重要となる．しかし，そのような論文タイトルを作成するのは難しいことである．そこで適切な論文タイトルの作成支援のために，論文タイトル候補を提示するシステムが必要とされている．本研究の先行研究では，抽出型文書要約手法を論文アブストラクトに適用することで，論文タイトルを生成する手法を提案した [1]．この手法ではまず，論文アブストラクトから論文タイトル内に含まれる可能性の高いキーワードを抽出する．次に，抽出したキーワードをいくつか連結することでタイトルに含まれる可能性の高い単語列を生成する．これらの単語列の並び替え順列をそれぞれ論文タイトル候補として生成する．これらの論文タイトル候補の文法的正誤を評価し，さらに文法的に正しいと判断された論文タイトル候補の論文タイトルとしての妥当性を評価する．最後に，この評価値が閾値より高い論文タイトル候補をユーザに提示する．キーワードの抽出と論文タイトルの妥当性の評価には，自然言語処理モデルである BERT[2] を組み込んだモデルを用いた．本稿では，論文アブストラクトからキーワードを抽出するためのモデルをキーワード抽出モデル，論文タイトルの妥当性を評価するためのモデルをタイトル評価モデルと呼ぶ．

本研究の先行研究で実装したシステムでは，トップカンファレンスの論文から構成されるデータセットから構築されたキーワード抽出モデルが，トップカンファレンスではない論文のアブストラクトからのキーワード抽出に失敗する傾向が見られた．本研究では，入力される英文の質と，キーワード抽出モデルに用いる英文の質を揃えた方が好ましいと考えた．本稿では，トップカンファレンスとそうでない会議から収集した英文を混合する比率が，キーワード抽出モデルとタイトル評価モデルの性能に与える影響について報告する．

2. 学習データによる影響の推定手法

機械学習において，どのような学習データを用いるかは、モデルの性能を決定づける一因になる．そのため，学習データがモデルに与える影響を解析する研究が広く行われている．Cook は，学習データがモデルに与える影響を調べる最も単純な手法として，データセット全体のデータを使って学習したモデルとデータセットから 1 つのデータを取り除いたデータで学習したモデルの予測値を比較することで影響値を推定する手法を提案した [5]．このような再学習の操作を leave-one-out retraining と呼ぶ．しかし，leave-one-out retraining を行うことはデータセットに含まれるデータの数だけ再学習することを意味する．これを DNN などの巨大なモデルに適用すると非常に長い計算時間を要する．

[1] 名古屋工業大学大学院工学専攻情報工学系プログラム
Computer Science Program, Department of Engineering, Graduate School of Engineering, Nagoya Institute of Technology
[2] 名古屋工業大学大学院情報工学専攻
Department of Computer Science, Graduate School of Engineering, Nagoya Institute of Technology
[a)] kaku@ozlab.org
[b)] masato@ozlab.org
[c)] ozono@ozlab.org

計算時間を短縮するために，影響値を近似的に推定する手法が提案されている．Koh ら [6] と Hara ら [7] は，データセットをいくつかのデータ集合に分けて，データセット全体で学習したモデルとデータセットから 1 つのデータ集合を取り除いたデータで学習したモデルの損失の差からデータの影響値を推定する手法を提案している．これにより，再学習の回数がデータ集合の数にまで減少した．また，小林ら [8] は dropout を利用した turn-over dropout という手法を提案している．この手法では，あるデータを学習する際に dropout を利用して決まったサブネットワークだけで学習することで，反転したサブネットワークの予測値との差からデータの影響値の推定を行うことができる．これにより，再学習を必要とせず，予測回数も 2 回だけとなった．

本研究では，トップカンファレンスとそうでない学会の論文データによるモデルへの影響を解析することが目的である．そのため，個々のデータの影響ではなくデータ集合の影響について解析したい．本研究で用いるデータセットをデータ集合に分割すると，AAAI のデータ集合と JSAI のデータ集合の 2 つになる．この 2 つのデータ集合の影響を解析するため，本研究では，2 つのデータ集合の比率を 4:0，3:1，2:2，1:3，0:4 と変えたデータで学習したモデルの予測や挙動の差を比較する．

3. 抽出型論文タイトル生成システム

本章では，先行研究 [1] において開発した抽出型論文タイトル生成システムについて述べる．まず，システムの構成について述べ，以下の 3 つの段階に分けて論文アブストラクトから論文タイトルを生成する手法について述べる．1 つ目は，論文アブストラクトから論文タイトルに含まれる可能性の高いキーワードを抽出することである．2 つ目は，抽出されたキーワードを並び替えることで複数の論文タイトル候補を生成することである．3 つ目は，生成された論文タイトル候補の論文タイトルとしての妥当性を評価することで，論文タイトル候補の中からより適切な論文タイトル候補を選出することである．

3.1 システム構成

先行システムのシステム構成図を図 1 に示す．先行システムは，キーワード抽出モデルにより構成されるキーワード抽出モジュール，キーワード連結部と並び替え機構により構成されるタイトル候補生成モジュール，文法確認部とタイトル評価モデルにより構成されるタイトル評価モジュールにより構成される．まず，ユーザはタイトルを生成したい論文のアブストラクトをシステムに入力する．入力されたアブストラクトはサーバに送信され，サーバでキーワード抽出モデルに入力される．ここで論文タイトルに含まれる可能性の高いキーワードがアブストラクトから抽出される．アブストラクトから抽出されたキーワードは，キーワード連結部に入力される．ここで，連結されたキーワードはクライアントに送信される．ユーザは，これらのキーワードを確認し，必要に応じてキーワードを修正することができる．修正されたキーワードは再度サーバに送信され，並び替え機構に入力される．並び替え機構は，送信されたキーワードを並び替えることで，キーワードの並び替え順列である論文タイトル候補を生成する．次に，並び替え機構で生成された論文タイトル候補を文法確認部に入力する．文法確認部は，文法的に正しくないタイトル候補を除外し，文法的に正しい論文タイトル候補のみを出力する．最後に，文法確認部に出力された論文タイトル候補は，タイトル評価モデルに入力される．タイトル評価モデルは，入力された論文タイトル候補の妥当性を評価し，評価値が閾値よりも高い論文タイトル候補だけを出力する．タイトル評価モデルにより出力されたタイトル候補のみが，最終的な論文タイトル候補としてクライアントに送信され，ユーザーに提示される．以降では，キーワード抽出モジュール，タイトル候補生成モジュール，タイトル評価モジュールについて述べる．

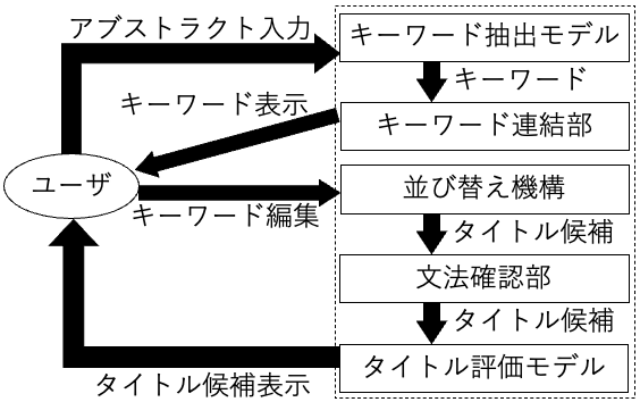

図 1 システム構成図

3.2 BERT によるキーワード抽出

キーワード抽出モジュールは，キーワード抽出モデルにより構成されている．キーワード抽出モジュールは，論文タイトルに含まれる可能性の高いキーワードを論文アブストラクトから抽出する．

キーワード抽出モデルは，論文アブストラクト内の各単語が論文タイトルに現れるか否かに基づく 2 値分類によって，論文タイトルに含まれる可能性の高い単語を抽出する．キーワード抽出モデルは，BERT を用いて開発した．キーワード抽出モデルの構成を図 2 に示す．まず、論文アブストラクトを BERT に入力可能にするために，BERT tokenizer を使用して論文アブストラクトの文字列を単語 ID 列に変換する．BERT tokenizer は，単語をサブワードに分割し，これらを単語 ID に変換する．次に，入力系列の

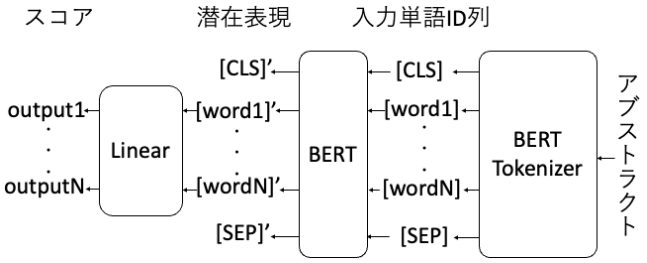

図 2 キーワード抽出モデル

先頭に [CLS] トークンを挿入し，各文の末尾に [SEP] トークンを挿入する．これらはそれぞれ，入力系列の意味と文の区切りを表す特別なトークンである．ここで，入力系列を BERT に入力し，その出力を得る．特殊トークンを除く論文アブストラクト内の各サブワードの潜在表現ベクトルに線形変換を施すことにより，各サブワードのスコアのベクトルが生成される．最後に，これらのスコアを閾値と比較することで，論文のタイトルに現れるサブワードとそうでないサブワードに 2 値分類する．学習の際には，タイトルに含まれるアブストラクト内のトークンには 1(正解のラベル) を，タイトルに含まれないものには 0(不正解のラベル) を割り当てた．

3.3 タイトル候補生成

タイトル候補生成モジュールはキーワード連結部と並び替え機構により構成される．タイトル候補生成モジュールでは，キーワード抽出モジュールにより抽出されたキーワードから論文タイトル候補を生成する．

キーワード連結部では論文アブストラクトの情報をもとに，キーワードを連結しサブワードから単語に復元したり，複数の単語を連結しイディオムを 1 つのキーワードとして扱うことができるようにする．本研究では，アブストラクト中で連続している単語を連結することとする．キーワード連結部の疑似コードをアルゴリズム 1 に示す．まず，アブストラクトからキーワードの最長一致の単語列を抽出する．この時，アブストラクト中の単語は BERT Tokenizer によってサブワードに分割された状態であり，1 つの単語が複数のサブワードに分解されている場合がある．このようなサブワードは単語の先頭にあるものを除いてはトークンの先頭に“##”というマーカーがついている．このマーカーがついている場合はマーカーを削除して前のサブワードと結合する．これにより単語を復元することができる．また，マーカーがついていない場合は前のサブワードとの間に空白を開けて結合する．これにより，イディオムなどの複数の単語で構成されるある程度まとまった単語列を 1 つのキーワードとして扱うことができる．ここまでの処理で生成されたキーワードのうち不要なキーワードは除外する．不要なキーワードとは，他のキーワードと一致するキーワードと “-” のみで構成されているキーワードである．最後に，ユーザはキーワードを確認し，必要があればキーワードを直接編集することができる．

アルゴリズム 1 キーワード連結部の擬似コード

```
1: function GenerateTitleParts(keywords, abstract)
2:     keywords = GetLM(keywords, abstract)
3:                 ▷ Extract the longest matching word sequence.
4:     keywords = Repair(keywords)
5:                                   ▷ Join the split words again.
6:     keywords = Dump(keywords)
7:                               ▷ Exclude unnecessary keywords.
8:     keywords = UserEdit(keywords)
9:                                       ▷ Perform user editing.
10:    return keywords
11: end function
```

アルゴリズム 2 並び替え機構の疑似コード

```
1: function Shape(keywords, title, title_list)
2:     for i = 1 to Length(parts) do
3:         Add(title, keywords[i])
4:         Delete(keywords, keywords[i])
5:         if Length(keywords) == 0 then
6:             if GrammarCheck(title) then
7:             ▷ Determine whether it is grammatically correct.
8:                 score = Evaluate(title)
9:                  ▷ Determine whether the title is appropriate.
10:                if score ≥ 0.5 then
11:                    Add(title_list, title)
12:                end if
13:            end if
14:        else
15:            title_list = Shape(keywords, title, title_list)
16:                    ▷ Call this recursively if keywords remain
17:        end if
18:    end for
19:    return title_list
20: end function
```

並び替え機構はキーワードを並び替えることで論文タイトル候補を生成する．並び替え機構の疑似コードを 2 に示す．初期状態では，引数はキーワードの集合と後に論文タイトル候補となる空の文字列，論文タイトル候補の一覧を格納する空の配列である．まず，キーワードの集合から 1 つのキーワードを取り出し，これを論文タイトル候補となる文字列へ追加する．これをキーワードの集合が空になるまで Shape 関数を再帰させ，キーワードの集合が空になった時に論文タイトル候補が完成したと考え，文法確認とタイトル評価を行うことにより論文タイトルとしてふさわしいか判断する．論文タイトルとしてふさわしいと判断された論文タイトル候補はユーザに提示するタイトル候補を格納する配列に追加する．最後にこれをユーザに提示する．

3.4 BERT によるタイトル評価

タイトル評価モジュールは文法確認部とタイトル評価モデルから構成される．タイトル評価モジュールでは，論文タイトル候補を評価することによって，論文タイトル候補

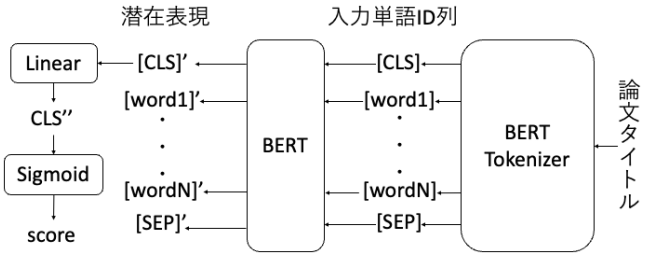

図 3 タイトル評価モデル

から評価値の高いものを選び出す．

文法確認部は論文タイトルの構文木の構造を元に論文タイトル候補の文法の正誤を検証する．本研究では，文法パターンというものを用いる．文法パターンとは，論文タイトルの構文木から単語や POS タグなどの構文木の葉にあたる部分を除き，句や節を表すタグのみで構成された構文木の一部分のことである．事前に実際の論文タイトルから文法パターンを収集しておき，実際の論文タイトルの文法パターンをタイトル候補の文法パターンと照合することによってタイトル候補の文法の正誤を判定する．収集した文法パターンにタイトル候補の文法パターンと一致するものが 1 つでもある場合，そのタイトル候補は文法的に正しいとしてタイトル評価モデルに入力される．また，収集した論文タイトルの中にタイトル候補の文法パターンと一致するものがない場合，そのタイトル候補は文法的に誤っているとして破棄される．実際の論文タイトルやタイトル候補から構文木を生成する際には構文解析器を用いた．本研究では，Berkeley Parser を用いた．

タイトル評価モデルは，論文タイトル候補の論文タイトルとしての妥当性を評価し，その評価値を閾値と比較することによりに論文タイトルとして適切であるか否かに基づく 2 値分類を行う．これにより，ユーザに提示するための適切な論文タイトル候補のみを得ることができる．タイトル評価モデルは BERT を用いて作成した．タイトル評価モデルの構成を図 3 に示す．まず，キーワード抽出モデルと同様に BERT Tokenizer を用いて論文タイトル候補を単語 ID 列に変換する．次に，論文タイトル候補の単語 ID 列を BERT に入力し，その出力を得る．この出力は各トークンの潜在表現ベクトルである．これらの潜在表現ベクトルのうち文の全体の意味を捉える [CLS] トークンが論文タイトルの妥当性に関する情報を持っていると考えた．そのため，[CLS] トークンの潜在表現ベクトルを線形変換することにより論文タイトル候補の妥当性のスコアを得る．このスコアを sigmoid 関数に入力することで 0〜1 の範囲の値を持つスコアに変換する．これらの値を閾値と比較することによって論文タイトルとして適切な論文タイトル候補とそうでない論文タイトル候補に分類する．学習の際には，実際の論文タイトルには 1(正解のラベル)，そうでないものには 0(不正解のラベル) を割り当てた．

4. 評価実験・考察

4.1 モデルの学習

本研究では，トップカンファレンスとそうでない学会の論文をデータセットとした時の抽出型論文タイトル生成システムへの影響を分析することが目的であるため，トップカンファレンスの論文データとそうでない学会の論文データを用いた．評価実験において，DBLP API を用いて収集した AAAI の論文データ 3,500 件と CiNii Articles OpenSearch を用いて収集した JSAI の論文データ 3,500 件を用いた．ただし，「:」が含まれるタイトルは文法が複雑になり BERT が学習できないと考え，これらのデータは除外した．キーワード抽出モデルとタイトル評価モデルの学習には AAAI と JSAI の論文データをそれぞれ 3,000 件ずつ用い，AAAI の論文データと JSAI の論文データの比率を 0:4，1:3，2:2，3:1，4:0 としたデータセットを作成した．これらのデータセットに含まれるデータの総数はそれぞれ 3,000 件となるようにした．これらのデータセットを学習データとし，これらのデータでキーワード抽出モデルとタイトル評価モデルを 5epoch ずつ学習した．また，残りの 500 件ずつの論文データを評価用データに使用した．ここでは，訓練用データと評価用データにそれぞれ含まれるトップカンファレンスの割合が性能に与える影響を調べるため，3 種類の評価用データを用いた．具体的には，AAAI と JSAI の両方の英文を含む，AAAI の英文のみ，および JSAI の英文のみの 3 種類である．

4.2 キーワード抽出モデルの性能評価

評価用のデータセットを用いてキーワード抽出モデルによって抽出された単語と実際の論文タイトルに含まれる単語から再現率，適合率，F 値を検証した．再現率を r，適合率を p，F 値を f，キーワード抽出モデルが抽出した単語の集合を A，実際の論文タイトルに含まれる単語の集合を B とした時，それぞれ次式のように表すことができる．

$$r = \frac{number(A \cap B)}{number(B)}, p = \frac{number(A \cap B)}{number(A)}, F = \frac{2pr}{p+r}$$

評価用データを用いた時のキーワード抽出モデルの予測結果の F 値を表 1 に示す．評価データに AAAI を用いた時と JSAI を用いた時では，学習データと評価データの会議が一致している場合の方が F 値が高く，学習データと評価データの会議が一致していない場合の方が F 値が低くなっている．また，AAAI 論文に基づくキーワード抽出モデルは，IJCAI 論文のアブストラクトからのキーワード抽出性能が高い [1]．このことから，キーワード抽出モデルは学習するアブストラクトの質と入力するアブストラクトの質が近い時により正確にキーワードを抽出することができると考えられる．また，多様な会議から英文を収集することで，会議に依存しない汎用的なキーワード抽出モデルが

表 1 キーワード抽出モデルの予測結果の F 値

評価データ	学習データの比率 (AAAI:JSAI)				
	0:4	1:3	2:2	3:1	4:0
AAAI	0.424	0.420	0.431	0.445	0.455
AAAI+JSAI	0.419	0.396	0.405	0.429	0.406
JSAI	0.413	0.372	0.380	0.414	0.358

表 2 タイトル評価モデルの予測結果の F 値

評価データ	学習データの比率 (AAAI:JSAI)				
	0:4	1:3	2:2	3:1	4:0
AAAI	0.804	0.808	0.884	0.792	0.905
AAAI+JSAI	0.810	0.807	0.858	0.765	0.712
JSAI	0.816	0.805	0.830	0.737	0.385

構築可能であるという可能性もある．

4.3 タイトル評価モデルの性能評価

評価用のデータセットを用いてタイトル評価モデルにより論文タイトルの妥当性を評価した際の正解率，再現率，適合率，F 値を検証した．正解率を a，再現率を r，適合率を p，F 値を f とした場合，それぞれ次式により表すことができる．

$$a = \frac{TP}{TP + FP + FN + TN}, r = \frac{TP}{TP + FN},$$

$$p = \frac{TP}{TP + FP}, f = \frac{2pr}{p + r}$$

評価用データを用いた時のタイトル評価モデルの予測結果の F 値を表 2 に示す．AAAI の論文データのみで学習した場合は，AAAI の評価データを評価した際の F 値が高くなり，JSAI の評価データを評価した際の F 値は低くなっている．このことから，AAAI の論文データで学習したタイトル評価モデルはトップカンファレンスの論文タイトルのような質の高い論文タイトルを高く評価し，それ以外の論文タイトルを低く評価していると考えられる．また，JSAI の論文データでのみ学習した場合は，評価データによらず F 値はほぼ一定である．このことから，JSAI の論文タイトルは当たり障りのない論文タイトルであれば高く評価しているように考えられる．タイトル評価モデルはトップカンファレンスの論文タイトルのような質の高い論文タイトルのみを高く評価することが必要であるので，タイトル評価モデルの学習にはトップカンファレンスの論文タイトルを使用した方がよいと考えられる．

5. おわりに

本稿では，抽出型タイトル生成システムのキーワード抽出モデルとタイトル評価モデルの学習データにトップカンファレンスの論文データとそうでない論文データを使用した時の影響について分析した．評価実験の結果から，キーワード抽出モデルは入力するアブストラクトと学習するアブストラクトの質を揃えた方がよいことがわかる．また，キーワード抽出モデルの学習データには，多様な会議の論文データを用いることで会議に依存しない汎用的なキーワード抽出モデルを構築できる可能性があることが示唆された．また，タイトル評価モデルは，より質の高い論文タイトルのみを高く評価することが必要であるので，トップカンファレンスの論文データを学習することが望ましいといえる．

謝辞 本研究の一部は JSPS 科研費 JP19K12266 の助成を受けたものです．

参考文献

[1] K. Kaku, M. Kikuchi, T. Ozono, T. Shintani. “Development of an Extractive Title Generation System Using Titles of Papers of Top Conferences for Intermediate English Students”. 10th International Congress on Advanced Applied Informatics, pp.59-64, 2021.

[2] J. Devlin, M. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol.1, pp.4171-4186, 2019.

[3] 大部達也, 大囿忠親, 新谷虎松, “論文執筆支援エージェントのためのタイトル生成機構,” 合同エージェントワークショップ＆シンポジウム 2017, pp.232-237, 2017.

[4] P. Mishra, C. Diwan, S. Srinivasa, G. Srinivasaraghavan., “Automatic Title Generation for Text with Pre-trained Transformer Language Model,” 15th IEEE International Conference on Semantic Computing, pp.17-24, 2021.

[5] R. D. Cook, “Detection of Influential Observation in Linear Regression,” Technometrics, vol.19, no.1, pp.15–18, 1977.

[6] P. W. Koh, P. Liang, “Understanding Black-box Predictions via Influence Functions,” Proceedings of the 34th International Conference on Machine Learning, vol.70, pp.1885-1894, 2017.

[7] S. Hara, A. Nitanda, T. Maehara, “Data Cleansing for Models Trained with SGD,” Annual Conference on Neural Information Processing Systems 2019, pp.4215-4224, 2019.

[8] 小林颯介, 横井祥, 鈴木潤, 乾健太郎, “訓練事例の影響の軽量な推定,” 自然言語処理, vol.28, no.2, pp.573-597, 2021.