Vol.2016-IS-136 No.6
2016/6/11

講義内容の要約字幕作成支援システム ──意思決定手法とバスケット分析に基づく支援方法の提案──

古宮 誠一[†1] 上之薗和宏[†2] 八重樫理人[†3]

概要 講師の発話情報を要約した文章を講義の映像に字幕として付与することは，日本語初心者が講義内容を理解するのに効果的であると思われる。しかし，要約字幕の作成には多くの労力が必要であり，講義を担当する講師以外の人間が要約を作成すると，講師の意図とは異なる要約が作成されてしまう可能性がある。著者らは，意思決定手法とアソシエーションルールを利用して，講師の発話テキストから重要文を抽出することにより，講師の意図を反映した要約字幕の作成を支援する方法を提案している。

キーワード: 講義内容の要約，要約字幕，発話情報，重要文抽出，意思決定手法，アソシエーションルール

A System to Help with Making Subtitles Condensed the Content of a Lecture: Proposing a Method to Help with Making them by Using Decision-Making Technique and Basket Analysis

Seiichi KOMIYA[†1] Kazuhiro UENOSONO[†2] and Rihito YAEGASHI[†3]

Abstract It is one of effective means for a beginner of Japanese language to understand the content of a lecture conducted in Japanese to attach abridged sentences of an instructor's utterance information as subtitles to the video of a lecture conducted in Japanese. However, the means have the following two problems: One problem is to take a lot of work to make subtitles condensed the content of a lecture. Another problem is to threaten to create abridged sentences to disagree with what the instructor intended, if anybody but the instructor make abridged sentences. The authors propose a method to help with making subtitles based on the instructor's intention, by eliciting key sentences from the utterance texts of the instructor with use of decision-making technique and association rules.

Keyword Summary of Lecture Content, Abridged Subtitles, Utterance Information, Key Sentence Elicitation, Decision-Making Technique, Association Rules

1. はじめに

マレーシア人学生が日本の工学系大学へ留学するための予備プログラムとして，JAD プログラム（Japan Associate Degree Program）[6][7]と呼ばれる制度がある。このプログラムでは，現地(マレーシア)で 1 年目は日本語の習得を目的とした教育が行われ，2 年目以降は工学系のほとんどの授業が日本語で講義される。現地に在住する教員だけでは対応できない科目の講義は，日本で収録された講義の映像を講義コンテンツの形で配信することによって行われる。講義コンテンツはストリーミングサーバに保管され，学生はこれを繰り返し閲覧することができる。

しかし，学生は日本語を学び始めて 1 年しか経っていないので，講義コンテンツを見るだけでは内容を完全に理解することは難しい。そのため，講義コンテンツに要約字幕(講師の発話を要約した字幕)を付与することで学生の理解を支援する試みがなされている。高田ら[13]は留学生向けの映像コンテンツに対する字幕の有用性を検証し,『日本語を非母国語とする学生に，日本語による講義を理解させるのに，日本語による講義の発話を要約した字幕が有効である』と述べている。従って，講義内容の理解を容易にするために，講師の発話テキストから要約字幕を作成することが本研究の目的である。

ところで，講義コンテンツから要約文を作成する作業は，多大な労力がかかるので，コンピュータ処理によって自動的に要約文を作成できるようにしたい。

要約文の作成方法には以下の 2 種類がある。

†1 国立情報学研究所 先端ソフトウェア工学・国際研究センター
National Institute of Informatics, Tokyo, 101-8430 Japan
†2 青山学院大学 Aoyama Gakuin University, Kanagawa, 252-5258 Japan
†3 香川大学 Kagawa University, Kagawa, 252-5258 Japan

(A)　文意の抽象化による方法

これは｛赤，青，黄，……｝という情報から，これらが意味しているのは『色』であると要約する手法である。この方法は文章の圧縮率を高める上では有効であるが，文章を要約する過程で文の意味を抽象化しているので，コンピュータ処理には不向きである。このため，この方法による自動要約の実現は困難である。

(B)　重要文の抽出による方法

これは，文の集合から，重要だと思われる文だけを抽出することにより，要約文を作成する方法である。この方法は，必要な文を抽出する方法なので，コンピュータ処理は可能だと思われる。

要約を実現する上記 2 つの方法を比較すると，(A)による実現は不可能であるが，(B)による実現は可能だと思われるので，本研究では(B)の『重要文の抽出による方法』を採用し，講師の発話を要約する過程をコンピュータで自動化する方法を考える。

この方法を採用して講師の発話テキストを要約する過程を自動化する際に，解決しなければならない課題として下記の 3 つがある。

(B1)　講義を担当した講師以外の者が要約文を作成すると，講師の意図とは異なる要約文が出来上がる可能性があること。

(B2)　文中に冗長な語や字句が残ってしまう可能性があること。

(B3)　文と文とのつながりが不自然になってしまう可能性があること。

本稿では，上記の問題点(B1)を解決することだけに絞って議論を進める。

上記の問題点(B1)を解決するために我々が採用した方法は，講師の発話テキストを要約するために，講義の内容を表しており，重要だと思われるキーワードを，講義を担当した講師に選んで貰うとともに，各キーワードの重要度を与えて貰うことにより，重要度の情報を基に重要文を自動抽出するというアプローチを採用する。このとき，講師の負担を少しでも軽くするために，講義の発話テキストの中からキーワードの候補(キーワードとなり得る語句)をコンピュータ処理により自動抽出して，それらの中から重要と思われるキーワードの候補を講師に選んで貰うとともに，各キーワード候補の重要度も与えて貰うという方法を採用する。

システムが自動生成する要約字幕の編集方針として，次の 2 つのモードが考えられる。

(a) 重要度の下限となるフラグ名を指定する方式

(b) 要約字幕の編集に必要な最大文字数を指定する方式

本稿では，編集方針(a)の場合における実現方法に絞って議論する。

2. 提案する要約の方法とその手順

講師の発話テキストの自動要約は，重要文の抽出による方法を採用し，その処理過程を**図 1 に示す 8** つに分解するとともに，そのそれぞれの過程を自動化することによって自動要約の実現を図る。

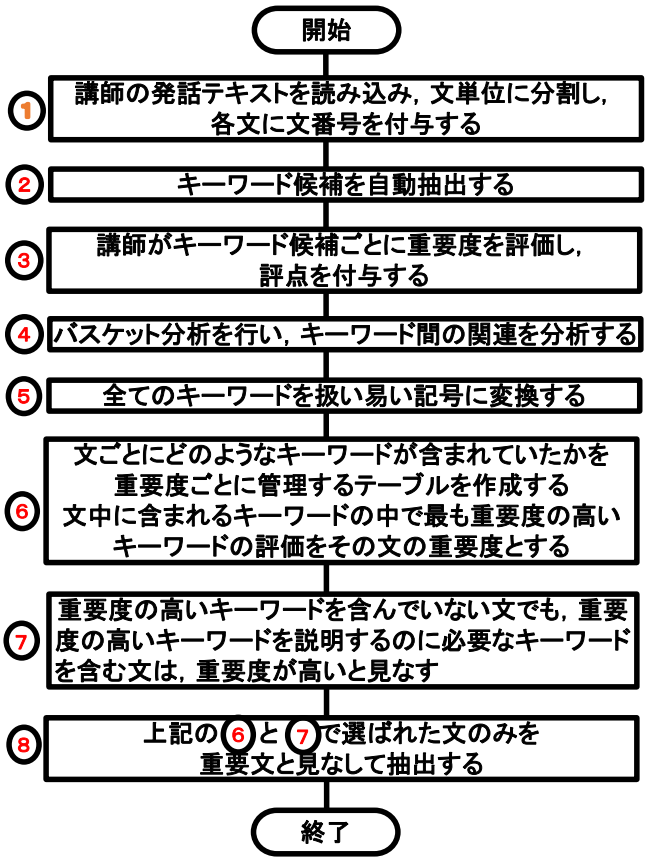


図 1　処理の流れ
Figure 1　Processing flow.

上記の⑥と⑦で選ばれた文のみを重要文と見做して抽出する。上記の 8 つの過程のそれぞれを,次節以降に節を分けて具体的に説明する。

2.1 講師の発話テキストの読み込み

要約の対象となる文章は，業者から購入した専用の『音声情報文字化ソフト』を用いて，講師の発話(音声)情報をテキスト化することによって得られる，テキスト化された講師の発話(音声)情報である。講師の発話(音声)情報の例を図 2 に示す。図 2 の情報を**文(1 sentence)**ごとに分解して，図 3 のように文番号を付与したものを用意し，これを要約の対象とする。

1:19:25　さて、これでおおよそ時間になりました。今日の授業はここまでにいたします。
1:19:33　来週少し残ったディジタル変調方式についてお話したいと思っています。
1:19:42　実は日本は11月の21日、私は明日マレーシアに発ちます。
1:19:50　明後日みなさんとお会いできる予定になっています。
1:19:55　えー、是非マレーシアで元気にお会いしたいと思っています。

図 2　テキスト化された講師の発話(音声)情報の例
Figure 2　An example of incoming information.

図 2 のようなテキスト化された講師の発話(音声)情報を，その順序を変えることなく，文(1 sentence)単位に区切って取り出し，それに文番号を付与して図 3 のように並べたものが発話テキストである。この発話テキストから，その順序を変えることなく，重要と思われる文のみを取り出し，新たにこれにテキストを加えたり削ったりすることなく，文と文とを物理的に繋げたものが，本稿で目標とする要約字幕である。

文番号	発話テキストを文単位に区切って取り出し，順序を変えずに並べたもの
文番号1	最初の文
文番号2	2番目の文
文番号3	3番目の文
·····	·····
文番号n	n番目の文(最後の文)

図 3　文番号を付与した発話テキストの例
Figure 3　An example of incoming information.

2.2 キーワードの候補となる語句の自動抽出

要約対象となる講師の発話テキストにタイムコードを付加した情報を入力し，chasen [1][2]を用いてこれを形態素解析して得られた結果(図 4 と表 1 の例を見よ)からキーワード候補を自動抽出する。

```
要約字幕の作成支援を行う。
要約    ヨウヤク        要約    名詞-サ変接続
字幕    ジマク  字幕    名詞-一般
の      ノ      の      助詞-連体化
作成    サクセイ        作成    名詞-サ変接続
支援    シエン  支援    名詞-サ変接続
を      ヲ      を      助詞-格助詞-一般
行う    オコナウ        行う    動詞-自立       五段・ワ行促音便        基本形
。      。      。      記号-句点
EOS
```

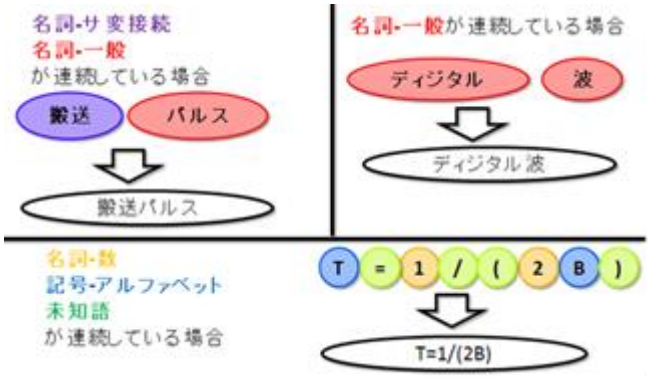
図 4　Chasen による形態素解析の結果の例
Figure 4 An example of analysis result with the use of the Chasen.

形態素解析の結果からキーワード候補を自動抽出するには，表 1 に示す(1)～(13)のようなパターンを持った語句を抽出すればよいことを実験によって明らかにした。より具体的に言えば，1 つまたは 2 つ以上連続している語句の品詞が，表 1 に示すパターンを持った語句のいずれかの組み合わせであれば，キーワードの候補となることが判明した。

表 1 に示す 13 種類のタパーン同士の組み合わせによるキーワード候補選出の具体例は次のとおりである。例えば，『搬送パルス』という語句は，『名詞_サ変接続』という品詞の語句『搬送』と，『名詞_一般』という品詞の語句『パルス』とが連続しているのでキーワード候補となる。また，『ディジタル波』という語句は，『名詞_一般』という品詞の語句『ディジタル』と『名詞_一般』という品詞の語句『波』とが連続しているのでキーワード候補となる。この様子を図 5 に示す。

表 1　キーワード候補となる語句とその品詞の例
Table 1　Examples of words and parts of speech which are a keyword candidate

項番	品詞	具体例
(1)	名詞_一般	パルス　高周波
(2)	名詞_固有名詞_一般	富士山
(3)	名詞_固有名詞_地域_一般	東京
(4)	名詞_サ変接続	サンプリング　搬送
(5)	名詞_数	0 1 2 3 4 5 6 7 8 9
(6)	記号_アルファベット	A B C D E F a b c d e f
(7)	記号_一般	＋　－　×　÷　＝
(8)	記号_括弧開	（
(9)	記号_括弧閉	）
(10)	未知語	+ - * / ( ) =
(11)	名詞_接尾-一般	値　系
(12)	名詞_接尾_サ変接続	化
(13)	名詞_接尾_助数詞	個　ビット



ワード候補の例
Figure 5　An example of keyword candidates.

2.3 抽出されたキーワード候補の評価と分類

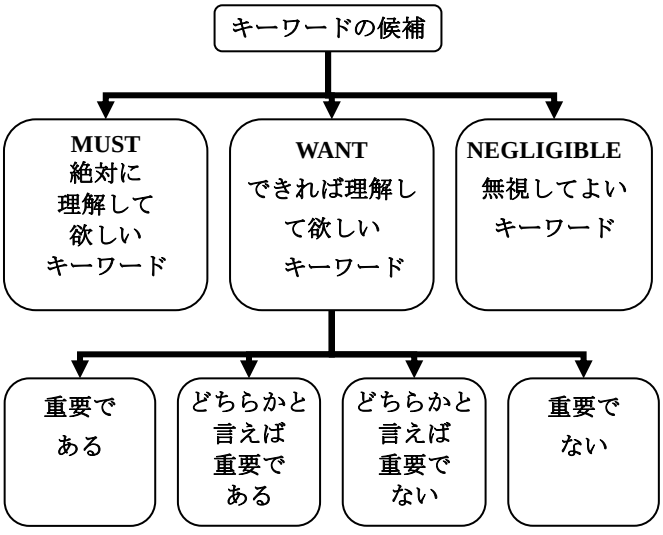


図 6　各キーワードの評価とそれに基づく分類
Figure 6　Evaluation of each keyword and evaluation-based classification of keywords.

システムが自動抽出したキーワード候補を，講師が『絶対に理解して欲しいキーワード』『できれば理解して欲しいキーワード』『無視して良いキーワード』の 3 種類に分類する。そして『できれば理解して欲しいキーワード』に対しては，さらに『重要である』『どちらかと言えば重要である』『どちらかと言えば重要でない』『重要でない』の 4 種類に分類する。この分類方法を**図 6** に示す。

2.4 記号名称への各キーワードの置き換え

各キーワードを，講義の中で使用されている表現を変えずに，そのまま使用するのでは管理し難い。このため，各キーワードを次のような置き換え規則で記号名称に置き換える(置き換え前の表現と置き換え後の表現との対応表を作っておく)。

(1)『絶対に理解して欲しい』キーワードの場合
それが『絶対に理解して欲しい』に分類されたキーワードであることを示す M で始まる記号名称を使用することとし，講義の中に出てきた順に追い番で M1, M2, M3, ・・・という名称を付与する。

(2)『重要である』に分類されたキーワードの場合
それが『重要である』に分類されたキーワードであることを示す A で始まる記号名称を使用することとし，講義の中に出てきた順に追番で A1, A2, A3, ・・・という名称を付与する。

(3)『どちらかと言えば重要である』に分類されたキーワードの場合
それが『どちらかと言えば重要である』に分類されたキーワードであることを示す B で始まる記号名称を使用することとし，講義の中に出てきた順に追番で B1, B2, M3, ・・・という名称を付与する。

(4)『どちらかと言えば重要でない』に分類されたキーワードの場合
それが『どちらかと言えば重要である』に分類されたキーワードであることを示す C で始まる記号名称を使用することとし，講義の中に出てきた順に追番で C1, C2, C3, ・・・という名称を付与する。

(5)『重要でない』に分類されたキーワードの場合
それが『重要でない』に分類されたキーワードであることを示す D で始まる記号名称を使用することとし，講義の中に出てきた順に追番で D1, D2, D3, ・・・・・という名称を付与する。

(6)『無視してよい』に分類されたキーワードの場合
『無視してよい』に分類されたキーワードは，キーワードとは認めないので無視する。

2.5 キーワードの重要度に基づく各文の評価方法

文ごとの重要度を評価するには，文ごとに，どのようなキーワードが含まれているかをチェックする必要がある。このため，文ごとに，そこにどのような重要度のどのようなキーワードを含まれているかを管理するために，表 5 のような『**文管理テーブル**』(後述する)を作成する。(テキスト情報と文管理テーブルとは，文番号で対応が付くようになっている。)

各文の重要度評価は，その文に含まれる最も重要度の高いキーワードの重要度をもってその文の重要度と見なす。

『絶対に理解して欲しい』に分類されたキーワードを含む文は，この文が無条件に重要文であることを示す『フラグ M』を，この文に対応する『**文管理テーブル**』に，システムが自動的に付与する。その文に含まれる最も重要度の高いキーワードが,『重要である』に分類されたキーワードであれば， そのことを示す『フラグ A』を,『どちらかと言えば重要である』に分類されたキーワードであれば，そのことを示す『フラグ B』を,『どちらかと言えば重要でない』に分類されたキーワードであれば, そのことを示す『フラグ C』を,『重要でない』に分類されたキーワードであれば，そのことを示す『フラグ D』を,『無視してよいキーワード』に分類されたキーワードであれば，そのことを示す『フラグ N』を，その文に対応する『**文管理テーブル**』にシステムがそれぞれ自動的に付与する。どのランクのかということと評点との対応関係を**表 2** に示す。

表 2　文中に含まれるキーワードの重要度に基づく文の重要度評価
Table 2　The important degree of a sentence based on important degree of the keywords contained in the sentence.

その文に含まれる最も重要度の高いキーワード		フラグ
(学生に)絶対に理解して欲しいキーワード		M
(学生に)できれば理解して欲しいキーワード	重要である	A
	どちらかと言えば重要	B
	どちらかと言えば重要でない	C
	重要でない	D
無視してよいキーワード		N

表 3　各文を格付けし評点する方法の具体例
Table 3　An example of a method for ranking and scoring each sentence.

文番号	発話テキストの例	文の評点(得点)
1	A1○M1○B1 N1○D1○○○○	M
2	○A1○C1○D2○ N2○○A2○○○	A
3	B1○N3○B2○B3 N4○D3○C1○○	B
4	○○○C1○○C2 ○○C3○D4○○	C
5	○D5○○N3○○ N4○○D6○○○	D

(注) ○印は無視してよいキーワードの語句を表す。

表 2 のようなキーワードの分類方法と文の評価方法を採用したときに，キーワードの重要度を基にどのように文を格付けして評価するのかを，**表 3** に発話テキストの例とそれに対応する文の評価例を示す。

3. バスケット分析

重要度が高い 1 つのキーワードの内容を説明するのに，そのキーワードを含む文が 1 つだけで済むケースは多くない。寧ろ，複数の文を要するケースのほうがずっと多いのではないかと思われる。このような場合に，キーワードの重要性の観点だけから重要文を抽出すると，そのキーワードを説明するのには必要な文なのに，重要度の高いキーワードを含んでいないために，選出されない文が出てくる可能性がある。このような問題点を解決するために，バスケット分析(basket analysis)[14]を用いる。バスケット分析とは，購入する商品の傾向をバスケット(買い物籠)単位で分析することによって，或る商品を消費者が購入した場合に別の或る商品を一緒に購入する傾向がどれだけあるのかを分析する方法である。講師が重要と認めたキーワードを S とし，S が出現するときに，キーワード T がどれだけ出現するかを，バスケット分析を使って次のように分析する。

バスケット分析では，1 つの発話テキスト(買い物籠に相当する)に含まれる，2 つのキーワード S と T の関係の深さをそれらの出現頻度に基づいて調べる。

表 4　　キーワード S と T の出現度数
Table 4 Occurrence rate of keywords S and T.

		T が		合計
		有り(出現)	無し(出現せず)	
S が	有り	n1 個	n2 個	n1+n2 個
	無し	n3 個	n4 個	n3+n4 個
合計		n1+n3 個	n2+n4 個	n1+n2+n3+n4 個

キーワード S と T の出現度数が表 4 の通りだったするときに，表 4 の数値を使ってどのようにアソシエーションルールを求めるかを以下に示す。

(1) 前提確率(Antecedent)
S が出現する確率のことで，p(S)と表記される。
p(S)は次式で求められる。
p(S) = **(n1+n2) / (n1+n2+n3+n4)**

(2) 支持度(**Support**)
S と T が同時に出現する確率のことで，p(**S∩T**)または p(S, T)と表記される。p(**S∩T**)は次式で求められる。
p(S∩T) = **n1 / (n1+n2+n3+n4)**

(3) 信頼度(**Confidence**)
S が出現する集合の中で T も出現する（条件付き）確率のことで，p(T|S)と表記される。p(T|S)は次式で求められる
p(T|S) = p(**S∩T**)/P(S)
= **{n1/(n1+n2+n3+n4)}/{(n1+n2)/(n1+n2+n3+n4)}**
= **n1 / (n1+n2)**
p(T|S)の値が大きければ，S が出現すると高い確率で T も出現することになり，アソシエーションルールとして採用される可能性が高くなる。

(4) 期待信頼度(**Expected Confidence**)
T が出現する確率のことで，p(T)と表記される。
p(T)は次式で求められる。
p(T) = **(n1+n3) / (n1+n2+n3+n4)**

(5) リフト値(**Lift**)
p(T) の値が大きければ，自ら p(T|S) の値も大きくなるので，p(T)の値とは無関係に p(T|S)の値が大きいものだけをアソシエーションルールとして採用すべきである。このため，p(T)の値による影響を取り除いた『**リフト値**』と呼ばれる数値が利用される。リフト値は次式で求められる。
p(T|S)/p(T)
= **{n1/(n1+n2)} / { (n1+n3)/(n1+n2+n3+n4)}**
リフト値の(少なくとも 1 よりも)大きいものがアソシエーションルールとして採用される。

上記のバスケット分析によって，『重要度の高いキーワード S が出現すると，キーワード T も出現する確率が高い』というアソシエーションルールが抽出されたとする。T も講師が重要だと認めたキーワードであれば，アソシエーションルールを用いなくても，T を含む文は，これまでの方法で重要文として抽出される。このため，アソシエーションルールの適用によって，T を含む文を重要文と見なして抽出するのは，キーワードの重要度の分析の際に，T を含む文が重要だと見なされなかった場合である。

アソシエーションルールを用いることによって初めて重要文だと見なされるのは，重要度の高いキーワード S が出現したときの，キーワード T を含む文だけである。従って，キーワード T を含む文が重要文と見なされる可能性を示す情報は、キーワード T を含む文に対応する文番号の『**文管理テーブル**』に，S の重要度がどのレベルであるかという情報とともに，この文が S との関係で重要文と見なされる可能性があるということが示されなければならない。つまり，キーワード S の重要度が M ならば，そのことを示すフラグ**m**と，S と T の関係を示すm（S, T）という情報が必要である。同様にして，S の重要度が A ならば a(S, T), B ならば b(S, T), C ならば c(S, T), D ならば d(S, T）という情報が必要である。

4. 文管理テーブル

文管理テーブルの形式を表 5 に示す。文管理テーブルは，その文にどのようなキーワードが含まれているかを，キーワードの種類別に整理して表現しているテーブルである。

表 5　文管理テーブルの形式
Table 5 A format of a table for showing information on keywords containing in each statement.

文番号	文の重要度	その文に含まれている重要度別のキーワード					その文に含まれている説明の為のキーワード				
		M	A	B	C	D	m	a	b	c	d
文番号1											
文番号2											
文番号3											
……	……	……	……	……	……	……	……	……	……	……	……	……
文番号n											

表 5 の文番号は，図 3 の『文番号を付与した発話テキスト』の文番号と対応付けがなされている。『その文に含まれている重要度別のキーワード』の欄には，重要度 M, A, B, C, D ごとに，どのようなキーワードが含まれているかが示されている。『その文に含まれている説明の為のキーワード』の欄には，重要度の高いキーワードを説明するためのキーワードとして，どのようなキーワードがその文に含まれているかが示されている。

4.1 その文に含まれている重要度別のキーワードの欄について

M，A，B，C，D の欄はそれぞれ，その重要度に分類されるキーワードとして，どのようなキーワードがその文に含まれているかを具体的に示している欄である。

重要度 M の欄の場合で，その具体例を以下に示す。

(1) M の欄は，現時点では 4Byte の長さを考えているが，先頭から 4Byte 目だけに絞って，その表記とその意味を示す。2 進数表現で

① 2 の 0 剰ビットが 1 のとき，つまり 00000001 のとき，M1 というキーワードがその文に含まれていることを意味する。

② 2 の 1 剰ビットが 1 のとき，つまり 00000010 のとき，M2 というキーワードがその文に含まれていることを意味する。

・・・・・・・・・・・・・・・・・・・・

2 の 7 剰ビットが 1 のとき，つまり 10000000 のとき，M8 というキーワードがその文に含まれていることを意味する。

(2) 先頭から 3Byte 目も先頭から 4Byte 目と同様の方法で，M9～M16 というキーワードがそれぞれその文に含まれていることを示す。

(3) 先頭から 2Byte 目も先頭から 4Byte 目と同様の方法で，M17～M24 というキーワードがそれぞれその文に含まれていることを示す。

(4) 先頭から 1Byte 目も先頭から 4Byte 目と同様の方法で，M25～M32 というキーワードがそれぞれその文に含まれていることを示す。

重要度 A, B, C, D の欄(現時点では，それぞれ 4Byte の長さを考えている)については，M を A, B, C, D にそれぞれ置き換えて考えれば良い。

4.2 その文に含まれている説明の為のキーワードの欄について

m，a，b，c，d の欄はそれぞれ，重要度 M，A，B，C，D のキーワードを説明するキーワードとして，どのようなキーワードがその文に含まれているかを具体的に示している欄である。それ故，**m，a，b，c，d** の欄については，**M を m に，A を a に，B を b に，C を c に，D を d に**それぞれ置き換えて考えれば良い。

5. 文ごとの評価に基づく要約文の編集方法

システムが自動生成する要約字幕の編集方針として，次の 2 つのモードが考えられる。

(a) 重要度の下限となるフラグ名を指定する方式
(b) 要約字幕の編集に必要な最大文字数を指定する方式

本稿では，編集方針(a)を採用している。

編集方針(a)は，フラグ M (最重要の文)から最低何処までの範囲の文を重要文として残すかを，文ごとに付与されたフラグ名を使って指定する方式である。これには次の 5 種類が用意されている。

M： M または m のフラグが付与されている(優先順序が 1 の)文のみを重要文として残す方式である。

A： M または m のフラグと A または a のフラグが付与されている(優先順序が 1 と 2 の)文のみを重要文として残す方式である。

B： M または m のフラグ，A または a のフラグ，B または b のフラグが付与されている(優先順序が 1 ～ 3 の)文のみを重要文として残す方式である。

C： M または m のフラグ，A または a のフラグ，B または b のフラグ，C または c のフラグが付与されている(優先順序が 1 ～ 4 の)文のみを重要文として残す方式である。

D： M または m のフラグ，A または a のフラグ，B または b のフラグ，C または c のフラグ，D また d のフラグが付与されている(優先順序が 1 ～ 5 の)文のみを重要文として残す方式である。

編集方針(a)の場合における処理の詳細を図 7 に示す。

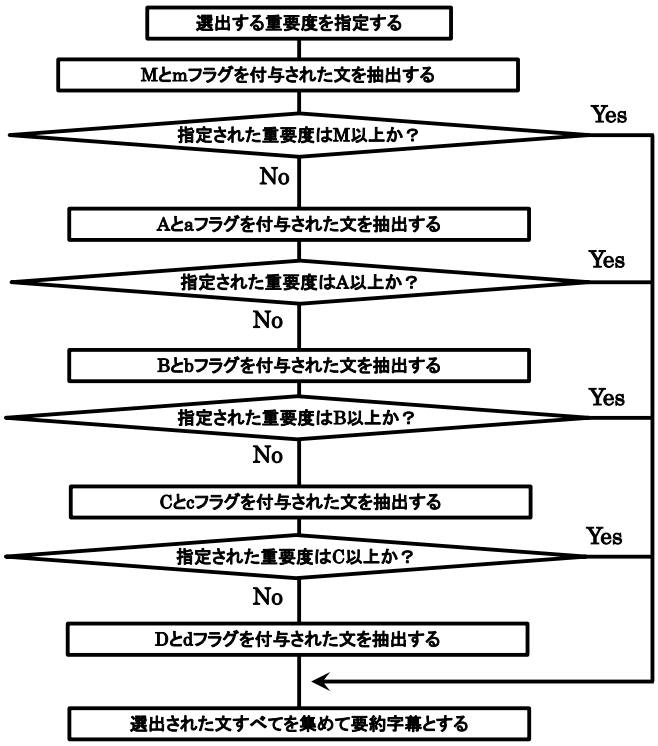


図 7　重要度の下限となるフラグ名を指定する方式の処理の流れ
Figure 7　Processing flow of the system to specify a flag name represented lower limit of important degree.

重要文の編集に際しては，講師の発話情報(テキスト)の順序を変えることなく，各文の取捨選択を行う。

このような準備をした上で，発話された順序を変えずに，入力バッファから，順次必要な文を取り出して編集用バッファに埋めて行くことにより，(冗長な部分を含んだままの)要約字幕の作成処理を完成させる。

6. 関連研究

解説[8]には，「要約は，原文の大意を保持したまま，テキストの長さ，複雑さを減らす処理だとも言える」と書かれているので，本稿で扱っている処理は，明らかに『テキスト要約』である。また，解説[8]には，要約処理の過程は，(1)テキストの解釈(文の解析とテキスト解析結果の生成)，(2)テキスト解析結果の，要約の内部表現への変形(解析結果の中の重要部分の抽出)，(3)要約の内部表現の要約文としての生成，の大きく３つのステップに分けられるとある。しかし，本稿で提案している処理は，これには全く適合しない。一方，解説[11]には，「情報抽出処理では，『テキスト解析(自然言語処理における構文解析や意味解析など)』の難しい処理は行わずに，抽出対象の特徴を指定する情報を与え，それとのパターンマッチングによる情報抽出が基本である」という意味のことが書かれている。本稿で提案している技術は，重要文を特定するために，キーワードの重要度を指定することにより，これを含む文を重要文と見なして抽出する方法が，パターンマッチングによる情報抽出に該当する。また，バスケット分析を用いて，重要なキーワードを説明するために使用されているキーワードを特定し，重要なキーワードを含まない文でも，これを含む文は重要であると見なして抽出する方法も，パターンマッチングによる情報抽出に該当する。従って，本稿で提案している技術は，情報抽出処理技術の利用によるテキスト自動要約である。しかし，解説[8][9]には，本稿で提案している手法と類似の手法に関する記述はない(新規性がある)。

3 章において，その文に講師が重要だと評価したキーワードが１つも含まれていないために，重要だと評価されなかった文でも，講師が重要だと評価したキーワードを説明するために必要なキーワードを含んでいれば，その文も重要文であると主張した。そして，講師が重要だと評価したキーワードを説明するのに必要なキーワードを検出するために，バスケット分析のアソシエーションルールを使用する方法を提案した。つまり，講師が重要だと評価したキーワードを S とし，重要だと評価されなかったキーワードを T とするとき，信頼度 p(T|S)の値が大きければ，S が出現すると高い確率で T も出現することになるので，T は S を説明するためのキーワードである可能性が高いと評価した。このとき，T が S を説明するためのキーワードであるか否かを判断するための条件として， p(T|S)/p(T)の値(リフト値)が 1 よりも大きいことを挙げた。

上記の目的で，アソシエーションルールの代わりに相互情報量[12]を用いるのは正しくない。何故なら，相互情報量を求めることは，p(T|S)と p(T|S)/p(T)が条件を満たすことを要求するだけでなく，同時に p(S|T)と p(S|T)/p(S)も条件を満たすことを求めていることになるからである。

我々がこれまでに行った先行研究[10][3][4][5]では，いずれも**図 6** のような 6 段階で講師がキーワードの重要度を与え，その文に含まれるキーワードの中で，最も重要度の高いキーワードをもって，その文の重要度をランク付けするところまでは，本稿と全く同じである。しかし，文ごとの重要度の評価方法が本稿とは異なっている。これらの論文では，『できれば学生に理解して欲しい』キーワードに対しては，『とても重要』『どちらかと言えば重要』『どちらかと言えば重要でない』『あまり重要でない』に分類されたそれぞれのキーワードごとに，その重みとして 4 点，3 点，2 点，1 点をそれぞれ付与し，その文中に含まれるこれらのキーワードの出現個数をも評価していた。つまり，各キーワードの重みとその出現個数との積和計算によって，各文の重要度を評価していた。そして，『学生に絶対に理解して欲しいキーワード』を含む文を優先して重要文と見なすとともに，『できれば学生に理解して欲しい』キーワードのみを含む文に対しては，積和計算の値の大きいものほど重要な文であると評価していた。何故なら，これらの論文では，システムが自動生成する要約字幕の編集方針として，(b)

要約字幕の編集に必要な最大文字数を指定する方式での実現を指向していたからである。つまり，要約字幕の文字数に制限があったので，どちらの文がより重要度が高いかが，そこでは重要なテーマであったからである。(従って，この時点では，そのキーワード自身の重要度は低いが，重要度が高いキーワードの説明には欠かせないキーワードを含む文も重要だとする考えは無かった。)

しかし，『学生に絶対に理解して欲しい』キーワードを含む文だけでも指定された要約字幕の最大文字数を超えてしまう場合にはどうするかが問題となり，『学生に絶対に理解して欲しい』キーワードを含む文に対しても，『できれば学生に理解して欲しい』キーワードを含む文と同様の積和計算を採用して，『学生に絶対に理解して欲しい』キーワードを含む文同士での優先順序を求める方法を[3][4][5]で提案した。このとき，文のランク付けよりも積和計算の計算結果を優先すると，計算結果の上では文の重要度が逆転する場合があっても，文のランク付けを優先することにした。

しかし，要約字幕の最大文字数を指定する方式では，最大文字数の制限を満足する中で，上記の方法で，より重要度の高い文を選ぶことができたとしても，日本語を母国語としない人達にとって，理解に有効な要約字幕ができるかどうかが問題となった。この問題に対しては，そもそも講義内容の要約字幕に必要な文字数は，講義内容ごとに異なる筈である。しかし，講義内容ごとに必要な文字数は，具体的にそれぞれ幾つが適切なのかが判らない。であるのに，要約字幕の最大文字数を指定する編集方針は良くない。このように考え，システムが自動生成する要約字幕の編集方針として，(a)重要度の下限となるフラグ名を指定する方式を採用することにした。それが本稿で採用した編集方式である。

7. おわりに

マレーシア人学生の理解を支援するために，講義内容の要約字幕を映像コンテンツに付与する試みがなされている。しかし，作成に労力がかかり過ぎるという問題点と講師の意図が要約字幕に反映されていないという問題点があった。我々は，講義内容を表すキーワードを講師に選出して貰うとともに，重要度の視点からキーワードの重要度を 6 種類に分類して貰い，文中に含まれる最も重要度の高いキーワードを基に文の重要度を決定する方法を提案した。また，重要度の低い文でも，重要度の高いキーワードを説明するキーワードを含む文も重要文だと見なして抽出する方法を提案した。これにより，これらの問題点を解決できるという見通しを得た。

謝辞　本研究で用いた講義コンテンツ及び発話テキストは，芝浦工業大学の三好匠准教授(2008 年当時)に提供して戴いた[10]。ここに記して感謝申し上げます。

参考文献

[1][ChaSen <2009 年 1 月現在>, http://chasen-legacy.sourceforge.jp/
[2] 松本裕治，形態素解析システム『茶筌』，情報処理 Vol.41, No.11, pp.1208-1214 (November 2000).
[3] 古宮誠一，工藤永貴，上之園和宏，八重樫理人，“講義内容の要約字幕作成支援システム―意思決定手法に基づく支援方法の提案,”信学技報, Vol. 112, No. 496, KBSE 2012-86, pp.103-108 (March 14-15, 20139.
[4] 工藤永貴，千葉亮太，八重樫理人，上之園和宏，古宮誠一，“講義内容の要約字幕作成支援システム―重要文自動抽出手法の提案―,” 研究報告 コンピュータと教育(CE), 2012-CE-114(15), pp,1-8 (March 9, 2012).
[5] 工藤永貴，千葉亮太，八重樫理人，上之園和宏，古宮誠一，“講義内容の要約字幕作成支援システム―重要文自動抽出手法の提案(その 2)―，第 9 回教育学習支援情報システム研究発表会，情報処理学会 (Feb. 1-2, 2013).
[6] マレーシア高等教育基金事業 <2009 年 1 月現在> https://office.shibaura-it.ac.jp/kokusai/06malaysia.html
[7] 日本国際教育大学連合『JAD プログラム』<2009 年 1 月現在> https://office.shibaura-it.ac.jp/kokusai/jucte/program/b,ackgraound.html
[8] 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).
[9] 奥村学，難波英嗣,"テキスト自動要約に関する最近の話題, "自然言語処理, Vol.9, No.4, pp.97-116 (2012).
[10] 大澤勇基，上之園和宏，八重樫理人，三崎貴裕，榎津秀次，古宮誠一，“ 要約字幕作成支援システム―重要文自動抽出手法の検討―，情報システム学会，第 4 回全国大会・研究発表大会, A1-4 (Dec. 12-13, 2008).
[11] 関根聡, “テキストからの情報抽出―文書から特定の情報を抜き出す―, “情報処理, Vol.40， No.4, pp.370-373 (1999).
[12] 相互情報量 https://Ja.m.wikipedia.org/wiki/相互情報量
[13] 高田充，三好匠，八重樫理人，國弘保明，尾沼玄也: e-Learning における日本語理解度と授業集中度を考慮した字幕作成手法，2008 年電子情報通信学会総合大会，分冊情報システム，D-15-33, p. 227 (March 2008).
[14] 山口和範，高橋淳一，竹内光悦, “図解入門 よくわかる多変量解析の基本と仕組み, “ (株)秀和システム (June 1, 2004).

付録

なし