VOD講義用字幕文の簡易化とやさしい表現への変換手法

則本　達哉[1]　羅　毅剛[2]　椎名　広光[3]　北川　文夫[4]
norimototatsuya@gmail.com[1], luoyigang@hotmail.co.jp[2],
shiina@mis.ous.ac.jp[3], kitagawa@mis.ous.ac.jp[4]
岡山理科大学大学院　総合情報研究科[1,2]
岡山理科大学　総合情報学部[3,4]

1　まえがき

現在，講義を動画で配信するVOD（Video On Demand）講義は様々な大学で利用されており，VOD講義の利用者は日本語を母国語とする学生だけではなく，各国の留学生も利用している．そのようなニーズに応えるべく，VOD講義と共に字幕を添付するシステムも存在する．

しかし，VOD講義の講演者の発言を字幕にしただけでは，講義内容が難しく字幕表示量も多いため，日本語を母国語としない留学生にとって理解は困難であると考えられる．そこで，学習者の理解をより深めるために，講演者の大まかな意図が表現できればよいとして，表層的な手法による字幕のやさしい表現への換言と，字幕文の簡易化を提案する．初めのやさしい表現への換言は，日本語能力試験[1][2]の旧試験で使用されている単語の級数データを用いて字幕文の動詞の級数を判定し，その動詞に対する日本国語辞書[3]の説明文にある文末動詞との級数を比較して換言する手法である．次に字幕文の簡易化は，変形した$tf \cdot idf$[6]を表示している字幕の前後に含まれる字幕文を一つの文書として計算し，係り受け構造と単語の$tf \cdot idf$値から不要な文節を削除する手法である．

また，提案した各手法を岡山理科大学で開講されているVOD講義「データベース」の字幕に対してユーザーアンケートによる評価を行った．

2　字幕文のやさしい表現への換言

講義で使用される専門用語には名詞が多く，講義内の名詞を換言すると講義の内容が大きく変わる可能性が高いと考えられる．そこで，本研究では字幕文の名詞は換言せず，字幕文の動詞のみを対象にしている．

動詞の換言は，日本語能力試験の級数データを用いて動詞の級数が低くなるように置き換える．また，換言の前後で難易度が同様の場合，よく使われる単語に換言する．換言後の単語がよく使用される単語であれば，やさしくなったとしている．

表1: 換言対象の種類数と出現回数

級別	換言対象種類数	換言対象出現回数
1級	2	2
2級	34	76
3級	26	93
4級	29	222
合計	91	393

2.1　換言対象の抽出

日本語能力試験の1級や2級の認定目安は，高度な文法や語彙が理解できることである．また，基本的な日本語を理解するには3級相当の能力が必要である．

本研究では，日常的な場面で使われる日本語の理解を促すためにも，1級または2級の動詞を換言対象として同一級以下の単語に換言する．換言対象の抽出手順を以下に示し，字幕文に含まれる動詞の級別の種類数と出現回数を表1に示す．

[換言対象の抽出手順]
(1) 形態素解析器ChaSen[4]を用いて字幕文を解析し，字幕文の動詞を抽出する．
(2) 抽出した動詞を日本語能力試験の級数データから検索し，動詞の級数を取得する．
(3) 級数が1級または2級である場合，その動詞を換言対象とする．

上記の手順を適用した字幕中に出現する換言対象の抽出例を以下に示す．なお，例文中の下線部が換言対象の動詞である．

[換言対象が1級の抽出例]
永続性を保つということは仕組みとしてなされてい

表 2: 日本語辞書による換言候補（下線部）の抽出例

見出し語	説明文（意味）
喋る	ものを 言う。
用いる	道具や方法をその用にあてて 使う。
取り出す	中から取って外へ 出す。
示す	はっきりとわかるように出して 見せる。
区切る	広さや長さをもつものに境目を入れて いくつかに 分ける。

表 3: 換言候補の出現回数（括弧内は種類数）

		換言前		
		1 級	2 級	合計
換言後	1 級	0(0)	—	0(0)
	2 級	0(0)	25(9)	25(9)
	3 級	1(1)	3(3)	4(4)
	4 級	0(0)	32(13)	32(13)
	合計	1(1)	60(25)	61(26)

るわけです。
[換言対象が 2 級の抽出例]
次にＢさんが三千円 引き出す わけですから、〜。

2.2 換言候補の抽出

字幕文から換言対象を抽出した後，その換言対象に対するやしい表現の換言候補の抽出を行う．日本語国語辞書の文末動詞の級数が換言対象の級以下の場合，その単語を換言候補とする．以下に換言候補の抽出手順を示し，日本国語辞書から抽出した換言候補の例を表 2 に示す．

[換言候補の抽出手順]
(1) 換言対象の単語を日本語国語辞書で検索し，換言対象の説明文を取得する．
(2) 説明文（意味）を ChaSen を用いて解析する．
(3) 解析した結果から説明文の文末動詞を抽出し，動詞が換言対象の級以下の場合は，換言候補とする．

字幕中に出現した換言対象のうち，日本語能力試験級が同一級以下になった換言候補の出現回数と種類数を表 3 に示す．

2.3 換言候補の活用変化

日本語国語辞書から抽出した換言候補の活用形は基本形が多く，換言対象の活用形との整合を取る必要がある．以下に ChaSen を用いた換言対象に対する換言候補の活用変化手順を示し,「取り出してみます」に対する変形例を図 1 に示す．

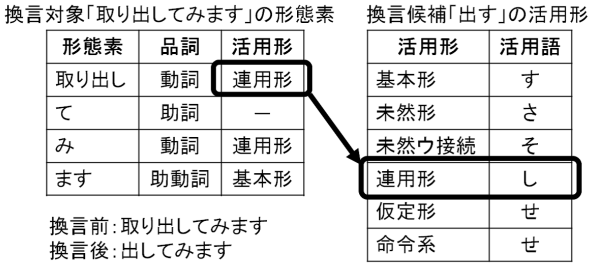

図 1: 動詞活用形の変換例

表 4: 級が下がった単語の難易度変化の評価

評価者	やさしくなった	変わらない	難しくなった
日本人 1	34	1	1
留学生 1	18	18	0
留学生 2	17	9	10
留学生 3	33	0	3
留学生 4	13	16	7
中国人 1	24	3	9

表 5: 級が下がった単語の換言文の妥当性評価

評価者	正しい	正しくない
日本人 1	26	10
留学生 1	27	9
留学生 2	21	15
留学生 3	27	9
留学生 4	27	9
中国人 1	36	0

[換言候補の活用変化手順]
(1) 換言候補の活用語を消去する．
(2) 換言対象の活用形を取得する．
(3) 換言対象の活用形に対応する換言候補の活用語を取得する．
(4) 取得した活用語を換言候補に付加する．

2.4 やさしい表現への換言評価

本研究で提案したやさしい表現への換言により，日本語能力試験の級数が下がった 36 単語と，換言の前後で級数が変わらない 25 単語が得られた．それらの単語に対して，難易度の変化と妥当性を評価する学生アンケートを行った．学生の内分けは，日本人 1 人，漢字語圏からの留学生 4 人，中国に在住している学生 1 人の計 6 人である．なお，留学生の日本滞在歴は，留学生１は 3 年，留学生２は 1 年，留学生３は 4 年，留学生は 3 年である．中国人１は中国で日本語の専門学校に通い，4 年間日本語を勉強している．

表 6: 同一級の換言の難易度変化の評価

評価者	やさしくなった	変わらない	難しくなった
日本人 1	5	12	8
留学生 1	18	7	0
留学生 2	6	13	6
留学生 3	10	11	4
留学生 4	6	8	11
中国人 1	16	1	8

表 7: 同一級の換言文の妥当性評価

評価者	正しい	正しくない
日本人 1	13	12
留学生 1	22	3
留学生 2	15	10
留学生 3	20	5
留学生 4	22	3
中国人 1	24	1

2.4.1 換言によって級が下がった単語に対する評価

換言後，級数が下がった単語に対して評価した結果を表 4 と表 5 に示す．

表 4 から，日本人 1 と留学生 3 の評価が近く，留学生 1 も含めて多くの換言文についてやさしくなったと評価している．一方，留学生 4 や中国人 1 は難しくなったと評価した割合が大きい．このことから，日本語の能力によって評価が分かれ，勉強期間が長くなると厳しい評価をする傾向にあると考えられる．逆に，留学生 1 のように勉強期間が短い場合は，難しくなった評価が低く，やさしい換言が有効であると考えられる．また，表 5 では，日本人 1 の評価に近い評価者が多いなか，中国人 1 は全ての換言文が正しいと評価している．換言元の字幕は話し言葉であり，留学経験のない中国人 1 は話し言葉に慣れていないためではないかと考えられる．

2.4.2 換言後同一級の単語に対する評価

換言の前後で級数が同一である単語に対して評価した結果を表 6 と表 7 に示す．

表 6 の難易度変化のアンケートでは，表 4 に比べてやさしくなったと評価した割合が減少する傾向にある．日本人 1 や留学生 4 などはあまりやさしくなったと評価していない．しかし，今回の評価者の中で留学経験のない中国人 1 は，多くの換言についてやさしくなったと感じている．また，表 7 より日本人 1 は約半数が否定的であるのに対して，中国人 1 はほとんどの換言文が正しいと評価している．これは，文を単語で理解しようと試みているためであると考えられる．

3 係り受け解析を利用した文短縮

VOD 講義での講演者の発言をそのまま字幕にしただけでは，字幕の表示量が多く日本人でも目で追うことは困難である．

本章では日本語を勉強中の留学生を対象として，VOD 講義の字幕の表示量を抑えるために，文の係り受け構造と $tf \cdot idf$ 値を利用した文節単位の文短縮手法を提案する．

3.1 字幕の重要単語抽出

VOD 講義の字幕から講義内で重要である単語を抽出する．VOD 講義全体の重要語以外にも，講師が今話している付近から得られる重要語も必要であると考えられる．そこで，$tf \cdot idf$[6] を表示している字幕の文の前後から計算するように変更する．

字幕を表示している講義 D_j の文 s の前後 d 時間の字幕を一つの文書 $D_j^{s,d}$ とみなして，$tf \cdot idf(i, D_j^{s,d})$ を次のように定義する．

[字幕文の前後を用いた $tf \cdot idf$ の定義]

単語 w_i が文書 $D_j^{s,d}$ に含まれる数 $tf(i, D_j^{s,d})$，時間区間 $2d$ で区切った文書数 N，単語 w_i を中心とした時間区間 $2d$ で区切った文書の数 $df(i)$ とするとき，以下のように定義する．

$$tf \cdot idf(i, D_j^{s,d}) = tf(i, D_j^{s,d}) \cdot \log \frac{N}{df(i)}.$$

3.2 係り受け解析による文短縮手法

字幕文の係り受け構造を係り受け解析器 CaboCha[5] を用いて解析し，係っている側の文節の重要度を前節で計算した $tf \cdot idf(i, D_j^{s,d})$ から求め，文節を削除する順序を計算する．次に，表示する文字量から削除の優先順位が高い文節を削除した文を新しい字幕として表示する．以下に $tf \cdot idf_c(i)$ と評価値 $T_c(i)$ の定義と係り受け短縮アルゴリズムを示す．また，短縮例を図 2 に示す．

[定義]

文 s，文 s の文節 $C(s) = C_1, C_2, \ldots, C_n$，各文節 C_i がかかる先 $A(C_i)$，各文節 C_i の $tf \cdot idf$ 値 $tf \cdot idf_c(i)$，各文節 C_i の評価値 $T_c(i)$ とする．

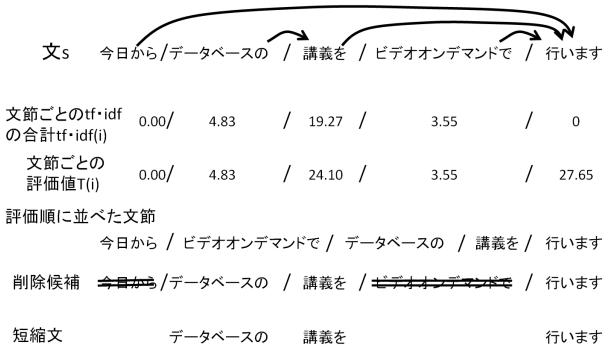

図 2: 字幕の文節の評価と削除例

表 8: 係り受け短縮の文字削減量

字幕文字数	短縮後文字数
20 文字以下	対象外
24 文字以下	20 文字以下
30 文字以下	24 文字以下
36 文字以下	30 文字以下
48 文字以下	36 文字以下
49 文字以下	対象外

[文節削除アルゴリズム]
(1) 各文節 C_i 内の単語 w_j の $tf \cdot idf$ の合計を計算する．

for $i = 1..n$ do
$$tf \cdot idf_c(i) = \sum_{w_j \in C_i} tf \cdot idf(j, D_j^{s,d})$$

(2) 各文節が係っている先に合計した $tf \cdot idf_c(i)$ を足し合わせる．

(2-1) for $i = 1..n$ do
$$T_c(i) = tf \cdot idf_c(i)$$
(2-2)for $i = 1..n$ do
if $A(C_i) = C_j$ then
$$T_c(j) := T_c(j) + T_c(i)$$

(3)$T_c(j)$ の小さい順に C_i を並び替え，優先的に削除する．

3.3 係り受け解析を利用した文短縮の評価

3.2 節で述べたアルゴリズムを用いて，短縮文の評価を行う．字幕の中には文短縮する余地のない短い文や，係り受け構造が複雑である長い文が存在する．特に長い文については係り受けを利用した文短縮は適用しにくいことが分かっている．そこで，短縮対象を元の字幕が 20 文字以上 48 以下の字幕文に限定し，表 8 の文字数に文字を削減する．係り受けを利用した文短縮を適用した短縮文を，日本人 1 人，中国人留学生 1 人に対してアンケートを実施した結果を表 9 に示す．

表 9: 係り受け短縮の評価

	短縮成功文	意味が変わった文	非文になった文	成功割合
日本人 1	138	16	14	0.821
留学生 1	137	18	13	0.815

表 9 から，日本人と留学生ともに 8 割程度の文に対して上手く短縮できていると答えた．しかし，本来の意味が変わってしまった文や非文も存在するため，係り受け構造や $tf \cdot idf$ 値以外にも格構造などを利用することを考えている．

4 まとめと今後の課題

本研究では，VOD 講義の字幕に対するやさしい表現への換言手法と，係り受け解析を利用した文短縮手法を提案し，VOD 講義「データベース」の字幕に対して評価した．

字幕の換言後，日本語能力試験の級数が下がった単語の評価では，個人差はあるものの概ねよい評価が得られた．しかし，換言後同一級の単語には，個人の日本語能力も影響して，よい評価が得られなかった．

また，係り受けを利用した文短縮について，8 割以上の精度で文短縮に成功したと言える．

今後の課題として，換言手法に関して，換言候補の抽出に単一の国語辞書だけではなく複数の辞書やウェブの組み合わせについても考慮している．また，係り受け構造を利用した文短縮についても，評価値 T_c の改良や，より多くのユーザアンケートによる評価が必要である．

参考文献

[1] 日本語能力試験, http://www.jlpt.jp/

[2] 徳弘康代, 日本語学習のためのよく使う順漢字 2100, 三省堂.

[3] 北原保雄, 明鏡国語辞典, 大修館書店.

[4] 形態素解析システム茶筌, http://chasen.naist.jp

[5] 日本語係り受け解析器南瓜, http://code.google.com/p/cabocha/

[6] 北, 津田, 獅々子, 情報検索アルゴリズム, 共立出版，2002.