E-032

テキスト構造に着目した学術論文の要旨自動生成のための重要文抽出

Key-sentence Extraction Using Text Structure for Automatic Summarization of the Scientific Paper

徳永 康次＊　延澤 志保‡　太原 育夫†
Koji Tokunaga　Shiho Nobesawa　Ikuo Tahara

1. はじめに

論文には内容を簡潔にまとめた要旨が付属しており，論文の要旨は，多くの論文の中から自分の研究に関連する論文を見つけ出す際に有用である．自動要約は自然言語処理における重要なテーマの一つであり，論文の要旨の自動生成についても様々な手法が提案されてきた[1]．本研究では論文というテキストの構造に着目し，要旨を動機，手法，結果の3つによって成り立つ文章と定義し，表層的な情報を利用した重要文抽出を行い，要旨を自動生成する手法を提案する．その際，手がかりとなる単語には各単語の頻度ではなく，動機，手法，結果を示す意味を含む単語の頻度を利用することを試みる．

2. 要旨の構成文とその抽出

テキストには様々なジャンルがあり要約処理にはそのジャンル固有の構造に着目することが有効な手段である[2]．学術論文から要旨を生成する場合，要旨がどのような文から構成されるかに着目する必要がある．

理想的な要旨の書き方としていくつかの指摘がなされている[3]．論文に付属する要旨は，主要な事実と結論とを総括し，それ自体で本文を見ないでも了解可能であり，完結したものであるべきである．すなわち，

- 論文で扱った主題と論文の目的を示す．
- 新しく観察された事実，実験的あるいは論理的な発見，結論その他特徴的なことを総括する．
- できるだけ具体的で量的なデータを示す．
- 論文中の新しい実験データを得た方法を示す．

などに注意して書くことが望ましい．そこで以上のことから，要旨は「何故その研究を行い，何を目指すのかがわかる文」である動機文，「その論文の筆者が行った手法がわかる文」である手法文，「具体的な数値などの結果がわかる文」である結果文から構成されるものと定義する．

動機文，手法文，結果文の各文を論文中から抽出する場合，抽出する箇所の選択に論文の構造が有効である．学術論文の構成上，『序文』（はじめに等）に動機文，手法文が載っており，論文の最後に書かれる『結論』（おわりに，むすび等）には簡単な結果が載っている．このことから『序文』から動機文，手法文を『結論』から結果文を抽出する．これにより論文全体を処理する必要がなくなり，抽出の精度を向上することができる．

＊東京理科大学院理工学研究科情報科学専攻
‡武蔵工業大学知識工学部情報科学科
†東京理科大学理工学部情報科学科

3 動機文，手法文，結果文の特徴

3.1 重要文の特徴

重要文を決定する特徴素として文の位置情報，タイトル上の単語，手がかり語の3つの特徴素を利用することができる．このとき文sの重要度W(s)は，C(s)，L(s)，T(s)をそれぞれ文s に対する手がかり語，位置情報，タイトル上に含まれる単語に基づく重要度とすると次式のように与えることができる．

$$W(s) = C(s) + L(s) + T(s) \qquad (1)$$

この式を用いて重要文を抽出すると，テストデータに対する自動要約と人手で抽出した文を比較して評価した結果，最も人手で抽出した文に近い文を抽出できることがわかっている[4]．

3.2 動機文，手法文，結果文の特徴

動機，手法，結果の3つの抽出対象の文の特徴を把握するため，実際の論文集から様々なジャンルの理系論文(FIT2006年第5回情報科学技術フォーラムの中から無作為に選んだ論文)172件についてそれぞれの特徴を調査した．具体的には，位置情報は抽出対象の文が第何段落の第何文にあるか，動機，手法，結果を示す手がかり語はどのようなものがあるか，タイトル上の単語はどのような文に使われているかを調査した．このとき，動機文，手法文，結果文の判断や手がかり語の選出には文の意味を判断する必要があるので人手で行った．

(1)動機文について

動機文は論文の問題提起であることから第一段落に書かれているものが多く，またその段落中でも第一文，第二文に動機が書かれている．有効な手がかり語として社会背景を述べる文に"近代"，"現代"，などの手がかり語が，さらにそこに存在する問題を述べる文には"課題"，"問題"等が含まれており，それらを手がかり語として利用する．

(2)手法文について

手法文は筆者の考案した手法を告げる文であり，第何段落に含まれるかはばらつきがあり，位置情報は有効ではない．手がかり語として"本研究"，"本論文"，などの"本～"という手がかり語が含まれる文が筆者の最も主張する箇所であり，手がかり語として有効である．さらに"提案した"，"試みた"，などの"本論文"に対する述語部に意味としての共通点がみられ，手がかり語として利用した．また，手法の説明が含まれる文にはタイトル上に含まれる単語を含む文が多く，タイトル上の単語は有効な特徴素である．

(3)結果文について

結果文は実験の結果や，具体的な数値を示す文であり，結論自体があまり長い文章ではないので段落が一つだけのものも多くみられ，結果文が第一段落に含まれているもの

が多く見られたことから位置情報は有効であると考えられる．有効な手がかり語として"結果"，"実験"や"精度"が多くみられた．また，精度を"%"で表した文は具体的な数値として要旨に載せるべきである．他の手がかり語として"確認した"，"できた"，"わかった"などの結果に対する評価を含む文があり，それらを手がかり語として利用する．以上をまとめると表 1，表 2 の様になる．

表 1:位置情報

	動機		結果	
	段落	文	段落	文
第一	92%	92%	80%	16%
第二	4%	69%	6%	54%
第三	3%	26%	3%	16%
第四	1%	8%	0%	5%
なし	0%	0%	11%	8%

表 2:手がかり語

動機	近年，困難，しかし etc
手法	本研究，方法，述語部 etc
結果	実験，結果，精度，評価部 etc

4. 重要度の計算

実際の論文の調査により有効な特徴素とそうでない特徴があることがわかった．そこで，動機文，手法文，結果文の重要度を$W_{moti}(s)$，$W_{meth}(s)$，$W_{conc}(s)$とすると，各重要度計算式は以下のようになる．

$$W_{moti}(s) = C(s) + L(s) \quad (2)$$

$$W_{meth}(s) = C(s) + T(s) \quad (3)$$

$$W_{conc}(s) = C(s) + L(s) \quad (4)$$

本研究ではテキスト構造と文としての意味を考慮し，目的とする文によって利用する特徴素を変える計算式を利用した．

4.1 重要度算出方法

重要度の数値を求める方法として統計的尺度を用いる方法を用いる．学術論文データの序文，結論で出現した特徴素の頻度をもとに単語ごとの重要度を決定する．

$$\text{特徴素の重要度} = \frac{\text{特徴素の出現するテキスト数}}{\text{調査論文の総テキスト数}} \quad (5)$$

この式をもとに文sに対する重要度を決定する．また手がかり語については人手で意味での選別を行ったことから"近代"と"現代"等の同じ意味を示す単語はひとつの単語グループとした．

4.2 抽出文の出力

本システムでは各文の$W_{moti}(s)$，$W_{meth}(s)$，$W_{conc}(s)$を算出し，その中から数値の高いものを順に重要文とし，規定の文字数を越えるまで抽出する．

5. 実験結果

FIT2006 第 5 回情報科学技術フォーラムの論文の中から様々なジャンルの理系論文 80 件に対して実験を行った．評価の方法は，まず人手で動機文，手法文，結果文として抽出されるべき文を選定し，システムが同じ文を抽出していれば正解とする．この際，本研究では書き換えなどの文の一貫性を向上させる処理を行っていないので文章の意味の欠如は考慮しない．さらに一般的に自動要約に用いられるオフライン評価の方法を基に以下の式で精度を示した結果を表 3 に示す．

$$\text{抽出精度} = \frac{\text{システムが正解をだしたテキスト数}}{\text{調査論文の総テキスト数}} \quad (6)$$

表 3:実験結果

	抽出精度
動機	79%
手法	76%
結果	68(93)%
総合	39(53)%

動機文には 79%，手法文には 76%と概ね高い精度が得られ，結果文には 68%という結果になった．これは実験に用いた論文が結果をだしていない論文が 22 件あったことが影響しており，現状の手法の問題点や研究の今後の課題を抽出していた．それら結果を示していない論文を対象外とすると 93%という精度になる．動機，手法，結果の 3 文とも抽出できた文は 39%(結果欠如の文を考慮すると 53%)という結果が得られた．

6. おわりに

本研究では理想的な学術論文のアブストラクトは動機，手法，結果が書かれている文と定義し，それに基づいて利用する特徴素を変える事でアブストラクト自動生成のための重要文抽出を行った．また，動機，手法，結果と文章の意味を考慮することから意味を考慮した手がかり語を利用するアプローチを試みた．実験の結果，動機の抽出率 79%，手法の抽出率 76%，結果の抽出率 93%が得られ，全体で 53%の抽出率を得ることが出来た．

参考文献

[1] 奥村学, 難波英嗣, 知の科学　テキスト自動要約, オーム社, 2005.

[2]Seki, Yohei, "Automatic Summarization Focusing on Document Genre and Text Structure," ACM SIGIR Forum, vol.139, No,1, pp.49-56, 2005.

[3] 電子情報通信学会，学術論文の書き方・発表の仕方，コロナ社，2005.

[4]Edmundson , H.P, "New Methods in Automatic Extracting," Journal of the ACM , Vol.16 , No.2 pp.264-285, 1969.