セグメント構造を考慮した学術論文の包括的要約の自動生成の提案

SHIN Wonha　　白井 清昭
北陸先端科学技術大学院大学 先端科学技術研究科
{s1510026,kshirai}@jaist.ac.jp

1 はじめに

研究のサーベイは，多くの学術論文を読む必要があり，労力の大きい作業である．このため，調査対象とする全ての論文の全文を読むのではなく，まず論文の概要を読み，重要な論文を選別した後，それらの論文の全文を読むという方法が効率的である．このとき，最初に読む概要としては，論文の冒頭に掲載されているアブストラクトや，自動作成された要約が考えられる．しかしながら，これらはサーベイに適していない場合がある．

論文に記載されているアブストラクトは一般に簡潔である．一方，サーベイのために論文を読む場合は，論文の内容をある程度深く理解するために，extended abstract のような長めの要約が必要とされることも多い．これに対し，自動要約によって要約を生成すれば，ユーザが望む長さの要約を得ることができる．ところが，サーベイのために読む要約としては，その論文の目的や提案手法の概略だけでなく，先行研究に対する位置付け，実験の設定やその結果，論文の貢献など，論文の主な要点が全て含まれていることが望ましい．従来の自動要約としては重要文抽出型の手法が主流であるが，上記のような論文の要点を全て含むかという観点で重要文が選ばれているわけではない．

本論文では，学術論文のセグメント構造に着目し，学術論文の背景，目的，関連研究との位置付け，提案手法，評価実験など，論文の要点を全て含む要約を「包括的要約」と定義し，これを自動生成する手法を提案する [5]．セグメント構造とは，ここでは章や節によって定義される学術論文の部分テキストの集合と定義する．提案手法は重要文抽出型の単一文書要約手法と位置付けられる．従来の単一文書要約手法 [1, 2, 3, 4] と異なり，学術論文が持つ典型的なセグメント構造を考慮し，各セグメントから重要文を抽出することで包括的要約を作成する点に特徴がある．

2 提案手法

提案手法の処理の流れを図 1 に示す．本研究では，論文として LaTeX のソースファイルが与えられるものと仮定する．まず，論文が与えられたとき，その論文のセグメント構造を解析する．本研究では，学術論文の典型的な構造を考慮し，「序論」「関連研究」「提案手法」「評価実験」「結論」の 5 つのセグメントに分割する．次に，各セグメントから重要文を抽出する．ただし，「結論」のセグメントからは重要文を抽出しない．「結論」では論文のまとめや今後の課題が書かれていることが多いが，論文のまとめは多くの場合「序論」の内容と重複し，また今後の課題はサーベイのための要約に含める必要性は低いと本論文では考える．また，重要文を抽出する際には，それぞれのセグメントに適した重要文抽出手法を開発し，それを用いる．最後に，各セグメントから抽出された重要文を統合し，要約を得る．このように，論文における複数のセグメントから重要文を抽出することで，論文の主要な内容を全て含む包括的要約を作成する．

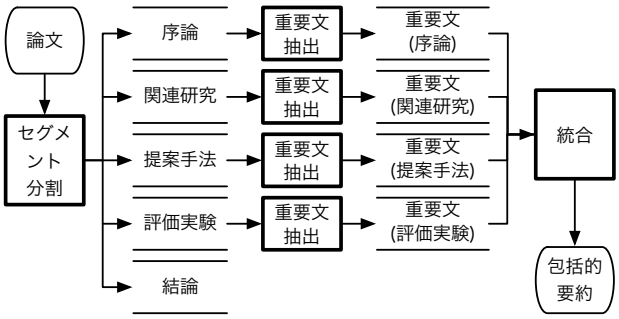

図 1: 提案手法の概要

以下，提案手法の詳細について述べる．ただし，「提案手法」と「評価実験」のセグメントからの重要文抽出，ならびに「統合」のモジュールは実装が完了していない．そのため現時点での構想を述べる．

2.1 セグメント分割

2.1.1 節のタイトルを手がかりとする手法

「提案手法」を除くセグメントについて，そのセグメントの節のタイトルによく使われると思われるキーワードのリスト (表 1) をあらかじめ用意する．\section と

P1: われわれ | 我々 | 本 (研究|手法|論文|稿) | 特徴 | 具体
P2: これ (まで|ら) の (研究|手法|方法) | 提案 | 比較 | 研究 | 方法 | 手法 | CITE
P3: しかし | 一方 | ただ | 違い | 異なる | 異なり | (で|て)(ε|は) ない | いない | できない | でき (る|た)

図 2: 「関連研究」のセグメントを検出するための手がかり句

表 1: セグメントのキーワードの一覧

セグメント	キーワード
序論	はじめに, まえがき, 序論, はしがき, 背景, 緒論
関連研究	関連研究
評価実験	実験, 評価, 評価実験, 評定実験
結論	考察, 結論, おわりに, 終わりに, 結び, むすび, まとめ, あとがき

いう LaTeX コマンドでマークアップされているテキストがキーワードを含むとき，その節をセグメントとして抽出する.「提案手法」のセグメントについては，様々な単語がタイトルに出現し，これらをあらかじめ網羅的に収集することが難しいことから，キーワードのマッチングでは抽出しない．代わりに，他の 4 つのセグメントのいずれにも該当しない節を「提案手法」のセグメントとして取り出す．なお，この手法で取り出されるセグメントは節を単位とする．

2.1.2 関連研究の手がかり句に基づく手法

予備実験では,「関連研究」のセグメントが検出できない誤りが多かった．そのため，節のタイトルに対するパタンマッチで「関連研究」のセグメントが抽出されなかったとき，手がかり句を用いて抽出する．ここでの手がかり句とは,「関連研究」のセグメントで典型的に使われると考えられる表現を表わすパターンであり，図 2 のように定義する．P1 は論文の特徴を述べる文に，P2 は先行研究との比較を述べている文に，P3 は先行研究の問題点を指摘する文にマッチすることを想定している．また，P2 における CITE は，論文を引用する LaTeX のコマンド \cite にマッチすることを表わす．図 2 のパターンにマッチする文があれば，その文を含む段落，及びその前に出現する 2 つの段落を「関連研究」のセグメントとして抽出する．したがって，この手法で抽出されるセグメントは段落を単位とする．

2.2 重要文の抽出

2.2.1 序論のセグメントからの重要文抽出

「序論」のセグメントからの重要文抽出は教師あり機械学習の手法を用いる.「序論」における重要文は，論文の冒頭に記載されているアブストラクトと重複することが多いと考えられる．そこで，論文のアブスラクトを正解とみなし，文を要約に含めるべき重要文か否かを判定する二値分類器を機械学習する．

訓練データは以下の手続きで作成する．論文から「序論」のセグメントを抽出し，それに含まれる文を s_i とおく．s_i が式 (1) の条件を満たすとき，その文を重要文と，そうでないときは非重要文とタグ付けする．

$$\max_{s_a \in A} \ sim(s_i, s_a) \ > \ T \tag{1}$$

$$sim(s_i, s_a) = \sum_{x \in TG(s_i), y \in TG(s_a)} \delta_{x,y} \tag{2}$$

A は論文のアブストラクトにおける文の集合，s_a はその要素となる文である．$sim(s_i, s_a)$ は文間の類似度であり，これが閾値 T より大きい文が A に存在するとき，文 s_i と同じ内容の文がアブストラクトに出現するとみなし，重要文であると判定する．本論文では $T = 6$ と設定した．文間類似度は式 (2) のように定義する．TG は文中に含まれる単語 3-gram の集合であり，$\delta_{x,y}$ はクロネッカーのデルタ[1]である．つまり，2 つの文に共通して現われる単語 3-gram の数を類似度と定義する．

次に，重要文か否かを判定するモデルの学習について述べる．学習アルゴリズムは Support Vector Machine(SVM) とする．SVM の学習には LIBSVM[2]を用いる．カーネルは線形カーネルを使用し，学習パラメタはデフォルト値とする．SVM を学習するためには，訓練データの文を素性ベクトルに変換する必要がある．素性は文中に含まれる単語の n-gram(n=1,2,3) とする．ただし，$n = 1$ のときは自立語のみを素性とする．素性の重みは，単語 n-gram が文に出現すれば 1，それ以外は 0 とする．また，簡単な素性選択を行う．具体的には，訓練データにおける出現頻度が 1 の素性を削除する．

2.2.2 関連研究のセグメントからの重要文抽出

「関連研究」のセグメントから重要文を抽出する手法について述べる．セグメント内の各文 s_i に対して式 (3) のスコアを求め，そのスコアの上位 N 件の文を重要文として抽出する．

$$Score(s_i) \ = \ Score_{tfidf}(s_i) + Score_{par}(P) \tag{3}$$

$$Score_{tfidf}(s_i) \ = \ \sigma(\ \textstyle\sum_{w \in s_i} \mathrm{TF} \cdot \mathrm{IDF}(w)\ ) \tag{4}$$

$$Score_{par}(P) \ = \ \sigma(\ \textstyle\sum_{s_j \in P} \ pat(s_j)\ ) + 1 \tag{5}$$

[1] $x = y$ のとき 1, $x \neq y$ のとき 0.
[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

$Score_{tfidf}(s_i)$ は文中に含まれる単語の TF·IDF の値によって決まるスコアで，式 (4) で定義される．w は文 s_i に含まれる単語，TF·IDF(w) は w の TF・IDF の値である．一方，$Score_{par}(P)$ は，s_i を含む段落 P に応じて与えられるスコアであり，式 (5) のように定義する．$pat(s_j)$ は，図 2 のパターンに応じて与えられるスコアであり，段落中の文 s_j が P1, P2, P3 にマッチしたとき，それぞれ 10, 3, 2 点とする．σ は bipolar sigmoid function[3] であり，$Score_{tfidf}(s_i)$ を $[0,1]$，$Score_{par}(P)$ を $[1,2]$ の範囲の値に変換する働きをする．式 (3) は，関連研究の手がかり句にマッチする段落に含まれる文を優先的に選択し，それ以外の文は文中の単語の TF·IDF 値の和によって重要度を決めることを意味する．

2.2.3 その他のセグメントからの重要文抽出

「提案手法」のセグメントからは，手法の概略を説明する文を重要文として抽出することを考えている．多くの場合，手法の概要は節や項 (subsection) の先頭に書かれることが多いため，重要文抽出の際には文の位置が有効な手がかりになると思われる．また，提案手法の処理の流れはしばしば図で提示される．そのような図を抽出し，要約に含めると，サーベイの際に提案手法の概略を理解しやすくなる．

「評価実験」のセグメントからは，実験の設定や結果を説明する文を重要文として抽出することを考えている．実験の設定については，評価実験の節の冒頭に書かれることが多い．また，実験結果はしばしば表やグラフで表わされる．表やグラフは実験結果を把握しやすいため，これを検出して要約に含めることは有望である．

2.3 重要文の統合

各セグメントから抽出した重要文を元の論文での出現順に結合することで，包括的要約を得る．この際，セグメント毎に重要文をまとめて提示することで，ユーザが異なる観点から論文の要点を把握できるようにする．また，要約率のコントロールは重要な課題である．要約率を満たすように重要文を選択する際，各セグメント毎に要約率を満たすように選択するのか，あるいは「序論」のような重要と考えられるセグメントからより多くの重要文を選択するのかは，今後検討する必要がある．

3 評価実験

3.1 実験データ

本論文では，実験データとして，言語処理学会論文誌 LaTeX コーパス[4]を用いた．このコーパスは，会誌「自然言語処理」に掲載された論文の LaTeX のソースファイルを集めたデータ集である．実験では，ランダムに選択した 30 件の論文をテストデータとし，388 件の論文を訓練・開発データとした．訓練・開発データは，2.2.1 で述べた「序論」のセグメントの重要文を選択する SVM の学習に用いるほか，提案手法を設計する際に参照した．例えば，表 1 のキーワードの選定や図 2 のパターンの作成は，訓練・開発データの論文を参照して行った．

3.2 セグメント分割の評価

まず，節のタイトルを手がかりとする手法でセグメントを決定する手法を評価する．評価基準はセグメント抽出の精度，再現率とする．また，抽出率を 1 つ以上の節がセグメントとして抽出できた論文の割合と定義し，これも評価基準とした．結果を表 2 に示す．精度は高いが，「関連研究」と「評価実験」の再現率が低い．「関連研究」については，独立した節ではなく節内の一部の段落で関連研究について論じている論文が多かった．「評価実験」は，提案手法の節がいくつかの項から構成され，その最後の項に評価実験の内容が書かれている論文があり，このような場合に「評価実験」のセグメントの抽出に失敗していた．

セグメントの抽出率は「関連研究」を除いて高い．「関連研究」については，30 件中 10 件の論文しかセグメントを抽出できなかった．残りの 20 件の論文については，図 2 に示した手がかり句のパタンマッチによって，関連研究について述べている段落を取り出すことを試みた．抽出されたセグメント (段落) が関連研究に関する記述を含む場合には正解とみなし，その正解率を人手で算出した．その結果，正解率は 65%であった．

3.3 序論からの重要文抽出の評価

2.2.1 で述べた重要文抽出手法を評価する．正解の要約，すなわち正解の重要文のセットは，訓練データと同じ方法で作成した．すなわち，テストデータの「序論」のセグメントに含まれる文が，その論文のアブストラクト内の文とほぼ同じ意味を持つとみなせるとき，それを要約に含めるべき重要文とした．重要文抽出の精度，

表 2: セグメント分割の結果

セグメント	序論	関連研究	提案手法	評価実験	結論
精度	1.0	1.0	0.83	1.0	1.0
再現率	1.0	0.62	1.1	0.73	0.91
抽出率	1.0	0.33	1.0	0.77	0.97

[3] $\sigma(x) = 2/(1+e^{-x}) - 1$

[4] http://www.anlp.jp/resource/journal_latex/index.html

表 3: 「序論」のセグメントからの重要文抽出結果

精度	再現率	F 値
0.31	0.29	0.30

表 4: 「関連研究」のセグメントからの重要文抽出結果

	精度	再現率	F 値
全体 (30 論文)	0.21	0.24	0.22
タイトル (10 論文)	0.20	0.32	0.25
手がかり句 (20 論文)	0.21	0.22	0.22

再現率，F 値を表 3 に示す．いずれも 30%程度であり，改善の余地がある．

3.4 関連研究からの重要文抽出の評価

2.2.2 で述べた重要文抽出手法を評価する．テストデータの論文から重要文と思われるものを人手で選択し，これを正解の要約とした．この際，正解とする重要文の数に制限を設けず，その論文の先行研究と提案手法の違いを述べているとみなせる文を全て選択した．また，「関連研究」のセグメント以外でも，例えば「序論」や「結論」のセグメントに関連研究について言及している文があれば，それも正解の重要文とした．提案手法によってセグメントから 4 個の文を選択したときの重要文抽出の精度，再現率，F 値を表 4 に示す．2 列目の「全体」はテストデータ 30 論文に対する評価結果，3 列目の「タイトル」は節のタイトルに対するパタンマッチによってセグメントを抽出した 10 論文に対する評価結果，4 列目の「手がかり句」は手がかり句のパタンマッチによってセグメントを抽出した 20 論文に対する評価結果を示す．精度は，タイトルのパタンマッチで検出されたセグメント，手がかり句のパタンマッチで検出されたセグメントのいずれも 20%程度であった．一方，再現率は，前者のセグメントが 32%，後者のセグメントが 22%で，前者の方が 10%程度高かった．タイトルのパタンマッチの方が手がかり句によるパタンマッチと比べてセグメント抽出の精度が高く，また手がかり句によるパタンマッチで抽出されるセグメントは節ではなく段落であり，関連研究とは関係のない文が含まれていることが多いことが原因として考えられる．

4 おわりに

本論文では，サーベイの労力を軽減させることを目標とし，学術論文の主要な要点を含む包括的要約を生成することを提唱した．包括的要約を自動生成するために，論文のセグメント構造を解析し，「序論」「関連研究」「提案手法」「評価実験」のそれぞれから重要文を抽出し，最終的にこれらを統合する手法を提案した．さらに，セグメント構造を解析する手法，ならびに「序論」と「関連研究」のセグメントから重要文を抽出する手法を実装し，その有効性を評価した．

今後の課題を以下に述べる．まず,「提案手法」「評価実験」のセグメントからの重要文抽出手法を実装する．それぞれのセグメントから取り出すべき重要文は何か，その重要文を抽出するための言語的，構造的特徴は何かを分析し，それぞれのセグメントに適した重要文抽出手法を探究する.「序論」「関連研究」のセグメントについても，現在の手法の重要文抽出の精度や再現率は高くないため，更なる改善が必要である．

複数のセグメントから抽出された重要文を統合する手法も開発する必要がある．特に，2.3 節で述べたように，ユーザによって与えられた要約率を満たすように重要文を選択する手法は十分に検討する必要があると考えている．

作成された包括的要約の評価も重要な課題である．正解の包括的要約を人手で生成し，ROUGE のような指標で評価することが考えられる．また，作成された包括的要約がサーベイにどの程度役に立つのか，すなわち包括的要約の生成が複数の論文の内容を短時間で把握するのにどれだけ貢献するかを確認するための被験者実験も必要である．

参考文献

[1] John M. Conroy and Dianne P. O'leary. Text summarization via hidden markov models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 406–407, 2001.

[2] H. P. Edmundson. New methods in automatic extracting. *Journal of the ACM*, Vol. 16, No. 2, pp. 264–285, 1969.

[3] Julian Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 68–73, 1995.

[4] 奥村学, 難波英嗣. テキスト自動要約に関する研究動向. 自然言語処理, Vol. 6, No. 6, pp. 1–26, 1999.

[5] Wonha Shin. セグメント構造に基づく学術論文の自動要約. 修士論文, 北陸先端科学技術大学院大学, 2017.