Vol.2013-CLE-9 No.6
2013/2/1

講義内容の要約字幕支援システム ～重要文自動抽出手法の提案（その 2）～

工藤永貴[†1]　千葉亮太[†1]　八重樫理人[†2]
上之薗和宏[†3]　古宮誠一[†1]

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の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．本稿では，意思決定手法を利用して，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．

A System to Help with Making Subtitles Condensed a Lecture Content: A Proposal for a Method to Elicit Key Sentences Automatically (Part 2)

HISAKI KUDO[†1]　RYOTA CHIBA[†1]　RIHITO YAEGASHI[†2]
KAZUHIRO UENOSONO[†3]　SEIICHI KOMIYA[†1]

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

1. はじめに

マレーシア人学生の工学系大学への留学のための予備プログラムとして JAD プログラム（Japan Associate Degree Program）[1][2]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 年しかたっておらず，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，日本語による講義の発話を要約した字幕が，日本語を非母国語とする学生に講義を理解させるのに有効である[4]と述べている．

HELP1　予備教育２年　日本に大学１年から入学

	1年目	2年目	3年目	4年目	5年目	6年目
コース	予備教育		大学教育			
年次	予備1年	予備2年	大学1年	大学2年	大学3年	大学4年
場所	マレーシア		日本			

⇩

HELP2　2+3ツイニング　大学２年次編入

	1年目	2年目	3年目	4年目	5年目
コース	予備教育	大学教育	大学教育		
年次	予備1年	予備2年	大学2年	大学3年	大学4年
場所	マレーシア		日本		

⇩

HELP3　3+2ツイニング　大学３年次編入

	1年目	2年目	3年目	4年目	5年目
コース	ディプロマコース			大学教育	
年次	予備1年	大学1年	大学2年	大学3年	大学4年
場所	マレーシア			日本	

図 1　JAD プログラムの教育システム

Figure 1　The education system of JAD Program.

そこで，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

†1 芝浦工業大学理工学研究科
Graduate School of Engineering and Science, Shibaura Institute of Technology
†2 香川大学
Kagawa University
†3 青山学院大学
Aoyama Gakuinn University

多大な労力がかかるので手作業ではなく，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．

(A)　帰納推論に基づく抽象化による方法

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，帰納推論を用いているため，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抜粋することにより，要約文を作成する方法である．この方法は，必要な文を自動的に抜粋する処理なので，コンピュータには比較的容易な処理だと思われる．

上記の 2 つの方法を比較すると，要約の実現は(A)よりも(B)のほうが容易だと思われるので，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．

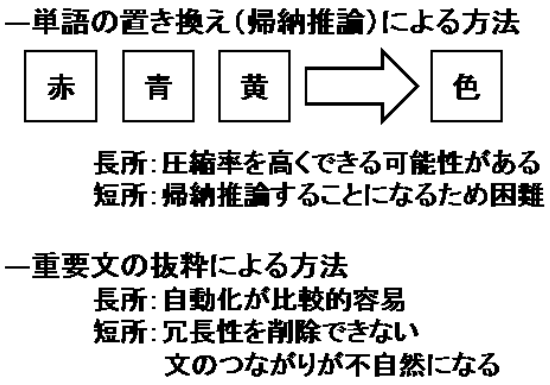

図 2　それぞれの要約作成方法における長所と短所
Figure 2　The merits and demerits of each summing-up method.

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．

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いる重要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，こ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となり得る語句を自動抽出して，それらの中から重要と思われるキーワードを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という方法を採用する．

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，その処理過程を下記の 5 つに分解するとともに，それぞれの過程を自動化することにより実現する．

1.　要約対象となる文章(＝講師の発話テキスト)を読み込む
2.　キーワードの候補となる語句を自動抽出する
3.　抽出されたキーワードを講師が評価・分類する
4.　各キーワードに与えた評価値を文ごとに集計する
5.　文ごとに集計された評価値に基づき要約文を自動生成する

次節から具体的に説明していく．

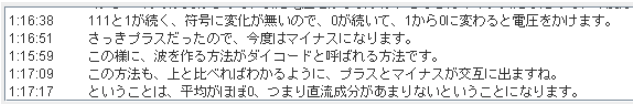

図 3　入力画面
Figure 3　Entry screen.

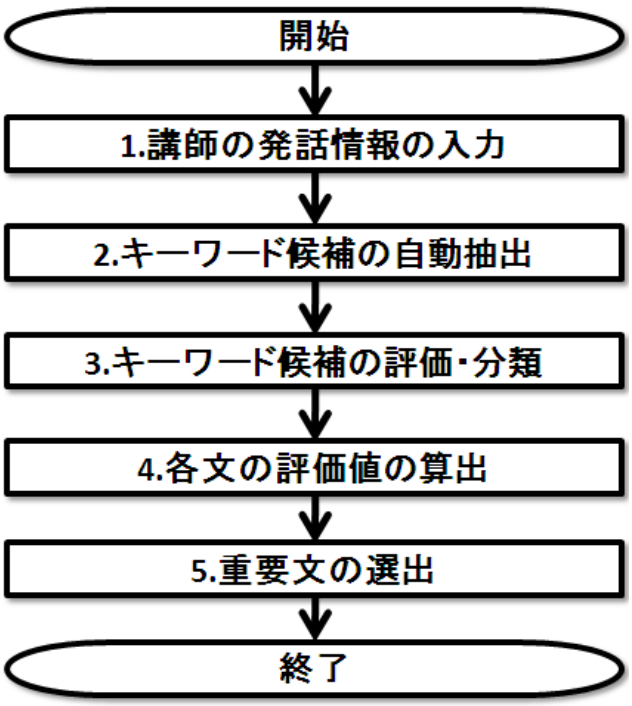

図 4　処理の流れ
Figure 4　Processing flow.

2.1　要約対象となる文章（講師の発話テキスト）の読み込み

システムへの入力情報は，既存の音声情報文字化ソフト

を用いて，講師の発話(音声)情報をテキスト情報へ自動変換することによって作成する．そして，このようにしてできた各文に，発話タイミングの情報を付与したものを入力情報とする．

入力情報の例を図 5 に示す．

```
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 5　入力情報の例

Figure 5　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

入力情報として与えられた文章から，キーワードの候補となる語句を自動抽出する．形態素解析には chasen [8]を用いた．事例となる文を形態素解析して得られた結果からキーワードの候補を抽出するには，下記のような特徴を持った語句を抽出すればよいことが実験によって判明した．

- 1 つまたは 2 つ以上連続している語句の品詞が，下記のいずれかの組み合わせであること
  - 名詞_一般
  - 名詞_固有名詞_一般
  - 名詞_固有名詞_地域_一般
  - 名詞_サ変接続
  - 名詞_数
  - 記号_アルファベット
  - 記号_一般
  - 記号_括弧開
  - 記号_括弧閉
  - 未知語
  - 名詞_接尾_一般
  - 名詞_接尾_サ変接続
  - 名詞_接尾_助数詞

```
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 6　Chasen による解析の結果の例

Figure 6　An example of analysis result with the use of the Chasen.

上記のルールによるキーワード候補の選出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．

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.

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

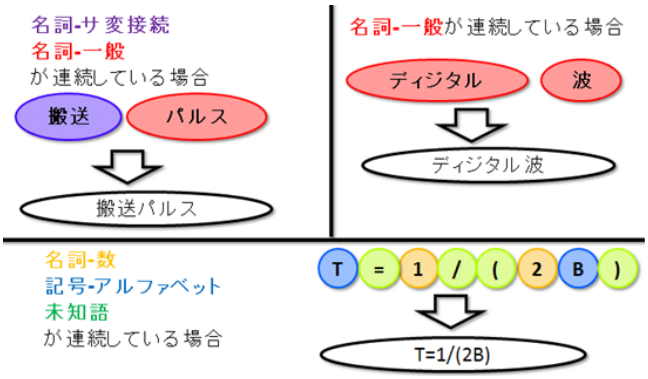

図 7　キーワード候補の例

Figure 7　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

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」については，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 段階に分類する．

「絶対に理解して欲しい」に分類されたキーワードには，この文が無条件に重要文と見なされることを示す「MUST フラグ」を，この文にシステムが自動的に付与するのみで，キーワードには得点を与えない．「できれば理解して欲しい」に分類されたキーワードのうちで，「重要である」に分類されたキーワードには 4 点を，「どちらかと言えば重要で

ある」に分類されたキーワードには 3 点を,「どちらかと言えば重要でない」に分類されたキーワードには 2 点を,「重要でない」に分類されたキーワードには 1 点をシステムが自動的に付与する.

講師がキーワードとして選ばなければ，システムが自動的に「無視してよいキーワード」に分類する.

評価の段階と評価値との対応関係を表 2 に示す.

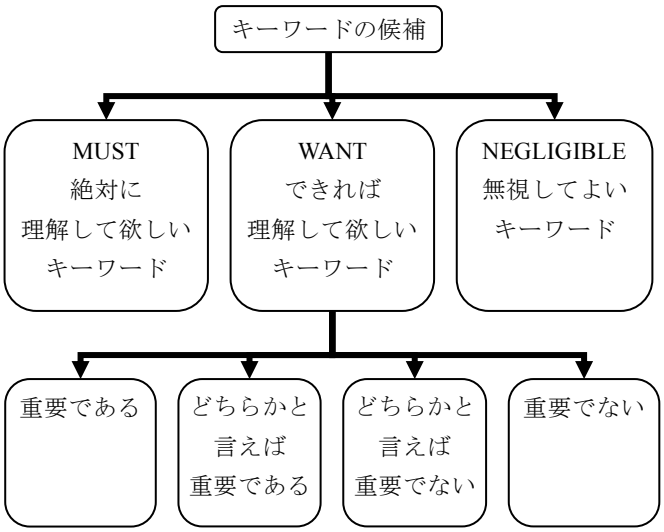

図 8　各キーワードの評価とそれに基づく分類

Figure 8　Evaluation of each keyword and evaluation-based classification of keywords.

表 2　段階別の評価値

Table 2　Graded evaluation value.

段階		評価値	フラグ
学生に絶対に理解して欲しいキーワード		MUST フラグを付与	M
学生にできれば理解して欲しいキーワード	重要	4 点を付与	A
	どちらかといえば重要	3 点を付与	B
	どちらかといえば重要でない	2 点を付与	C
	重要でない	1 点を付与	D
無視してよいキーワード		なし	なし

さらに，文中に含まれる最も高い得点のキーワードが「絶対に理解して欲しい」に分類されたキーワードであれば，そのことを示す MUST フラグ（M）を,「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Ａフラグ（A）を,「どちらかと言えば重要」に分類されたキーワードであれば，そのことを示すＢ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Ｃ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Ｄフラグ（D）を，文ごとにそれぞれ付与する.

本稿では,「絶対に理解して欲しい」に分類されたキーワードは M1, M2,･･･で,「重要である」に分類されたキーワードは A1, A2, ･･･で,「どちらかと言えば重要である」に分類されたキーワードは B1, B2, ･･･で,「どちらかと言えば重要でない」に分類されたキーワードは C1, C2, ･･･で,「重要でない」に分類されたキーワードは D1, D2, ･･･で,「無視して良いキーワード」に分類されたキーワードは N1, N2, ･･･で，それぞれ表記するものと約束する．キーワードの表記と講師が与えた評価との対応関係を表 3 に示す.

表 3　キーワードの表記と講師が与えた評価との対応関係

Table 3　Correspondence between the keyword notation and evaluation by given an instructor.

キーワードの表記	絶対に理解して欲しい	できれば理解して欲しい				無視してよい
		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	重要でない	
M1, ...	○					
A1, ...		○				
B1, ...			○			
C1, ...				○		
D1, ...					○	
N1, ...						○

2.4 キーワードの評価値に基づく文ごとの集計方法

まず，文中に含まれるキーワードを調べる．その結果，文中に含まれる最も評価値の高いキーワードに対応するフラグを文に付与する．即ち，文中に含まれる最も評価値の高いキーワードが,「絶対に理解して欲しいキーワード」に分類されたキーワードであれば，そのことを示す「MUST フラグ」を,「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Ａフラグ（A）を,「どちらかと言えば重要」に分類されたキーワードであれば，そのことを示すＢ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Ｃ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Ｄフラグ（D）を，文ごとにそれぞれ付与する.

次に，評価値を集計する.「絶対に理解して欲しいキーワード」が文中に含まれている文では，評価値の集計を行わない.「絶対に理解して欲しいキーワード」が文中に含まれていない文は，次のように集計していく.「重要である」に分類されたキーワードには４点,「どちらかと言えば重要である」に分類されたキーワードには３点,「どちらかと言えば重要でない」に分類されたキーワードには２点,「重要でない」に分類されたキーワードには１点を与える（「無視し

てよい」に分類されたキーワードは０点とする)．そして，それぞれに分類されたキーワードが，１つの文中にそれぞれ何回出現したかを数え，集計（積和計算）して得られた値を，その文の評価値とする．

文ごとに付与するフラグと評価値の求め方(積和計算)の具体例を表 4 に示す．

表 4　文ごとに付与する評価値の求め方の具体例
Table 4　A concrete example of a method for calculating the evaluation value to give each sentence.

文番号	発話テキスト(テキスト)	フラグ	文の評価値（得点）
1	○○○M1○○○N1○D1○	M	評価値の集計は行わず，その文を無条件に選出する
2	○A1○○○D5○N2○○A2○	A	4×(n1+n3)+n2
3	○B1○○B2○○N3○D3	B	3×(n4+n5)
4	○○○C1○○C2○○B3○D4○○	B	2×(n6+n7)+3×n8+n9
5	○D5○○N3○○N4○○D6○○	D	n10+n11

（注）n1, n2, … は文中での各キーワードの出現回数

発話テキストがその順序を変えずに，表 5 に示す表形式の入力バッファの上に読み込まれ，表 4 に示された方式に従って，文の文字数がカウントされ，文の評価値(評価値)，フラグが付与される．

表 5　発話情報の入力バッファの形式
Table 5　The format of input buffer of utterance information.

優先順序	フラグ	文の評価値	文の文字数	発話テキスト

2.5 文ごとの評価に基づく要約文の編集方式

システムが自動生成する要約字幕の編集方針としては，次の２つのモードがシステムに用意されている．

－　重要度の下限となるフラグ名を指定する方式
－　要約字幕の編集に要する最大文字数を指定する方式

これらの編集方式の処理詳細を以下に示す．

(1)　重要度の下限となるフラグ名を指定する方法

これは，MUST フラグ(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：　MUST 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：　MUST フラグとＡ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：　MUST フラグ，Ａフラグ，Ｂ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：　MUST フラグ，Ａフラグ，Ｂフラグ，Ｃ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：　MUST フラグ，Ａフラグ，Ｂフラグ，Ｃフラグ，Ｄ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．

重要文の編集に際しては，講師が発話した順序を変えることなく，各文の取捨選択を行うことに注意されたい．

この方式の処理の流れを図 9 に示す．

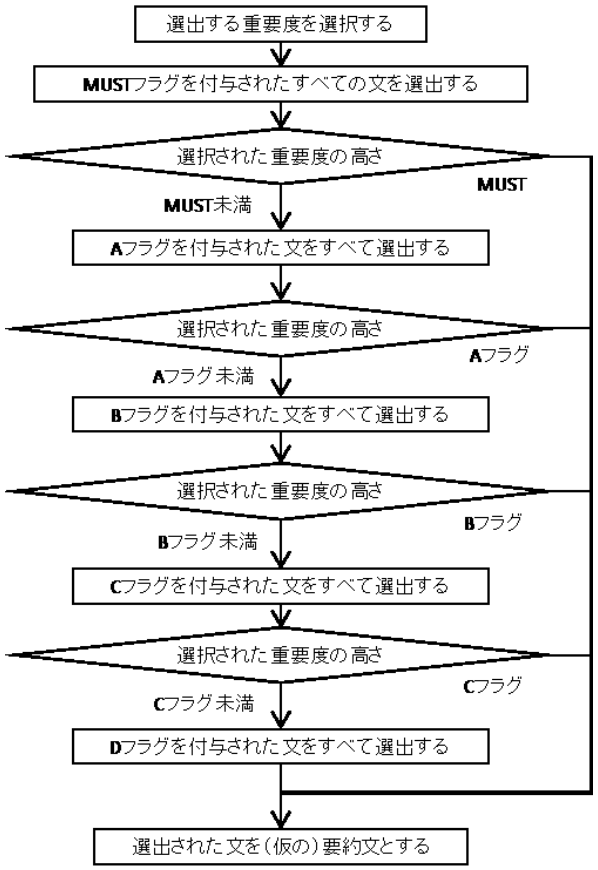

図 9　重要度の下限となるフラグ名を指定する方式の処理の流れ
Figure 9　Processing flow of the system to specify a flag name represented lower limit of important degree.

(2)　要約字幕の編集に要する最大文字数を指定する方式

最初に，指定された要約字幕の最大文字数の 2 倍を要約

字幕編集用のバッファメモリとして確保する．しかる後に「MUST フラグ」を付与された文のすべてを重要文と見なして選出候補とする．次に，文ごとに付与された評価値(得点)の最も高いものを１番とし，評価値(得点)が低くなるにつれて数値が大きくなるように，重要文を選出する際の優先順序を表す番号を付与する．これらの処理を，各文を構成する文字数を加算しながら行う．そして，文字数の合計が，指定された要約字幕の最大文字数の 2 倍に達したところで，この処理を終了する．このとき，オーバーフローした文は，バッファ内に収納される部分を含めて，その文の全体を廃棄する．指定された要約字幕の最大文字数の 2 倍弱を重要文の候補として残すので，このようにしても問題はないと考えられる. 何故なら, このとき廃棄される文と，優先順序が同じ文だけで，文字数が指定された要約字幕の最大文字数を超えるとは考えにくいからである．

このような準備をした上で，発話された順序を変えずに，入力バッファ表 5 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．

この方式の処理の流れを図 10 に示す．

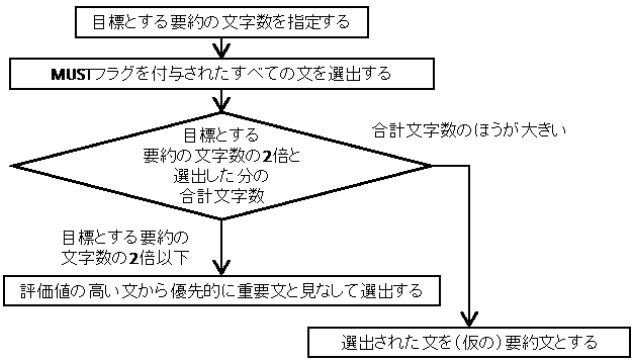

図 10　要約字幕の編集に要する最大文字数を指定する方式の処理の流れ

Figure 10　Processing flow of the system to specify the maximal number of characters required to summing up subtitles.

3. 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ていた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かったという問題点があった．我々は，講師の発話テキストからキーワードの候補を計算機で自動抽出し, キーワード候補を講師に選択，６種類に分類してもらい，それと出現回数を基にキーワードの重要度を決定，キーワードの重要度を基に自動的に要約文を作成することでこれらの問題を解決した．要約文における冗長部分の削除方式は後日，別の論文で発表する．

謝辞　本研究は文部科学省平成 18 年度サイバーキャンパス整備事業における「バーチャルワンキャンパス計画：芝浦工業大学」の支援を受けた．また本研究において用いた講義コンテンツ及び発話テキストは，芝浦工業大学システム工学部電子情報システム学科三好匠准教授に提供いただいた．記して感謝を申し上げます．

参考文献

1)　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/backgraound.html
2) 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
3) 八重樫理人, 佐々木良造, 石松純, 尾沼玄也, 山下哲生, 橘雅彦, 小林孝郎: JAD プログラムにおける日本語学習進捗状況共有　システムの提案及びその実装方法, メディア教育研究　第3巻　第2号 Journal of Multimedia Aided Education Research 2007, Vol. 3, No. 2, 143－150
4) 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.
5) 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, Vol9, No.4, pp.97-116
6) H. Luhn: The Automatic Creation of Literature Abstracts, IBM Journal of Research and Development, Vol.2, No.2, pp.159-165 (1958).
7) JUMAN <2009 年 1 月現在>, http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
8) ChaSen <2009 年 1 月現在>, http://chasen-legacy.sourceforge.jp/
9) KNP <2009 年 1 月現在>, http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html
10)　Richard Alterman, Lawrence A. Bookman, Reasoning about a semantic memory encoding of the connectivity of events, Cognitive Science　Vol 16, Issue 2, April-June 1992, Pages 205-232
11)　Edmundson, H: New methods in automatic abstracting, Journal of ACM, 16 (2), pp.264-285 (1969).
12)　Skorochod'ko: Adaptive Method of Automatic Abstracting and Indexing, in Proceedings of the IFIP Congress 71, pp. 1179-1182 (1972).
13)　J. R. Hobbs: On the Coherence and Structure of Discourse, CSLI Report No.CSLI-85-37, CSLI, (1985).
14)　Mann,W .and Thompson,S: Rhetorical Structure Theory: A Framework for the Analysis of Texts,Technical report, Techical Report ISI/RS_87_185, Marina delRey, California (1987).
15)　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.
16)　Weka Machine Learning Project <2009 年 1 月現在>, http://www.cs.waikato.ac.nz/ml/
17)　I.H. Witten, E. Frank: Data Mining: MORGAN KAUFMANN: ISBN 1-55860-552-5
18)　金明哲: WEKA と樹木モデル, ESTRELA　No.132 pp.64-69, 2005 年 3 月
19)　I.H. Witten, E. Frank: Data Mining, MORGAN KAUFMANN: ISBN 1-55860-552-5
20)　工藤拓, 松本裕治, チャンキングの段階適用による係り受け解析, 情報処理学会論文誌, Vol 43 No. 6, pp.1834-1842, June 2002.
21)　大津: 判別および最小２乗基準に基づく自動しきい値選定法, 電子通信学会論文誌, Vol. J63-D, No.4, pp.349-356 (1980).
22)　大津: パターン認識における特徴抽出に関する数理的研究, 電子技術総合研究所研究報告, Vol.818 (1981).