社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DD－64 (6)
2008－EIP－39 (6)
2008／1／31

教材スライドから自動抽出した重要語に基づく教材改良支援

下園 幸一†, 菅沼 明††

† 鹿児島大学学術情報基盤センター
††九州大学大学院システム情報科学研究院

我々は授業を効果的に行うために、授業で用いられるスライドの改善支援に関する研究を行っている。学生の理解モデルを計算機上で実装することにより、どのように学生がスライドを理解するかの情報を抽出し、授業スライドの改善に役立てたい。今回は、対数尤度比検定の一つであるG検定の値を用いて、授業スライドと語の共起関係から自動的に授業スライドから重要語と思われる語の抽出を行った。この抽出した重要語とスライド作成者の意図する重要語とを比較することで、教材スライド改善に役立てる事ができる。

Supporting System for Improvement of a Teaching Material based on Important Words Extracted from Teaching Material Slides

Koichi Shimozono†, Akira Suganuma††

† Computing & Communications Center, Kagoshima University.
†† Faculty of Information Science and Electrical Engineering, Kyushu University

For effective teaching, we are researching a supporting system for improvement of teaching material slides. We have tried to extract information how a student understands slides by the student model on the computer, and study it to use for improvement of teaching material slides. Now, the method of automatic extraction of the words that seem to be important words from teaching material slides is developed. It is possible to use it for improvement of teaching material slides by comparing them with important words which a teacher intends.

1. はじめに

近年、大学では、授業を効果的に行うためにさまざまな取り組みがなされている。その１つにe-Learning システムの導入がある。しかしながら、システムや設備に多大なコストがかかる事や、教員へ教材作成の負担がかかるため、一部の授業でのみ利用されている場合も多い。また、導入コストと比較して、その効果を疑問視する声も存在する。現状では、授業で利用する講義スライド(PowerPoint ファイル)を授業開始以前に WEB 上で公開するのみにとどまっている場合が多い。

従来の講義ノートを板書して解説を行う授業スタイルでは、学生は自分のノートに板書された

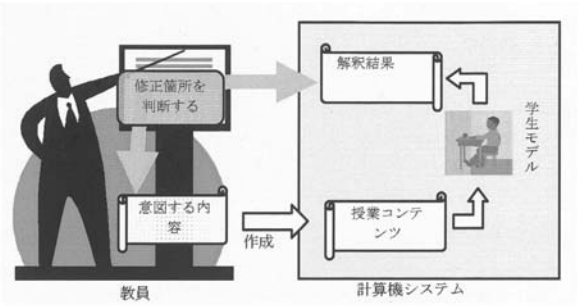

図 1: システムの概要

ものを書く時間がある。しかしながら、スライドを用いた授業スタイルでは、スライドの内容をノートに書く時間が十分にない。そのため、事前にスライドを印刷して授業を受けなければならない。また、事前にスライドが配布されていれば、予習を行うことも可能である。

このような場合、授業スライドの質を向上させることが重要となってくる。本研究では、計算機を用いて授業スライドの改善支援を行う。

2. スライド改善支援手法について

教員が授業スライドを作成する際、まず考えることは「学生が授業スライドをどのように理解するか」である。実際に授業を行ってみて、学生の反応を見ることにより、授業スライドに変更を加える場合もある。つまり、『授業スライドの質』とは「どれだけ学生がそのスライドを理解できるか」にかかっている。そのため、授業スライドの質を高めるには、まず、学生がどのように授業スライドの内容を理解するかを知ることが重要である。我々は、この「学生がどのように授業スライドの内容を理解するか」を計算機上で推測することにより、教員のスライド改良支援ができるのではないかと考えた。

支援手順としては以下のようになる(図 1)。

1. 教員はスライドを作成する。
2. 学生モデルが、スライドの内容を解析し、学生がどのようにスライドを理解するかを教員に提示を行う。
3. 提示された内容が教員の意図する内容と異なった場合は、スライドの修正を行う。

この 2,3 を繰り返すことによってスライドの質を高める。実際にスライドの修正を行うのは教員であり、システムは学生モデルに従って解析結果を提示するのみである。また、2,3 を繰り返しながらスライドの改善を行うため、計算機上の解析は、なるべく短時間で終わる方がよい。

3. 重要語の抽出法

学生モデルが持つ解析機能の一つとして、我々はまず重要語の抽出を考えた。重要語を抽出するため、実際の授業スライドの分析を行う。

まず、授業スライドに現れる文章から単語の抽出を試みた。日本語の文章は分かち書きをされていないため、単語を抽出するために形態素解析を行わなければならない。形態素解析エンジンには京都大学で開発された MeCab を、解析辞書には IPA で作成された辞書を用いた。

対象とした講義スライドは我々が実際に授業で利用したスライド(表 1)である。CS.1,2,3 は講

義形式の授業であり、CS.4,5 は実際にプログラミングを行う演習形式の授業である。

表 1:解析に利用した講義

科目名	講義回数
システム設計(CS,1)	12
オペレーティングシステム論 (CS.2)	11
情報ネットワーク論 (CS.3)	12
情報処理演習(電気情報) (CS.4)	12
情報処理演習(標準カリキュラム) (CS.5)	12

それぞれの授業で利用した講義スライド数を表 2 に示す。

表 2: 講義毎のスライド数

	No.1	No.2	No.3	No.4	No.5	No.6	No.7	No.8	No.9	No.10	No.11	No.12	合計
CS.1	6	19	20	18	15	16	15	17	13	13	7	6	165
CS.2	5	14	9	13	10	19	10	16	11	13	13		133
CS.3	4	12	21	19	26	12	28	48	32	23	13	9	247
CS.4	76	45	54	43	74	68	59	58	113	73	46	19	728
CS.5	29	22	22	22	24	17	16	14	20	17	15	19	237

授業によりスライド数が大きく異なることが分る。それぞれの講義毎のスライドに出現した名詞の数を表 3 に示す。

表 3: 講義毎の名詞の数

	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	No. 10	No. 11	No. 12	合計
CS. 1	178	520	386	440	369	479	219	300	522	545	139	145	4, 242
CS. 2	94	330	303	297	231	484	322	495	451	584	493		4, 084
CS. 3	122	256	460	468	855	245	843	1, 034	1, 099	773	288	292	6, 735
CS. 4	2, 046	1, 560	1, 199	1, 211	1, 886	1, 721	1, 794	1, 650	2, 286	2, 053	1, 288	799	19, 493
CS. 5	932	925	630	759	773	453	612	401	756	729	565	798	8, 333

それぞれの講義毎のスライドの異なり単語数を表 4 に示す。異なり単語数とはそれぞれの講義毎に何度も出現する単語を 1 つにまとめてカウントした場合の数である。

表 4: 講義毎の異なり単語数

	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	No. 10	No. 11	No. 12	合計
CS. 1	123	277	246	206	190	232	128	187	186	315	103	93	1, 464
CS. 2	76	187	140	130	105	255	190	219	266	315	251		1, 279
CS. 3	100	169	246	283	396	148	445	517	393	409	162	171	1, 885
CS. 4	628	762	205	185	305	226	242	216	234	304	255	324	1, 989
CS. 5	384	534	278	293	244	168	178	106	180	234	157	323	1, 667

これらスライドから取り出した名詞の中から重要語を特定する場合、単に出現頻度の多い単語を重要語として抽出することは簡単である。しかしなから、次のような誤りを侵すと考えた。

1. 全体を通して出現頻度は少ないが、ある特定の回で多く出現している重要語を抽出できない
2. 各回に頻繁に出現する一般名詞を重要語として抽出してしまう

そのため、統計的手法を用いて重要語の特定を行った。

大規模コーパスからある単語同士の共起関係を求める統計的手法として、尤度比検定の一つである G 検定(G スコア)が用いられている。G 検定の一般的な式は

$$G = 2\sum O_i \ln(\frac{O_i}{E_i})$$

で表すことができる。Oi が観測値、Ei が期待値である。2 × 2 分割表を用いて単語 a と単語 b の共起関係を求める場合は以下のような計算になる。

	b の出現数	b 以外の単語の出現数
a の出現数	O_{11}	O_{12}
a 以外の単語の出現数	O_{21}	O_{22}

1. $A=O_{11}\ln O_{11}+O_{12}\ln O_{12}+O_{21}\ln O_{21}+O_{22}\ln O_{22}$
2. $B=(O_{11}+O_{12})\ln(O_{11}+O_{12})+(O_{21}+O_{22})\ln(O_{21}+O_{22})+(O_{11}+O_{21})\ln(O_{11}+O_{21})+(O_{12}+O_{22})\ln(O_{12}+O_{22})$
3. $C=(O_{11}+O_{12}+O_{21}+O_{22})\ln(O_{11}+O_{12}+O_{21}+O_{22})$
4. $G=2(A-B+C)$

この G の値が大きいほど a と b の共起頻度が統計的に多く、関連性が深いといえる。

今回は、各講義で使用するスライドと、ある単語の出現数から G スコアを求めた。これより「ある単語 a は、ある講義とどの程度関係が深いか」を数値で表すことができる。例えば、CS.1 No.1 のスライドに出現する「システム」という名詞の G スコアは 2×2 の分割表(表 5)より 2.537 となる。

表 5:「システム」と「CS. 1 No. 1」との 2x2 分割表

	CS.1 No.1	CS.1 No.1 以外	合計
「システム」の出現数	7	81	88
「システム」以外の出現数	171	3,983	4,154
合計	178	4,064	4,242

この G スコアをスライドに出現する全ての名詞に対して計算を行い、各回毎に単語の順位付け

表 6: 順位付けされた語の例

CS. 2, No. 7	G スコア	39. 42	39. 42	30. 02	15. 53	15. 53	15. 27	15. 27
	単語	FAT	クラスタ	セクタ	MB	権	FD	領域
CS. 4, No. 11	G スコア	334. 21	312. 13	247. 43	141. 78	136. 30	130. 83	118. 75
	単語	node	木	探索	tree	height	二分	二

を行った。表 6 に例を示す。

例えば、CS.4 No.11 の node という単語の G スコアは 334.21 である。これは、CS.4 の他のスライドより No.11 での node という単語の出現頻度が非常に高いことを表している。つまり node という単語が No.11 での重要語であると考えられる。

4. 重要語の提示方法とスライドの修正方法

今回のシステムでは、学生モデルに基づき計算機が重要語と思われる語を各講義のスライド毎に抽出している。実際に重要語かどうかを判断するのは教員である。そのため、抽出結果を基に

した具体的なスライドの修正方法として、以下のような手順を考えた。

1. あるスライドの解析結果をGスコア順に20語程度表示する。
2. 教員が重要語にマークを行う。リストに存在しない重要語は入力する
3. マークされた語に関して、リストの上位に入るには何語程度増やせばよいかを再計算する。提示された情報を基にスライドを改良する
4. 再計算を行い最終的な値を提示する。満足できない場合は3,4 を繰り返す

3における何語程度増やせばよいかの計算では、マーク語総数より順位が低い語のみGスコアの再計算を行い、マーク語総数より順位が高くなるだけの語数と予想される順位を提示する。例えば、マークされた語が10語ある場合、順位10番以降でマークされた語に対して、順位10番の語のGスコア以上になるだけの語数をそれぞれ計算する。実際には、Gスコアから語数を求める逆関数を作成することは容易ではない。そのため、語数を少しずつ増やしながらGスコアを計算し、規定のGスコアを超えた場合の語数とその際の順位を提示する。

実際の例として、CS.2 No.7 を用いて修正手順を示す。まず、計算機が単語とその単語の順位を教員に提示する。教員は提示した語にマークを行う(表7, マークされた語は網掛で示す)。

表 7:重要語のマーク

順位	1	2	3	4	5	6	7	8	9	10
単語	FAT	クラスタ	セクタ	MB	権	FD	領域	ディスク	名	ファイル
順位	11	12	13	14	15	16	17	18	19	20
単語	Ccc	TB	xls	パーティション	ルート	予備	ディレクトリ	フォーマット	bit	エラー

マークされた語は9つあるので、9番目以降の「ファイル」、「パーティション」「ディレクトリ」「フォーマット」の語に対して何語増やせばGスコアで9番以内に入るかの再計算を行う(表8)。

表 8: 改善例の提示

順位	1	2	3	4	5	6	7	8	9	10
単語	FAT	クラスタ	セクタ	MB	権	FD	領域	ディスク	名	ファイル
改善例										1語増やすと4位になる
順位	11	12	13	14	15	16	17	18	19	20
単語	Ccc	TB	xls	パーティション	ルート	予備	ディレクトリ	フォーマット	bit	エラー
改善例				1語増やすと6位になる			2語増やすと4位になる	3語増やすと4位になる		

改善例では、スライド中にその単語をあと何語増やせばよいかと、その語数だけ増やした場合の順位を提示する。教員は改善例に従いそれぞれの語を増やすようにスライド中の文章の推敲を行う。その後、順位の再計算を行う(表9)。

表 9: 再計算結果の表示

順位	1	2	3	4	5	6	7	8	9	10
単語	FAT	クラスタ	セクタ	フォーマット	ディレクトリ	MB	権	FD	パーティション	領域
改善例										1語増やすと3位になる
順位	11	12	13	14	15	16	17	18	19	20
単語	ファイル	ディスク	名	Ccc	TB	Xls	ルート	予備	Bit	エラー
改善例	1語増やすと7位になる	1語増やすと7位になる								

再計算の結果、「領域」「ファイル」「ディスク」が 9 位以内に入っていない。そのため、教員は再度スライドの修正を行う。これを満足できるまで繰り返すことによってスライドを改善できる。

5. 問題点と改善方法

今回の手法では、複雑な文章解析を行わずに重要語を提示することによって教材の改良支援を行っている。しかしながらいくつかの問題点がある。

1. 形態素解析の精度
   スライドから文を抽出し、そのまま形態素解析エンジン(MeCab)により、単語に分解し、名詞のみを抽出している。スライド中に現れる文は、一般の文章に表れる文とは異なり、完全な文になっていない場合や、改行や記号により文が分断されている場合がある。そのためいくつかの語に関しては、間違った解析結果を出力している。例えば「二分木」という言葉が「二」と「分」と「木」に分断されてしまう場合があった。形態素解析時にこれらを考慮して、文の解析の調整を行う必要がある。

2. 語の重み付け
   実際のスライドでは、重要語は本文領域ではなくてタイトル領域に出現したり、視覚的効果(例:フォントサイズ、色、アンダーライン)が施されていたりする。スライドを見る側はこれら視覚的効果が施された語を重要語と考える。これらの情報を語の重みとしてその語の出現頻度に加算する必要がある。

3. 修正後の重要語の順位
   重要語に関して、改善例に示された通りに重要語を増やしても、改善例の順位にはならない場合がある。これは、スライドを改善する場合には、その重要語だけではなく、新しい文やスライドを追加するため、スライド全体の語数が変わり、全てのスライドにある全ての語の G スコアが変化する。このため、目標とした G スコアにならなかったり、目標とした G スコアになっても他の語の G スコアが上がってしまい、順位としては下がってしまう場合がある。また、別の授業スライドでの語の順位も変化する。実際の修正作業では、重要語の順位を上げるために何度も修正しなければならい。

また、G スコア順に並べた語を本当に学生が重要語と認識するのかの調査はまだ行っていない。2 の語の重み付けと関連して、実際の授業での調査が必要である。

6. まとめ

今回、スライドを解析することによって重要語の候補を提示し、その情報を用いてスライド改善に役立てる手法を提案した。今後、支援ツールとして使えるようにソフトウェア開発を行い、実用に耐えられるかどうかの検証を行う予定である。また、重要語の抽出だけでなく、他のスライド改善手法も考えていく。

参考文献

- 工藤拓, "MeCab: Yet Another Part-of-Speech and Morphological Analyzer," http://mecab.sourceforge.net/
- Robert R. Sokal, F. James Rohlf, "Introduction to Biostatistics," W. H. Freeman and Company, 1973.