講演からの重要文抽出のための評価データの選択と重要文の特徴の分析

森脇 雅人　　南條 浩輝　　吉見 毅彦

龍谷大学 理工学部 情報メディア学科

e-mail: moriwaki@nlp.i.ryukoku.ac.jp

1　はじめに

話し言葉（講演）からの重要文抽出を考える．話し言葉では重要文が明示的でないことが多く，その認定は容易ではない．このことは日本語話し言葉コーパス（CSJ）での人間 3 名による重要文の認定において結果が一致しないことが多いことからもわかる．このため，話し言葉からの重要文抽出においては完全な正解が存在することは稀であり，自動抽出手法の評価は人間の抽出精度（人間同士での一致度）との比較という形で行われるのが一般的である．ただし，無作為に抽出して作成した評価データの中には人間同士での一致度が低いデータが存在する．このようなデータには重要文が存在しない，すなわち重要文を適切に定義できないため，これらは評価データとして適切ではない．

このような背景に基づき，本研究では，はじめに人間同士で重要文認定に一定の一致度が得られるデータを重要文抽出の適切な評価データとして選択し，次にそれを用いて重要文抽出手法の評価を詳細に行う．最後に重要文の特徴を分析した結果について述べる．

2　講演からの重要文抽出

2.1　日本語話し言葉コーパス

本研究では，日本語話し言葉コーパス（CSJ）を利用して講演からの重要文抽出の研究を行う．CSJ は主に学会講演と模擬講演からなる大規模話し言葉コーパスであり，コアとよばれる一部の講演（学会講演 70，模擬講演 110）とコアには含まれないが音声認識のテストセットである 22 講演（学会講演 15，模擬講演 7）に対しては，作業者 3 名により重要文タグが抽出率 50%および 10%で付与されている．本研究では，これらの講演のうち 195 講演を用いる[1]．

次に CSJ における「文」の定義について述べる．CSJ においては形式上明示的な文末表現（絶対境界）と，いわゆる文末ではないが発話の大きな切れ目として考えられる節境界（強境界）を文境界として「文」を定義している [2]．これは話し言葉では文の定義自体が必ずしも自明でないためである．なお，CSJ での「節」は「述語を中心としたまとまり」であり，話し言葉においても明確に判定でき，また，係り受け構造としてもある程度独立な処理単位である．本研究での「文」はこの定義に基づく．

[1]これは，本研究では重要文抽出の先行研究 [1] の結果を一部利用しており，そこで用いられていたデータと整合性がとれないものを除いたためである．

表 1: κ 値の解釈表

κ 値	値の解釈
$\langle$ 0	Poor
.00〜.20	Slight
21〜.40	Fair
.41〜.60	Moderate
.61〜.80	Substantial
.81〜1.0	Near perfect

2.2　重要文抽出手法の評価方法

次に重要文の定義と重要文抽出の評価方法について述べる．本研究では，先行研究 [1] と同じ方法で重要文を定義する．すなわち，CSJ の重要文タグ付与者の 3 名から 2 名を選び，抽出率 10%の場合は 2 名のうち少なくとも一方が選んだ文，抽出率 50%の場合は 2 名がともに選んだ文を重要文（正解）とする．

正解データは講演ごとに 3 通りできるため，評価はそれぞれの正解を用いて算出した結果の平均とする．このとき，正解データ作成に使用しなかった，作業者 1 名が付与した重要文タグを用いて人間の抽出精度を算出することができる．これにより自動抽出精度と人間の抽出精度とを比較する．人間がもしくは自動で抽出した重要文と正解データの一致度（抽出精度）は κ 値を用いて評価する．κ 値とは 2 つのデータ間で偶然に一致する割合を除いた一致度の尺度である [3]．κ 値の解釈を表 1 に示す．

2.3　談話標識と話題語を用いた講演からの重要文抽出

次に重要文抽出手法について述べる．本研究では談話標識と話題語を利用する手法 [1][4] を用いる．談話

標識を利用する手法は，段落の冒頭に重要文が存在すると考え，段落の冒頭文に特徴的に出現する単語（談話標識）を用いて段落の推定と重要文抽出を同時に行う手法である．

談話標識の学習と，それを用いた重要文抽出は次のように行う．まず，学習データからポーズ長に基づき段落の境界候補を抽出し，その直後の文の集合をつくる．次に，この文集合に特徴的に出現する単語（名詞）を談話標識として選択する．具体的には式（1）で定義される統計量 DM に基づき選択する．

$$DM_m = wf_m * \log(\frac{Ns}{sf_m}) \quad (1)$$

ここで wf_m は段落の冒頭文の集合での単語 m の出現回数，sf_m は単語 m が現れた文数，Ns は学習セットの全文を表す．段落の冒頭文に頻出し，その他に出現しない単語に対して統計値 DM は大きい値をとる．各文 s_j の談話標識に基づく重要度を，その文に出現した異なる全ての談話標識の統計値の合計値 $S_{DM}(j)$（式（2））とする．この合計値 $S_{DM}(j)$ は段落の文頭らしさ（＝重要文らしさ）を表すので，この値の高い文から順に重要文として抽出する．

$$S_{DM}(j) = \sum_{m_i \in s_j} DM_{m_i} \quad (2)$$

次に話題語に基づく手法について述べる．ここではtf・idf 法に基づく手法を用いる．すなわち，話題語の重要度として式（3）で定義される単語（名詞）の統計値 KW_m（tf・idf 値）を用いる．

$$KW_m = tf_m * \log(\frac{Nd}{df_m}) \quad (3)$$

ここで tf_m は当該講演内での単語 m の出現回数，df_m は単語 m が現れた講演数，Nd は全講演数を表す．当該講演に頻出し，他の講演にあまり出現しない単語，すなわち，その講演を特徴づける単語ほどその tf・idf 値は高い値となる．文の重要度は各文 s_j に含まれる話題語の tf・idf 値の合計値（式（4））とし，この値に基づいて重要文を抽出する．

$$S_{KW}(j) = \sum_{m_i \in s_j} KW_{m_i} \quad (4)$$

最後に両者の統合について述べる．これはそれぞれの手法で得られた文の重要度の重みつき幾何平均（式（5））を新たな文の重要度とし，重要度の高い順に定められた抽出率に達するまで重要文抽出を行うものである．

$$S(j) = S_{KW}(j)^{\alpha} * S_{DM}(j)^{1-\alpha} \quad (5)$$

この統合手法を利用して，評価データ全体（195 講演）を対象として重要文抽出を行った．抽出率 50%および 10%の場合の結果を表 2 に示す[2]．人間の抽出精度に対する自動抽出精度の割合は，抽出率 50%の場合は 69.4%，抽出率 10%の場合は 47.4%であった．学会講演，模擬講演ではそれぞれ，抽出率 50%の場合は 75.5%，64.7%であり，抽出率 10%の場合は 59.7%，37.9%であった．模擬講演を対象とした場合は抽出率 50%，10%とも学会講演と比べて低いことがわかる．抽出率 10%の場合は自動抽出精度（κ 値）自体も低いことがわかる．

表 2: 評価データ全体での重要文抽出精度（κ 値）

抽出率 50%

対象（講演数）	自動	人間	自動/人間
全体（195）	0.315	0.466	69.4%
学会（85）	0.329	0.449	75.5%
模擬（110）	0.304	0.478	64.7%

抽出率 10%

対象（講演数）	自動	人間	自動/人間
全体（195）	0.168	0.402	47.4%
学会（85）	0.194	0.381	59.7%
模擬（110）	0.147	0.419	37.9%

3 重要文抽出手法の評価のためのデータの選択

話し言葉における重要文認定では，人間同士でもあまり一致しないことが多い．さらに人間同士での一致度が低いデータ（講演）は，そもそも正解とすべき重要文が存在しないデータであると考えられるため，これらは，自動抽出結果の評価データとしては不適切である．本章では，評価の対象として適切な講演のみを選択し，これらを用いて重要文抽出手法の評価を行う．

3.1 適切な評価データの選択

はじめに人間同士での一致度に基づく評価データの選択について述べる．講演ごとの人間の重要文抽出精度（抽出率 50%および 10%），すなわち一致度（κ 値）のヒストグラムをそれぞれ図 1 に示す．講演の中には人間同士でも重要と判断する文が一致するものと一致しないものがあることがわかる．特に，抽出率 10%の場合は人間同士での一致度が低い講演が多い．

本研究では，重要文抽出における人間の一致度（κ 値）の解釈が Moderate（0.41）以上の講演を人間同士でも重要と判断する文の一致がみられる講演と定義し，これらを評価データとして選択した．これにより，自動重要文抽出手法のより正確な評価が可能と考

[2]“自動” および “人間” は自動および人間の抽出精度（κ 値），“自動/人間” は人間の抽出精度に対する自動抽出精度（%）を表す．

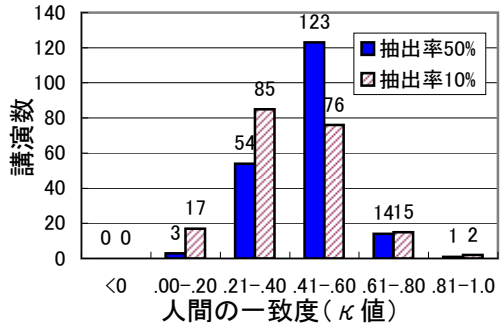

図 1: 講演ごとの重要文抽出における人間の一致度のヒストグラム

表 3: 選択した評価データでの重要文抽出精度（κ 値）

抽出率 50%（**D-50**）

対象（講演数）	自動	人間	自動/人間
全体（138）	0.399	0.519	65.4%
学会（53）	0.366	0.522	70.5%
模擬（85）	0.322	0.518	62.3%

抽出率 10%（**D-10**）

対象（講演数）	自動	人間	自動/人間
全体（93）	0.188	0.524	37.3%
学会（36）	0.231	0.519	47.2%
模擬（57）	0.161	0.527	31.1%

える．本研究では抽出率 50%用の評価データを **D-50**，抽出率 10%用の評価データを **D-10** と表記する．評価データの内訳は次のとおりである．

- **D-50**：138 講演（学会講演 53，模擬講演 85）
- **D-10**：93 講演（学会講演 36，模擬講演 57）

3.2 選択した評価データを用いた重要文抽出の評価

抽出率 50%および 10%の場合の重要文抽出精度の平均を表 3 に示す．全体で評価を行った結果（表 2 ）に比べて人間の抽出精度に対する自動抽出精度の割合が低いことがわかる．これは，評価データ全体の中には人間の抽出精度が低い講演も含まれており，自動抽出精度の絶対値が低いにもかかわらず，人間の精度に対する割合が高く算出されていたためである．この結果は，抽出率 50%の場合において，特に模擬講演で改善の余地が大きいことを示し，抽出率 10%では重要文抽出手法の抜本的な見直しが必要であることを示唆している．

D-10 と **D-50** を詳しく調べたところ **D-10** は **D-50** のサブセットでないことがわかった．すなわち，抽出率 10%の場合に一致がみられるが，抽出率 50%の場合に一致がみられない講演があることがわかった．

表 4: 重要文と非重要文の単語数の平均と標準偏差

		重要文			非重要文		
	抽出率	文数	平均	S.D.	文数	平均	S.D.
学会	50%	3477	31.4	19.0	3430	17.7	13.8
	10%	388	38.6	21.9	3815	24.0	17.2
模擬	50%	3960	28.3	16.8	3874	16.7	12.7
	10%	474	30.7	18.7	4573	21.3	14.9

S.D.: 標準偏差

また，その逆の傾向を示す講演もみられた．重要文抽出における人間の一致度が抽出率 50%の場合にのみ高い講演には，多数の重要文があるが非常に重要な文が存在しない（もしくは少ない）と考えられる．抽出率 10%の場合にのみ高い講演には，少数の非常に重要な文があるが重要な文がそもそも多くないと考えられる．このことは，重要文の分布の傾向にしたがって抽出手法を切替えることで精度の向上を得られる可能性があることを示している．

4 重要文の特徴の分析

最後に重要文抽出精度の向上の指針を得るために，**D-50** および **D-10** を用いて重要文の特徴の分析を行った．はじめに，重要文と非重要文の長さ（単語数）の比較について述べる．次に，重要文に頻出する単語および節について述べる．ここでの単語の単位は CSJ の短単位 [5] とした．重要文は作業者 2 名以上が重要と選んだ文と定義した．

4.1 長さに関する特徴

重要文と非重要文の長さ（単語数）の平均と標準偏差を表 4 に示す．すべての場合（学会講演，模擬講演の抽出率 50%，10%それぞれ）で，重要文での単語数が多かった．非重要文の単語数との差について 2 群の母平均の差の検定を行ったところ，すべての場合で有意水準 1%で有意であった．

4.2 単語と節情報に関する特徴

次に重要文に頻出する単語および節を調べる．本研究では重要文に頻出する単語を次の手順で抽出した．

1. 重要文に出現する単語 w について，講演ごとに重要文中での頻度と非重要文中での頻度の比を求める．
2. 全講演での頻度の比の平均（$Score_w$）を求める．
3. $Score_w$ があるしきい値 T 以上となる単語 w を重要文に頻出する単語とする．本研究では重要文と非重要文の文の数を考慮し，抽出率 50%，抽出率 10%の場合でそれぞれ T を 1，1/9 とした．

なお，重要文に頻出する節も同様の手順で抽出した．

表 5: 重要文に頻出する単語（上位 30）

学会講演	抽出率 50%	事	物	語	場合	中
		情報	意味	日本	研究	其れ
		様	其の	此の	考える	拠る
		行なう	出来る	居る	為る	分かる
		用いる	対する	言う	使う	見る
		有る	行く	就く	来る	無い
	抽出率 10%	今後	検討	研究	今回	目的
		為	課題	影響	音声	時
		以上	処理	結果	変化	問題
		程度	其処	有効	必要	可能
		何の	或いは	得る	考える	出来る
		分かる	思う	基づく	用いる	無い
模擬講演	抽出率 50%	所	時	人	事	後
		方	自分	今	何	私
		様	其の	もう	矢張り	その
		あー	ん	えー	成る	来る
		行く	居る	為る	有る	言う
		仕舞う	出来る	思う	良い	無い
	抽出率 10%	前	今	次	後	時
		生活	中	一緒	気	自分
		其処	此れ	みたい	先ず	一番
		色々	どう	結構	又	良く
		余り	入る	話す	出来る	就く
		終わる	仕舞う	良い	大きい	多い

4.2.1 学会講演および模擬講演の重要文の特徴

前述の手法を用いて学会講演，模擬講演の重要文に頻出する単語を求めたところ，学会講演では，抽出率 50%，抽出率 10%の場合にそれぞれ，139 単語，463 単語が抽出された．模擬講演では，それぞれ，71 単語，350 単語が抽出された．抽出した単語のうち，スコア $Score_w$ の高い上位 30 単語を表 5 に示す．また，表 6 に抽出した単語に含まれる名詞の割合を示す．抽出した単語の中には名詞以外にも様々な品詞の単語が含まれていることがわかる．2.3 節で述べた重要文抽出手法では名詞以外の情報を用いておらず，この結果は，名詞以外の品詞の情報を用いることで重要文抽出精度が向上する可能性があることを示している．表 6 より，模擬講演では，重要文に頻出する単語に名詞が占める割合が低いことがわかる．これは模擬講演からの重要文抽出では名詞以外の単語を用いることで精度向上を得られる可能性が学会講演よりも高いことを示唆している．

学会講演，模擬講演の重要文において頻出する節を求めたところ，学会講演では，抽出率 50%，抽出率 10%の場合にそれぞれ，18 種類，25 種類が抽出された．模擬講演では，それぞれ，17 種類，29 種類が抽出された．抽出した節のうちスコアの高い上位 5 種類を表 7 に示す．抽出率 50%ではトイウ節，テ節，抽出率 10%ではテモ節，テハ節が重要文となりやすいことがわかる．また，学会講演では連用節，模擬講演では条件節タラが抽出率に関わらず重要文となりやすいことがわかる．このことは，節の情報を用いることで重要文抽出精度の向上が期待できることを示している．ただし，実際に節の情報を重要文抽出に用いるためには，発声が明瞭でなく音声認識が困難な文末を誤らないように音声認識を行うことが必要と考えられる．

表 6: 重要文に頻出する単語のうち名詞が占める割合

	抽出率 50%	抽出率 10%
学会講演	0.453	0.540
模擬講演	0.254	0.429

表 7: 重要文に頻出する節（上位 5）

学会講演	抽出率 50%	トイウ節	連用節	テ節
		引用節	接続詞	
	抽出率 10%	テモ節	テハ節	間接疑問節-助詞
		理由節ノデ	連用節	
模擬講演	抽出率 50%	トイウ節	並列節デ	条件節タラ
		テ節	文末候補	
	抽出率 10%	条件節レバ	テモ節	連用節
		テハ節	条件節タラ	

5 まとめ

講演からの重要文抽出の検討を行った．CSJ の講演から重要文抽出の評価データとして適切な講演を選択して，従来の重要文抽出手法を評価した．自動抽出精度はこれまで考えられていたよりも低く，改善の余地が多く残されていることがわかった．重要文の特徴の分析を行ったところ，重要文と非重要文の判別に単語数が有効であることがわかった．また，重要文に頻出する単語および節について調べたところ，これまで利用されていない名詞以外の品詞や節の情報を用いることで重要文抽出精度の向上が得られる可能性があることがわかった．

参考文献

[1] 南條浩輝, 北出祐, 河原達也. CSJ の講演からの重要文抽出とベイズリスク最小化音声認識. 音講演, 3-1-3, 春季 2006.

[2] 高梨克也, 内元清貴, 丸山岳彦.『日本語話し言葉コーパス』における節単位認定　version1.0. http://www2.kokken.go.jp/csj/public/manuals/clause.pdf.

[3] 野畑周, 内元清貴, 高梨克也, 井佐原均.『日本語話し言葉コーパス』における自由要約・重要文抽出データについて　version1.0. http://www2.kokken.go.jp/csj/public/manuals/summarydata.pdf.

[4] 南條浩輝, 北出祐, 河原達也. 談話標識の統計的選択に基づいた CSJ の講演からの重要文抽出. 信学技報, SP2003-125, NLC2003-62 (SLP-49-13), 2003.

[5] 小椋秀樹, 山口昌也, 西川賢哉, 石塚京子, 木村睦子.『日本語話し言葉コーパス』の形態論情報の概要 ver.1.0.　http://www2.kokken.go.jp/csj/public/manuals/pos.pdf.