生徒の検索情報を利用した講義の重要語抽出

A Method for Keyword Extraction Using Retrieval Information from Students in Lectures

篠原 修二
Shuji Shinohara

アドバンストアルゴリズム & システムズ
Advanced Algorithm & Systems
shinohara.shuji@cocoa.plala.or.jp

入部 百合絵
Yurie Iribe

豊橋技術科学大学 情報メディア基盤センター
Information and Media Center, Toyohashi Univ. of Tech.
iribe@imc.tut.ac.jp

河嶋 宏明
Hiroaki Kawashima

豊橋技術科学大学大学院工学研究科知識情報工学系
Department of Knowledge-based Information Engineering, Graduate School of Engineering, Toyohashi Univ. of Tech.
kawashima@vox.tutkie.tut.ac.jp

桂田 浩一
Kouichi Katsurada

(同 上)
katurada@tutkie.tut.ac.jp

新田 恒雄
Tsuneo Nitta

(同 上)
nitta@tutkie.tut.ac.jp

keywords: learning support system, retrieval information, keyword extraction, lecture information, technical term

Summary

Recently, e-learning systems for self-learning with various types of retrieval functions have been developed. This paper describes a method for keyword extraction using retrieval information stored from many students through the retrieval functions. Firstly, we show that (1) teachers tend to consider technical terms as important, while students unfamiliar with the technical terms tend to retrieve the terms, therefore (2) there is a clear correlation between keywords extracted by the teachers and the retrieval words by the students. Secondly, we propose a method utilizing retrieval information from the students for keyword extraction, and show that the method can achieve quite better performance than a method extracting keywords using only lecture information.

1. ま え が き

近年、ネットワークの高速化やマルチメディア環境の普及に伴って、教育分野においてもコンピュータの有効活用が求められている [吉野 98, Harasim 99]。このような背景から、最近では Web 教材を用いた e-learning, すなわち WBT (Web-Based Training) の研究が盛んである [先進 03, 教育 01]。

講義の復習などに用いられる、自主学習向け WBT システムの多くは、授業中の教師音声やデジタルインクを用いた板書、講義スライドなどの講義情報を、Web 上で忠実に再現する機能を持つ。このような講義再現システムを利用することによって、生徒は授業中に聞き逃した部分や、理解が困難であった部分を繰り返し学習することができる。

最近では、講義再現機能に加えて、様々な学習支援機能を備えたシステムも開発され始めている [Katayama 99, 山田 02, Iribe 06a, 入部 06b]。たとえば VIEW Classroom [Katayama 99] は、文字列検索、マウスから

の書き込み箇所の検索など、多様な検索機能を提供する。Iribe ら [Iribe 06a] は、各生徒の理解状況に応じたきめ細やかな支援を実現するシステムを提案している。このシステムでは、受講中の各生徒から取得した情報 (覚醒度など) を、生徒自身が復習を行う際に利用する。たとえば、その生徒が居眠りしていた時間帯に行われた講義部分を色分けして表示したり、キーワード検索の結果として、居眠り時間帯の講義スライドを優先的に提示するなどの支援を行う。これらの機能をうまく活用すれば、必要な部分だけを集中的に学習できるため、生徒は効率的な復習を行える。

しかしながら、以上の支援機能は、全ての生徒にとって有効であるわけではない。たとえば授業を欠席してしまった生徒には、検索機能を使うにしても、そもそも何を検索すればよいかわからない、という問題がある。つまり、検索機能を用いるためには、講義の筋道を理解している必要があるため、通して聴いた後でないとその機能は使い難い。

そこで、もし「講義においてどのような語句が重要で

あり、各語句について理解するためには教材のどの部分を見ればよいのか」ということが予めわかれば、欠席者も効率よく学習できるだろう。またこのような情報は、講義をはじめて聴くことになる次年度の受講生の予習などにも同様に役立つと期待できる。我々はこれまで

(1) 講義における重要語を自動的に抽出し、
(2) 各重要語の理解に必要なスライドを提示するような機能をもつシステムについて検討してきた [篠原 07]。本稿では特に (1) に焦点を絞り、講義情報からの重要語抽出法について議論を行う。

2. 検索情報と重要語

重要語抽出には、これまで $tf \cdot idf$ をはじめ様々な手法が用いられてきた。最も単純な手法として、出現頻度の高い語を重要語とみなす方法が考えられる。しかし、この方法には一般的な語も重要語として抽出してしまうという問題点がある。これに対し $tf \cdot idf$ では、出現頻度が高いだけではなく、対象となる文書に固有な語句ほどその文書にとって重要であるという観点から重要語の抽出を行う。この手法では、ある語句がその文書に特徴的な語か否かを判定するために、大量のコーパスデータが必要となる。最近では、コーパスを用いずに、対象とする文書のみからの情報から重要語を抽出できる手法も提案されている。

たとえば、松尾ら [松尾 02] の手法では、対象とする文書から頻出語を抜き出し、その頻出語と各語の共起頻度を求める。この共起頻度の偏り具合を、その語が重要語であるかどうかの指標として用いることで、対象となる文書単独の情報のみから、比較的高い精度での重要語抽出を実現している。この手法は、論文などの文書を対象としたものであり、語の共起頻度を求める際、1 文をひとつのまとまりとして扱う。しかし、講義スライドを対象とする場合、スライド上に書かれた文字列は、文という単位になっていないことも多い。1 スライドをひとつのまとまりと捉える方法も考えられるが、教師がひとつのトピックを複数枚のスライドを用いて説明することもあり、文書をどのように分節化するかについては恣意的とならざるを得ない。またひとつの講義において使用されるスライド数は、論文などに出てくる文数よりも圧倒的に少ないため、統計処理の精度も悪くなると考えられる。以上の理由から、この手法を講義スライドからの重要語抽出に直接適用することは難しいと考えられる。

さて、先に示したように、最近の WBT システムの多くは検索機能を備えている。我々もまた、検索機能を備えた講義再現システムを開発してきた [入部 06b]。図 1 に、そのシステムの画面例を示す。画面右下の部分に、検索機能が実装されている。

講義再現システムは、授業後、復習をする際に講義ビデオと講義スライドを繰り返し見ることで、ノートの見

図 1 検索機能を持つ講義再現システム

直しだけではなくマルチメディア教材を利用した効率的な復習を支援するものである。このようにアーカイブ化した講義コンテンツを蓄積し、学習に役立てる試みは数多く行われている [村上 05, Med]。本稿ではこのような講義再現システムに搭載された検索機能により、生徒が授業中に聞き逃した箇所やよく分からなかった内容に関連するキーワードを入力し、検索された講義ビデオ（講義を収録したビデオ）と講義スライド（講義で利用した PowerPoint）を見直すことで復習を進める学習形態を想定している。学生は学習を効率的に進めるために、検索機能により各々が必要としている講義コンテンツを探し出すという行為を行っており、講義コンテンツの検索に関する研究や検索機能を備えた講義コンテンツ管理ソフトは数多く存在する [小林 05, 岡本 02, EZP, Pow]。本稿では、我々が構築した講義再現システム上で本手法を実現することを前提とし、重要語抽出に対する提案を行っている。

このようなシステムを利用すれば、多数の生徒が学習に利用する講義コンテンツ（講義ビデオおよび講義スライド）の検索行為から、検索語とその語句がどれくらいの生徒によって検索されたかという検索情報を容易に取得・蓄積できる。重要語抽出の際、この検索情報をコーパス代わりに利用することができないだろうか。

通常の文書と同様、講義においても、教師が特に意識しない限り、頻出語が必ずしも重要語となるわけではないだろうし、逆に重要語であっても教師がその語を何度も使うとは限らないだろう。つまり講義における重要語抽出においても、出現頻度は高いけれども重要ではない語を除去したり、あまり用いられないけれども重要である語を抽出することが求められる。

翻って、生徒がどのような語を検索するかを考えてみよう。生徒は、講義に多用される語であっても、意味が自明である一般語などは講義コンテンツの中からわざわざ検索しないと予想される。また、教師が一、二度しか使わなかった語であってもそれが重要であると考えれば、復習の際に検索機能を用いて、その語に関連する講義コ

ンテンツを再確認する可能性が高いと思われる。もし実際にこのような傾向があるならば、検索情報を語句の重要度に反映させることによって、重要語抽出の精度向上が期待できよう。

以上の観点から、本稿では、第一に被験者実験を行って、多くの生徒によって検索される語と教師がその講義において重要だと考える語の間にどのような関係があるのかについて分析する。第二に、生徒から取得した検索情報を、語の重要度に反映させることによって、重要語抽出の精度向上を試みる。

3. 実験

本章では、被験者実験を行って検索情報と語の重要度を抽出し、両者の関係について分析する。

3.1 実験設定

実験では、4 人の教師が 20～30 枚のスライドを使った 30 分程度の講義（各々 L1, L2, L3, L4 と呼ぶ）を行い、これらの講義を初めて聴く 23 人の学生被験者がノートを取りながら受講した。4 つの講義の内、L1～L3 は各々「ゲームプログラミング」「音声認識」「発達心理学」に関するものであり、専門用語が多く使われている。一方 L4 は対照実験として行われたものであり、あるシステムの内容を、一般の人にわかりやすく紹介するプレゼンテーションであった。このため、専門用語はほとんど使われていない。各被験者は、受講後、復習の際に検索したいと思う語^{*1}（「検索語」と呼ぶ）を抽出した。一方教師は、講義における重要語を抽出し、これらの語に対して三段階に区分された重要度（1：やや重要、2：重要、3：非常に重要）を付与した。なお、全ての検索語の内、重要語として抽出されなかった語に関しては、重要度 0 とした。また各講義の後、被験者の理解度を測定するためにテストを行った^{*2}。テストは、各講義を行った教師が各々作成し、穴埋め問題や選択問題、記述形式の問題を含む 6～10 の設問から構成されていた。

3.2 結果と考察

図 2～図 5 に、各講義における実験結果を示す。これらの図で、横軸は検索人数、縦軸は語数を表す。「検索語」で表されたヒストグラムは、横軸で表された人数の被験者によって検索語として挙げられた語が何個あったのかを表す。また「重要語」で表されたヒストグラムは、検索語の内、教師が重要語と判定した語が何個あったのかを表す。さらに「重要語率」で表された折れ線グラフは、

^{*1} 各被験者には、講義 L1～L4 を受講している際に、よく理解できなかった内容や聞き逃した箇所など、後で再度聞き直したい講義内容に関連する語（即ち、後で検索しそうな語）を全て記述してもらった

^{*2} 被験者の事情により、テストを受けたのは 23 人中 16 人であった。また、テストの制限時間は特に設けなかった。

図 2 被験者実験における検索情報と重要語の関係：L1

図 3 被験者実験における検索情報と重要語の関係：L2

検索語の内での重要語の割合（重要語数/検索語数）を表す。たとえば図 2 において、検索人数が「21 以上」の箇所で検索語数は 6、重要語数は 5 となっている。これは、23 人の被験者の内の 21 人以上によって検索語として挙げられた語は全部で 6 個あり、それらの内で教師が重要語と判定したものは 5 個あったことを示している。この時、重要語率は、 $5/6 \approx 83\%$ となる。これらの図から、検索人数の増加に伴って重要語率が上昇する傾向、つまり、多くの被験者によって検索される語は重要語である可能性が高い、という傾向が見て取れる。

「重要語」のヒストグラムに着目してみると、L1 においては、右上がりの大まかな傾向が見られるのに対し、L4 では左上がり、つまり検索人数の少ない語の語数が多い

図 4 被験者実験における検索情報と重要語の関係：L3

図 5 被験者実験における検索情報と重要語の関係：L4

図 6 語の重要度と検索率の関係

くなる傾向がある。これは、L4 ではあまり検索されない重要語が多くあるということを意味する。換言すれば、どのような重要語を検索するかは被験者毎にばらつきがある。

以上に示した結果から、「多くの被験者によって検索される語は重要である」とはいえるが、「重要語は多くの被験者に検索される」とは必ずしもいえないことがわかる。

次に図 6 に、各講義における語の重要度と平均検索率の関係を示す。この図で、横軸は教師が各語に付与した重要度を表す。また縦軸の「平均検索率」とは、横軸で表された重要度を持つ語が、平均でどれくらいの割合の被験者に検索されたのかを表す。この結果から、重要度の高い語は、多くの被験者に検索される傾向にあることがわかる。特に L1～L3 において、重要度 3 の語は、ほとんどの被験者（80% 程度）の検索対象となっている。例として、表 1 に L1 における重要度 3 の語と各語に対する検索率を示す。L1 は「ゲームプログラミング」に関する講義であるが、重要度 3 の語には、一般の人にとってはなじみのないものが多い。この点は、L2, L3 においても同様であった。つまり、専門性の高い大学の講義において、教師は専門用語に高い重要度を付与する傾向にあるといえる。

次に、重要語と検索語の関係を大まかに見てみよう。「多くの被験者に検索される語」における「多く」は曖昧であるが、ここでは被験者全体の 1/3 以上に検索された語

表 1 L1 における重要度 3 の語と検索率

重要語	検索率 (%)
ミニマックス法	100
- 法	96
二人零和確定完全情報ゲーム	91
カット	74
カット	74
ゲームプログラミング	61
値	57
値	57

表 2 重要語と検索語の関係

講義	$P(\text{重要語} \text{検索語})$	$P(\text{検索語} \text{重要語})$	調和平均
L1	.67	.95	.78
L2	.79	.82	.81
L3	.85	.88	.86
L4	.79	.43	.56

を「検索語」とみなす。表 2 に教師の抽出した重要語と検索語の関係を示す。表中で、 $P(\text{重要語} | \text{検索語})$ と $P(\text{検索語} | \text{重要語})$ は各々、検索語中での重要語の割合、重要語中での検索語の割合を表す。また表の右端に $P(\text{重要語} | \text{検索語})$ と $P(\text{検索語} | \text{重要語})$ の調和平均を示す。この表から、調和平均は、L1～L3 において 0.8 前後の高い値を示すことがわかる。一方、L4 においては、 $P(\text{検索語} | \text{重要語})$ の値が 0.43 と低いため、調和平均の値はそれほど高くない。

この理由について分析するために、L4 の内容について詳しく見てみよう。表 3 に、L4 における重要語の中で、検索率の最も低かった語（検索率 4.3%）の一部を示す。なお表には、各語の重要度も併せて表示している。L4 は、「パソコン初心者ソフト操作学習をサポートするために、適切な学習用教材を自動的に作成し、彼らの操作誤りを監視しながら、その誤りを的確に指摘するシステム」の紹介であった。このため教師は、「操作学習サポート」「教材自動作成」「監視」「誤り」などの語句を重要語として選んでいる。しかし被験者は、これらの語の意味は容易に推定できるため、わざわざ検索する必要を認めなかったと考えられる。これが L4 において、 $P(\text{検索語} | \text{重要語})$ の値が低くなった一因であると考えられる。

更に L4 について分析してみたところ、教師は「教材自動作成」という語を非常に重要な語として挙げたにもかかわらず、講義ではこの語句をほとんど使用していなかった。これは、教材を自動的に作成するシステムのことを、講義の中では「教材作成システム」とも呼んでいたためである。このように、一般的な概念を表す語は様々な言い換えが可能であり、教師が講義を行う際にも、生徒が検索を行う際にも表現のゆらぎが生じる。これが、L4 において、被験者の重要語の検索にばらつきが見られた要

表 3 L4 において検索率の低かった重要語

重要語	重要度
教材自動作成	3
操作学習サポート	2
監視	1
誤り	1

図 7 再現率と理解度テストの相関

因であると考えられる。一方、たとえば L1 における重要語「 - 法」で表現される概念などは、「 - 法」以外の語で言い換えることは困難である。つまり L1～L3 においては、専門用語には同義語が存在しないため、上記の問題が生じ難かったと考えられる。以上の結果から、専門用語が多く使われる大学の講義に限れば、検索情報が重要語抽出に有効であると期待できる。

次に図 7 に、各被験者の理解度テストの得点と再現率の関係を示す。ここで再現率とは、教師の抽出した重要語の内、各被験者が検索語として挙げた語の割合である。図で横軸は、各被験者の L1～L4 のテストの平均点を表す。また縦軸は、各被験者の L1～L4 に対する再現率の平均値を示す。16 人の被験者がテストを受けたので、図には 16 個の点がプロットされている。

図から、テストの得点が高い被験者ほど再現率も高いという大まかな傾向が見て取れる（相関係数 0.81）。また、各被験者が検索語として挙げた語の数と再現率の間にも、非常に強い相関が見られた（相関係数 0.93）。一方、適合率とテストの得点の間には相関は見られなかった（相関係数 0.096）。ただし適合率とは、被験者が検索語として挙げた語の内の重要語の割合である。ここで相関が見られなかったのは、各被験者の適合率が、テストの点数にかかわらず、ほぼ一定（平均 59.2%）であったためである。

これらの結果から、テストの得点の低い被験者の再現率が低かったのは、彼らが検索語として重要ではない語を数多く挙げていたためではなく、検索語として挙げた語の数自体が少なかったためであることがわかる。つまり、これらの被験者は、講義内容が理解できなかったというよりはむしろ、居眠りや他の考え事をするなどして

集中力を欠いていたため、講義の重要な部分を聞き逃していたと思われる。このように、テストの得点の低い被験者であっても、的外れの語を数多く検索語として挙げるわけではない。このため、様々な生徒から検索情報を収集する実用システムにおいても、重要語に関して十分精度のよい情報が得られると期待される。また、重要語の再現率と理解度テストの得点の間に正の相関があるということは、我々の提案する重要語提示システムが、生徒の理解度を高めるための有効な手段になり得ることを示唆する。

4. 検索情報を用いた重要語抽出

本章では、講義情報（教師の音声データ、スライドのテキストデータ）と生徒の検索情報から、講義における重要語を抽出する手法について検討する。

4.1 講義情報の取得

講義情報として、教師の音声データとスライドに記述されたテキストデータから、語とスライドの共起情報を抽出する。そのために、まず音声データ^{*3}とスライドデータに対して形態素解析を行って単語を取り出した。形態素解析には茶筌 [cha] を用いた。取り出す単語の品詞は基本的に、名詞と未知語とした。ただし、代名詞や非自立語を除くなどの設定を行っている。また、名詞や未知語が連続する場合はそれらを結合するという処理も行った。たとえば、「音声」「認識」という二つの名詞が連続する場合には、それらを結合し、「音声認識」という一つの語句として取り出している。

次に、このようにして教師の講義データから取り出された全ての語に関して、スライドとの共起頻度を求める。ここで語 w とスライド s の共起頻度を $Freq(w, s)$ と表記する。つまり $Freq(w, s)$ とは、スライド s を説明する際、教師が語 w を発話した、あるいはスライド s に語 w という文字列が書き込まれていた頻度である。この時、講義全体での語 w の出現頻度は、 $Freq(w) = \sum_{s \in S} Freq(w, s)$ となる。ただし、 S は教師が講義で用いたスライド集合を表す。

4.2 重要語抽出

本節では、講義情報と検索情報から、語の重要度を求める手法について述べる。まず、講義情報と検索情報から導出された語 w の重要度を、各々 $L(w)$ 、 $R(w)$ と表記

^{*3} 音声データに関して、音声認識誤りにおける実験の煩雑化を避けるため、教師自身による書き起こしデータを用いた。本提案を実装する図 1 の検索機能は、予め音素列でテキスト化した音声データから DP マッチングによりキーワードに該当する音声部分を検索している。DP マッチングは、音素弁別特徴ベクトル間の距離に基づいて行われるため、類似の音素列に置換された場合も近いキーワードを検出できる。上記の手法を利用することにより、手作業による書き起こし作業の負担は減少すると考えられる。

表 4 重要語抽出結果

講義	TH	α	再現率 (%)	適合率 (%)	F 値 (%)
L1	.46	.72	84(16/19)	84(16/19)	84
L2	.21	.58	85(28/33)	82(28/34)	84
L3	.28	.72	88(22/25)	96(22/23)	92
L4	.11	.86	66(29/44)	51(29/57)	57

する．これら二つの重要度から，以下の式を用いて語 w の重要度 $I(w)$ を算出する．

$$I(w) = (1 - \alpha)L(w) + \alpha R(w) \quad (1)$$

ここで α ($0 \leq \alpha \leq 1$) は，検索情報をどの程度，語の重要度に反映させるかを表すパラメータである． $\alpha = 0$ の場合，講義情報のみから語の重要度を算出するということを意味する．

$L(w)$ と $R(w)$ に関して，様々な尺度が考えられるが， $R(w)$ に関しては，何らかの語を検索した生徒の内，語 w を検索した生徒の割合を用いることにする． $L(w)$ に関して，講義データにおける語 w の出現確率 $P(w) = \text{Freq}(w) / \sum_{w' \in W} \text{Freq}(w')$ を用いる方法が考えられる．ここで， W は講義データから取り出された語の集合である．しかしこの方法では，講義が長時間になり，データに出現する語の種類が多くなればなるほど，語の重要度が相対的に低くなってしまふ．一般的には，講義時間の長さに比例して，重要語が多くなると考える方が自然だろう．この点を考慮し，本稿では最も出現頻度の高い語の L の値が 1 となるように $\text{Freq}(w)$ を正規化する．すなわち， $L(w) = \text{Freq}(w) / \max_{w' \in W} \text{Freq}(w')$ とする．以上に示した重要度を用いて， $I(w)$ の値が閾値 TH 以上ならば，語 w を重要語として抽出する．

4.3 評価と考察

本節では，3 章の被験者実験で行われた講義に対して，重要語抽出を行った評価結果を示す．

まず，提案手法によって重要語として出力された語の内，教師が重要語であると判定した語の割合，すなわち $P(\text{重要語} | \text{出力語})$ を適合率，教師が重要語と判定した全ての語の内，提案手法が出力できた語の割合，すなわち $P(\text{出力語} | \text{重要語})$ を再現率とする．次に，F 値として両者の調和平均をとり，手法の性能評価のために用いる．

表 4 に，各講義において最も F 値が高くなるようなパラメータ (TH, α) の値とその時の再現率，適合率，F 値を示す．なお，表中の再現率，および適合率の括弧内には頻度，すなわち (出力語数/全重要語数) (重要語数/全出力語数) も併せて示している．F 値は，L1～L3 において 80% 以上の高率を示す．図 8 に， $\alpha = 0.7$ に固定し，閾値 TH を変えた場合に F 値がどのように変化するかを調べた結果を示す．ただし TH は $[0, 1]$ 区間において 0.01 刻みで変化させた．L1～L3 に関しては，

TH の間が中程度の時，F 値が高くなる．L1～L3 の F 値の平均値は， $TH = 0.27$ の時に最大となる．一方 L4 においては， TH が 0.1 程度の低い値で F 値は最大となる．これは，L4 では被験者毎に検索語が分散しているため他の講義に比べて $R(w)$ の値が小さい，という理由による*4．

図 9 に， $\alpha = 0.7, TH = 0.27$ とした場合の講義毎の F 値を示す．また比較のため， $\alpha = 0$ ，すなわち検索情報を全く考慮せず，講義情報のみから重要語抽出を行った結果も併せて示す．ただし $\alpha = 0$ の場合の閾値は，L1～L4 の F 値の平均が最大となるよう $TH = 0.04$ に設定した．この結果から，語の重要度に検索情報を反映させることによって，重要語抽出の性能が大幅に改善されることがわかる．特に L1～L3 においてその傾向は顕著である．

図 9 では，パラメータを $\alpha = 0.7, TH = 0.27$ に固定した．一方，表 4 では，最も F 値が高くなるようなパラメータ (TH, α) の値にばらつきがある．しかし，図 9 の「検索情報考慮」の部分を見ると，F 値は表 4 に示した最高値から数%の下落に留まっている．また図 8 を見ると，L1～L3 では，かなり広い範囲の TH 値 (0.2～0.4) で，比較的高い F 値が維持される．より数多くの講義データを分析する必要があるが，L1～L3 のような専門的な講義に限れば， $\alpha = 0.7$ 程度， $TH = 0.3$ 程度の値で，比較的高いパフォーマンスが得られると考えられる*5．

また，十分な精度のオープンテストを行うためにクロスバリデーション (CV) による評価を行った．図 10 に，その結果を示す．図 9 の結果と同様に，L4 は他の講義と比べて F 値が低い，L1～L3 においては全体的に抽出精度が良い．この結果から，実際の検索情報を利用した場合も，専門用語がほとんど使用されていない講義を除いては，比較的高い精度の重要語を抽出できる可能性が高いことがわかる．

本稿では，講義データから重要度を抽出する際，最も単純に語の頻度確率を用いたが， $L(w)$ として $\text{tf} \cdot \text{idf}$ *6をはじめ，より高性能な様々な手法 [鈴木 06] を採用することも可能である．この点については今後の課題としたい．

ただし，本稿での主な主張点は「(何らかの手法で) 講義データから抽出された重要語情報に，検索情報を加味することによって，重要語抽出の性能を向上させる」ということであり， $\text{tf} \cdot \text{idf}$ などの手法とは相補的な関係にある．講義データから重要度を抽出する部分をより高性能な手法に置き換えれば，図 9 に示した「検索情報なし」の F 値が上昇すると予想される．しかしその場合でも，

*4 L1～L4 における重要語の R 値の平均は，各々 0.71, 0.55, 0.62, 0.30 であった．

*5 本稿では，講義データから重要度を抽出する際，最も単純に語の頻度確率を用いたが，より高性能な手法を採用すれば， α 値は下がることが予想される．

*6 現時点では，様々な教師が行った講義データ (スライド，音声) の入手が困難であり， $\text{tf} \cdot \text{idf}$ の評価は行っていない．

図 8 閾値 TH と F 値の関係 ($\alpha = 0.7$)

図 9 検索情報を考慮した場合としない場合の性能比較

検索情報は個々の講義に特化した情報であり、コーパスに比べてより直接的な情報を与えるため、検索情報を考慮することによって ($\alpha > 0$ とすることによって), 更なる性能向上が期待できる。

5. む す び

本稿では, 教師音声やスライドデータなどの講義情報と生徒の検索情報から, 重要語を抽出する手法について検討した。

まず被験者実験を行って, 教師が重要だと考える語と生徒の検索する語の関係について分析した。その結果, 専

図 10 クロスバリデーションによる性能評価

門用語が多く使われる講義においては, 両者の間に強い相関が見られることがわかった。より正確に言えば, 「多くの生徒に検索される語は重要語である」とはいえるが, 「重要語は多くの生徒に検索される」か否かは講義内容による。専門用語がほとんど使用されない講義においては, 重要語であるからといって, 多くの生徒に検索されるとは限らない。これは, 重要語が一般的な語である場合, 「一般語は他の語を使って様々な言い換えが可能のため, 教師や生徒の表現が一貫しない」, 「一般語は生徒にとってなじみのある語であるため, それを検索するか否かには個人差がある」などの理由によって, 検索語が生徒毎に分散するためである。一方, 専門用語が多く使われる講義では, 「教師は専門用語に高い重要度を付与する」, 「多くの生徒は, なじみのない専門用語を検索する」などの傾向があり, 検索語と重要語の間に強い相関が見られることになる。

次に講義情報とともに検索情報も考慮した重要語抽出手法を提案し, その評価を行った。その結果, 検索情報を考慮しない場合に比べて大幅な性能改善が見られた。

今後は, 我々が開発してきた講義再現システム上に本提案を実装し運用することで, 実践的评价を行っていく予定である。その際には, 音声の書き起こしデータを利用しない場合の本提案の性能評価も行うつもりである。

我々は, 生徒から取得した検索情報を重要語抽出に利用する手法について検討してきたが, この検索情報は教師の授業改善のためにも活用できると考えている。本稿の被験者実験では, 専門的な講義において, 生徒の検索語と教師の重要語の間に相関が見られた。しかし両者の間に大きなずれがあるような講義もあるかもしれない。復習の際, 生徒は重要語を検索しようとする傾向性を持つとすれば, このずれが大きいということは, 教師の意図が生徒に正しく伝わっていない可能性を示唆する。このような検索情報と重要語のずれに関する情報を, 教師の授業改善に役立てるシステムも開発していきたい。

謝 辞

本研究の一部は, 文部科学省 21 世紀 COE プログラム「インテリジェントヒューマンセンシング」の補助による。

◇ 参 考 文 献 ◇

- [cha] ChaSen's Wiki - FrontPage, <http://chasen.naist.jp/hiki/ChaSen/>
- [EZP] EZ-Plat: 製品概要 製品・システム | 日立 AD, <http://www.hitachi-ad.co.jp/ezplat/>
- [Harasim 99] Harasim, L.: A Framework for Online Learning: The Virtual U, *IEEE Computer*, Vol. 32, No. 9, pp. 44-49 (1999)
- [Iribe 06a] Iribe, Y., Shinohara, S., Matsuura, K., Taguchi, R., Katsurada, K., Sekiguchi, Y., and Nitta, T.: Development of Self-Learning Support System Using Arousal Level and Lecture Information, *International*

Journal of Computer Science and Network Security, Vol. 6, No. 8A, pp. 11–17 (2006)

- [入部 06b] 入部 百合絵, 松浦 喬一, 篠原 修二, 桂田 浩一, 新田 恒雄: 受講情報を利用した学習支援機能付き講義再現システムの開発, 信学技法, Vol. ET2005-95, pp. 41–46 (2006)
- [Katayama 99] Katayama, K., Kagawa, O., Kamiya, Y., Tsushima, H., Yoshihiro, T., and Kambayashi, Y.: Use of Action History Views for Indexing Continuous Media Objects, in *Proceedings of AMCP'98*, pp. 345–355 (1999)
- [小林 05] 小林 隆志, 村木 太一, 直井 聡, 横田 治夫: 統合プレゼンテーションコンテンツ蓄積検索システムの試作, 電子情報通信学会論文誌 D, Vol. J88-D1, No. 3, pp. 715–726 (2005)
- [教育 01] 教育システム情報学会 (編): 教育システム情報ハンドブック, 実教出版, 東京 (2001)
- [松尾 02] 松尾 豊, 石塚 満: 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム, 人工知能学会論文誌, Vol. 17, No. 3, pp. 213–227 (2002)
- [Med] 富士ゼロックス e ラーニングサービス MediaDEPO, <http://www.ubiquitous-media.com/technology/index.html>
- [村上 05] 村上 正行, 西口 敏司, 亀田 能成, 角所 考, 美濃 彦彦: 京都大学での実践に基づく講義アーカイブの調査分析, 日本教育工学会論文誌, Vol. 28, No. 3, pp. 253–262 (2005)
- [岡本 02] 岡本 拓明, 小林 隆志, 直井 聡, 横田 治夫, 古井 貞照: 講義講演シーン検索における音声データの利用, 信学技法, Vol. DE2005-107 (2005-7), pp. 1–6 (2002)
- [Pow] フォトロン: e-ソリューション | Power Rec, <http://www.photron.co.jp/products/e-solution/rec/index.html>
- [先進 03] 先進学習基盤協議会 (ALIC) (編): e ラーニング白書 2003/2004 年版, オーム社, 東京 (2003)
- [篠原 07] 篠原 修二, 入部 百合絵, 河嶋 宏明, 桂田 浩一, 新田 恒雄: 重要語句に対する関連スライド群抽出法の検討, 信学技法, Vol. ET2006-101 (2007-1), pp. 105–110 (2007)
- [鈴木 06] 鈴木 大介, 内海 彰: Support Vector Machine を用いた文書の重要文節抽出, 人工知能学会論文誌, Vol. 21, No. 4, pp. 330–339 (2006)
- [山田 02] 山田 博文, 松田 和彦, 田口 亮, 桂田 浩一, 小林 聡, 新田 恒雄: 講義再現システムにおけるスライド重要度抽出, 人工知能学会論文誌, Vol. 17, No. 4, pp. 481–489 (2002)
- [吉野 98] 吉野 孝, 井上 穰, 由井園 隆也, 宗森 純, 伊藤 士郎, 長澤 庸二: インターネットを介したパーソナルコンピュータによる遠隔授業支援システムの開発と適用, 情報学論, Vol. 39, No. 10, pp. 2788–2801 (1998)

[担当委員: 柏原 昭博]

2007 年 3 月 19 日 受理

著 者 紹 介

篠原 修二

1999 年神戸大学大学院自然科学研究科博士後期課程修了。博士 (理学)。2005 ~ 2007 年豊橋技術科学大学で生体情報をを用いた教育支援システムの開発に従事。現在、画像処理 (超解像) や交通渋滞シミュレーションのシステム開発に携わる。日本認知科学会, 進化経済学会各会員。

入部 百合絵 (正会員)

1999 年名古屋工業大学工学部卒業。2001 年名古屋大学大学院人間情報学研究所修士課程修了。2004 年同大学院人間情報学研究所博士課程単位取得退学。博士 (学術)。現在、豊橋技術科学大学情報メディア基盤センター助教。教育支援, ユーザインタフェース等に関する研究に従事。情報処理学会, 教育システム情報学会各会員。

河嶋 宏明

2005 年大島商船高等専門学校情報工学科卒業。2007 年豊橋技術科学大学工学部知識情報工学課程卒業。音声認識, Web 教育システムに関する研究に従事。

桂田 浩一 (正会員)

1995 年大阪大学基礎工学部情報工学科卒業。2000 年同大学院基礎工学研究科博士後期課程修了。同年豊橋技術科学大学工学研究科助手。2007 年同講師。現在に至る。博士 (工学)。音声・マルチモーダル対話, 知識処理に関する研究に従事。AAAI, 情報処理学会, 日本音響学会, 言語処理学会, ヒューマンインタフェース学会各会員。

新田 恒雄 (正会員)

1969 年東北大学工学部電気工学科卒業 (株) 東芝総合研究所, 同社マルチメディア技術研究所を経て, 1998 年豊橋技術科学大学大学院工学研究科教授。工学博士。音声認識・合成・文字認識, マルチモーダル対話システム, および概念獲得の研究に従事。IEEE, 電子情報通信学会, 情報処理学会, 日本音響学会各会員。