マルチメディア議事録生成における概念ベースに基づく会議構造化

別所克人　　大附克年　　廣嶋伸章　　松永昭一　　林良彦
日本電信電話株式会社　NTT サイバースペース研究所

bessho.katsuji@lab.ntt.co.jp

会議音声から話題の階層構造を抽出することにより議事録を生成するシステムを提案する。本システムでの議事録とは、大項目の下に小項目が列挙されているような階層的な話題構成をとるもので、各項目は簡潔な要約文になっているものである。話題構造の抽出は、会議音声の認識結果を単語ベクトルの系列ととらえ、この系列をトピック単位に分割し、得られたトピック単位の集合を階層的にクラスタリングすることによって行う。書き起こしテキストを入力としたときのトピック分割及びクラスタリングの評価実験を行い、その有効性を確認した。

1. はじめに

録音した会議音声を書き起こしたテキスト、あるいは音声認識技術により自動的にテキスト化したものは、会議の内容を把握する上で極めて有用なものであることはいうまでもないが、それらの分量は概して多く、内容を把握するのに時間がかかるため、内容を簡潔にまとめた要約が求められる。

三村の研究[1]においては、入力音声中の小区間に予稿の章題を話題ラベルとして対応付け、入力音声のトピック分割を行う。しかしながら、予稿等の資料が存在しない会議もあり、またトピック区間同士の関係性が得られない。秋田の研究[2]においては、議長発話から談話標識を含むキーフレーズの検出により議論の結論となる発話を特定する。結論部分は必ずしもキーフレーズを含むとは限らず、また結論部分以外にも重要な会議の情報が存在する場合もある。松村の研究[3]においては、入力テキストをトピック分割した後、各トピック区間から、時間的に後のトピック区間で類似度が閾値以上のものへリンクを張っていく。リンク付けられたトピック区間の集合を人間が把握することはできるが、そこでできているトピック区間群内部のより詳細な話題構成、及びトピック区間群同士の類似性を把握することは難しく、よりきめ細かな話題構成を把握することは困難である。

一般に議事録は、会議の内容が項目にまとめられ、階層的に整理されている。議事録作成者は、記憶に残っているもの、会議のときにメモをとったもの（ともに記録しておく必要があると考えた重要事項）を必ず項目にまとめ、階層的に整理しようとする。全ての項目を時系列順に忠実に並べようとはしないし、そもそも会議の模様を時系列に細かく追想するのは困難である。大局的な項目から局所的な項目まで階層的に整理されていることによって、議事録の読者も会議全体の内容を容易に把握することが可能となる。そこで、機械が議事録を作成するにあたっても、話題の集約・階層化が必要である。

我々は上記の要件を満たす議事録を会議音声から自動生成するシステムの構築を行っている。本システムでは会議音声の認識結果をトピックセグメントに分割し（トピックセグメンテーション）、得られたトピックセグメントの集合を階層的にクラスタリングする（セグメントクラスタリング）ことによって、話題のツリー状の階層構造を抽出し、このツリーにおける各ノードから要約文を抽出することによって、項目が階層的に配置された議事録を生成する。

本稿では、2 章において、システムの全体的な枠組みを述べる。3 章で話題構造抽出処理を構成するトピックセグメンテーションとセグメントクラスタリングについて述べ、4 章で書き起こしテキストを入力としたときのそれらの処理の評価実験結果を述べる。5 章でツリーの各ノードから特徴的な単語を抽出することによって、ツリーの概観を示し、6 章でまとめを述べる。

2. 議事録生成システムの概要

本システムの処理概要を図 2.1 に示す。

本システムでは、高精度の音声認識を実現するため、各話者に接話型マイクを割り当て、発声して得られる話者ごとの音声ファイルの集合を入力とする。

音響信号セグメンテーション部では、各音声ファイルを、音声・ノイズ・ポーズのいずれかの種別と判定された音響セグメントに分割する。

音声認識部では、検出された音声区間の音声の認識を行う。これにより、各話者ごとに、認識結果セグメント（ポーズで区切られた音声の認識結果）の列が得られる。

話者は、ある発話意図をもって言葉を発するので、一般に各認識結果セグメントは、それ以上分割し得ない意味の最小単位と考えられる。そこで、音声認識結果マージ部では、全話者の認識結果セグメント集合をマージして、各セグメントの開始時刻でソートする。

ソートされた認識結果セグメントの列から、話題の階層構造を抽出する話題構造抽出処理はトピックセグメンテーション部とセグメントクラスタリング部とからなる。

トピックセグメンテーション部では、ソートされた認識結果セグメントの列をトピック単位（トピックセグメントと呼ぶ）に分割する。トピックセグメントは、同一話者の連続する認識結果セグメントの列（発言区間と呼ぶ）に細分される。

内容の類似する異なるトピックセグメントはある大きなトピックにまとめられると考えられるため、セグメントクラスタリング部では、トピックセグメント集合の階層的なクラスタリングを行う。これにより、各トピックセグメントをリーフノードとする話題のツリー状の階層構造を抽出することができる。各リーフノードの配下には発言区間のノードがあると捉えることができる。

要約部では、ツリー上の各ノードに対し、そのノード配下のテキストから、語句相当・文相当の要約文を抽出する。

以上の処理により、人間が作成するような簡潔な要約文からなる項目が階層的に配置された議事録が生成される。ツリー上、上位ノードが議事録における大項目、下位ノードが小項目に相当することになる。上位ノード群により会議における主要項目を容易に把握でき、下位ノードを読むにつれ、各主要項目の詳細情報を知ることができる。このように会議の話題がトップダウン式に整理され構造化されているので、ユーザは容易にその内容を理解することが可能となる。

ツリー上の各項目を選択することにより、該項目配下の

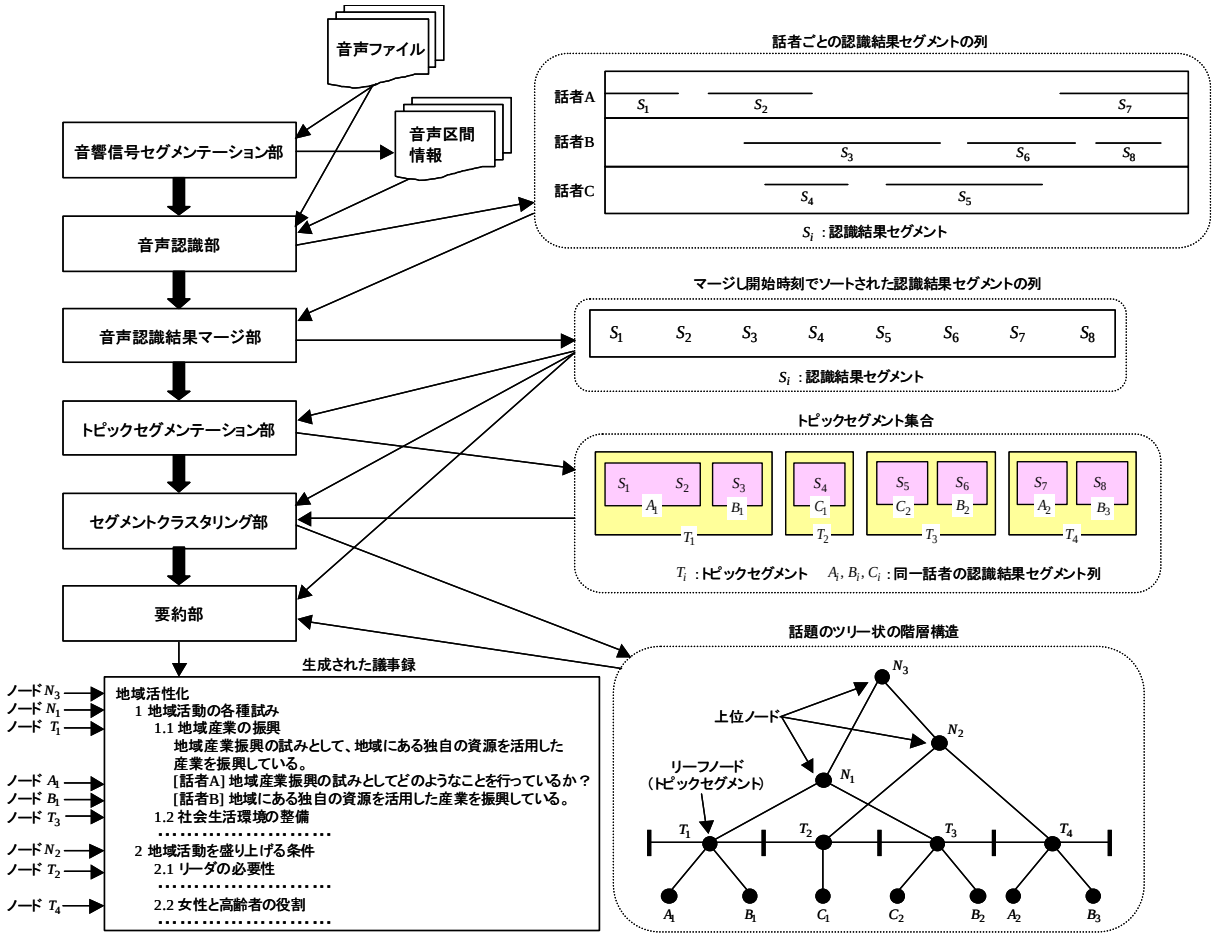

図 2.1

テキストを表示したり、該テキストに対応する時間区間の、収録音声・映像を再生したりする機能をつけることにより、ユーザは興味のある項目の詳細な情報にアクセスすることも可能となる。

3. 話題構造抽出処理

3.1. トピックセグメンテーション処理

トピックセグメンテーション処理は、単語とその意味表現である概念ベクトルの対の集合が格納された概念ベース[4][5]を用いて、クラスタ内変動最小アルゴリズム[6]によって行う。

音声認識処理の結果、各認識結果セグメントは単語の列となっている。認識結果セグメントの各単語を、対応する概念ベクトルに変換する。クラスタ内変動最小アルゴリズムでは、得られた概念ベクトルの系列を分割するクラスタ列で、クラスタ群として最適なものを動的計画法により求め、得られたクラスタ列をトピックセグメント列とする。

3.2. セグメントクラスタリング処理

各トピックセグメントを概念ベクトルの集合である初期クラスタとみて、トピックセグメントの集合をウォード法[7]により階層的にクラスタリングする。

以下、ベクトルを割り当てられた単語に、テキスト中の配列順に番号を振り、クラスタを単語番号の集合とみなす。

ウォード法では、クラスタ A のコストを

$$\mathrm{cost}(A)=\sum_{k\in A}\|v_k-M(A)\|^2$$

と定義する。ここで、v_k は番号 k の単語のベクトル、$M(A)$ は A の重心ベクトルである。またクラスタ群のコストを、構成する各クラスタのコストの和とする。

各初期クラスタに、初期クラスタ群のコストを、ツリーにおけるレベル値として割り当てる。

異なるクラスタを結合したときのコストの増分が最小となるクラスタの対を求める。このようなクラスタ対を結合して得られるクラスタを、対を構成する各クラスタの上位クラスタとする。この時点でのクラスタ群のコストを、新しいクラスタの、ツリーにおけるレベル値として割り当てる。最終的にクラスタが一つになるまで結合を行っていく。

このようにして、図 3.1 のような各クラスタ C_i をノードとする 2 分木が得られ、各ノードがそのレベル値（$nodeLevel(i)$ と表す）に位置付けられる。2 分木では階層数が非常に多くなるので、以下のアルゴリズムにより階層数を制限したツリーに変形する。

【アルゴリズム開始】

(1) ルートノード C_g の $nodeLevel(g)$ （$=e_1$ とする）とリーフノード C_i の $nodeLevel(i)$ （$=e_0$ とする）を端点とする区間を、指定した数 p で等分する。以下、等分点とは、等分して得られた点と端点を意味するものとする。

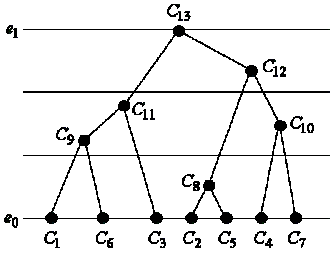

図 3.1

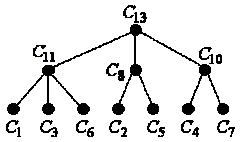

図 3.2

(2)ルートノードC_gを引数として関数Fを呼び出す。

関数F ：［引数のノードC_jがリーフなら終了する。C_jがリーフでないならば、$nodeLevel(j)$未満の等分点の最大値mを求める（$nodeLevel(j)=e_0$ならば$m=e_0$とする）。C_jを子ノードへ展開する。展開先ノードの$nodeLevel$がmより大きい限り展開先ノードを展開する。このようにして、$nodeLevel$がm以下となるノード群$\{C_{t_1},C_{t_2},\cdots,C_{t_q}\}$が得られる。$C_j$の新しい子を、ノード群$\{C_{t_1},C_{t_2},\cdots,C_{t_q}\}$の各ノードとする。各$C_{t_r}$の新しい親ノードを$C_j$とする。各$C_{t_r}$ $(1\le r\le q)$を引数として関数Fを再帰呼び出しする。］

【アルゴリズム終了】

上記のアルゴリズムにより、図 3.1 のツリーは、$p=3$のとき、図 3.2 のツリーに変形される。

4. 評価実験

4.1. 評価データ

認識誤りのない条件での話題構造抽出処理の精度をみるため、会議音声を人手で書き起こしたテキストに対するトピックセグメンテーションとセグメントクラスタリングの評価実験を行った。使用したデータは、情報通信審議会の書き起こしテキストと、構造改革についてのテレビ討論番組の書き起こしテキストである。本評価においては、句点で区切られた一文が一認識結果セグメントに相当すると考える。各データに対し、正解のトピック分割結果と話題の階層構造を人手で作成した。但し、テレビ討論番組の方は 7 トピックであり、話題の階層構造が、ルート直下に各トピックセグメントがあるような単純なものであるため、トピックセグメンテーションの評価のみ行った。

各データの正解のトピック分割結果、話題の階層構造に関する各種値は表 4.1 のとおりである。ベースラインとは、セグメンテーション精度のベースラインのことであり、文間の境界をランダムに選んだとき、それがトピック境界である確率であり、小さいほど分割が難しいことを意味する。ツリーのノード数には、リーフノード配下の発言区間ノードは含めていず、精度評価にあたり発言区間ノードは考慮しないものとする。

概念ベースについては、毎日新聞 2000 年版の記事 1 年分を基に、単語間の共起頻度をカウントした行列を特異値分解することにより得られる単語数 30000、ベクトルの次元数 750 の概念ベースを使用した。

4.2. トピックセグメンテーション精度

各データに対し、クラスタ内変動最小アルゴリズムにより、正解のトピック境界数分だけ、トピック境界を出力することによって得られた精度を表 4.2 に示す。出力結果に対し、精度は以下のように算出される。

表 4.1

	情報通信審議会	テレビ討論番組
文数	406	213
トピック数	67	7
ベースライン	16.3%	2.8%
ノード数	92	8
階層数	6	2

再現率＝正解の出力境界数/正解境界数

適合率＝正解の出力境界数/出力境界数

F値$=(2\times$再現率$\times$適合率$)/($再現率$+$適合率$)$

±0 文、±1 文、±2 文とは正解とする正解トピック境界からの範囲を意味する。

表 4.2

データ	正解範囲	再現率	適合率	F 値
情報通信審議会	±0 文	43.9%	43.9%	43.9%
	±1 文	72.7%	65.2%	68.7%
	±2 文	89.4%	75.8%	82.0%
テレビ討論番組	±0 文	66.7%	66.7%	66.7%
	±1 文	83.3%	83.3%	83.3%
	±2 文	83.3%	83.3%	83.3%

情報通信審議会の方は、全体的にトピックが似通っていることや、正解のトピック境界自体に 1～2 文程度の主観によるずれが考えられることから、±0 文の精度は 40%台になっている。しかしながら、±1～2 文のずれを許容するなら、約 70～80%の精度をもつ。

一方、テレビ討論番組の方は、司会者があらかじめシナリオに沿ってテーマを進行していくことから、トピックの切り替わりがある程度明確という特徴があり、ベースラインは情報通信審議会の方より低いにも関わらず、±0 文の場合でも 66.7%の精度をもつ。

なお、文献[6]では、コスト比の閾値を設定することにより正解トピック境界数を推定する方法を提案しているが、会議録のようなデータに対し適切な閾値を設定することは難しく、適切な正解トピック境界数推定は今後の課題である。

4.3. セグメントクラスタリング精度

情報通信審議会のデータに対し、初期クラスタ集合を正解のトピックセグメント集合とした場合と、トピックセグメンテーション結果におけるトピックセグメント集合とした場合のそれぞれに対し、3.2 節のアルゴリズムにおいて$p=5$としてクラスタリングを行い、精度を測定した。

A：正解クラスタ集合、B：出力クラスタ集合とする。$F(a,b)$ $(a\in A,b\in B)$ を、a,b のF 値とする。評価指標を以下のように定める。

・正解クラスタ実現率$=\frac{1}{|A|}\sum_{a\in A}\max_{b\in B}F(a,b)$

・出力クラスタ正解率$=\frac{1}{|B|}\sum_{b\in B}\max_{a\in A}F(a,b)$

●初期クラスタ集合を正解のトピックセグメント集合とした場合の精度は以下のようになった。正解クラスタ実現率は、Aからルート及びリーフのクラスタを除いて、また出力クラスタ正解率は、Bからルート及びリーフのクラスタを除いて算出している。

○クラスタの要素をトピックセグメントとして F 値を算出した場合
・正解クラスタ実現率：69.7%
・出力クラスタ正解率：60.6%
○クラスタの要素を文として F 値を算出した場合
・正解クラスタ実現率：79.1%
・出力クラスタ正解率：72.0%
●初期クラスタ集合をトピックセグメンテーション結果におけるトピックセグメント集合とした場合の精度は以下のようになった。正解クラスタ実現率は、A からルートのクラスタを除いて、また出力クラスタ正解率は、B からルートのクラスタを除いて算出している。
○クラスタの要素を文として F 値を算出
・正解クラスタ実現率：69.7%
・出力クラスタ正解率：65.9%

セグメンテーション誤りのない純粋なクラスタリング精度は文集合としては、約 70～80%の精度となっている。トピックセグメンテーション結果を用いた場合でも約70%の正解クラスタ実現率をもつ。

4.4. 話題構造抽出処理の構成の検証

セグメントクラスタリングが同一トピックの文をクラスタ化する処理ならば、トピックセグメンテーション処理を行わず、各文を 1 トピックセグメントとした上で、セグメントクラスタリングを行うことも考えられる。

各文を初期クラスタとした場合のクラスタリングの精度は、正解クラスタ実現率：54.2%、出力クラスタ正解率：28.7%であり、4.3 節の結果よりも低い。これは、異なるトピックに属する文で、類似性の高いものが同一クラスタに誤って分類されることがあるためである。一次元の配列という制約下でトピックセグメンテーションを行い、その結果得られたトピックセグメントをクラスタリングするという構成の方が適切だといえる。

5. 特徴単語の抽出によるツリーの概観

本システムでは、要約部によりツリー上の各ノードの要約文を抽出するが、本稿では、要約部の処理によらず、各ノードに特徴的な単語を抽出することにより、ツリーの概観を示す。

ツリー上の各ノードに対し、該ノードに含まれる異なり単語を意味的に重要な順にソートする。該ノードにおいて中心的で、兄弟ノードからは離れているような異なり単語が、該ノードを代表し、かつ兄弟ノードから差異化できるものとして、重要性が高いとする。

ノードに対応するクラスタを C_h、兄弟ノードに対応するクラスタの集合を $\{C_{i_{1h}}, C_{i_{2h}}, \cdots, C_{i_{n_{hh}}}\}$ とし、

$D = C_{i_{1h}} \cup C_{i_{2h}} \cup \cdots \cup C_{i_{n_{hh}}}$ とおく。

クラスタ C_h に含まれる異なり単語 w の概念ベクトルを v_w と表したとき、w に対し、以下の $den(w)$、$num(w)$、$score(w)$ を計算する。

$$den(w) = \sum_{k \in C_h} \| v_k - v_w \|^2$$

$$num(w) = \begin{cases} 0 & D = \phi \text{のとき} \\ \sum_{k \in D} \| v_k - v_w \|^2 & D \neq \phi \text{のとき} \end{cases}$$

$$score(w) = \frac{num(w)}{den(w)}$$

異なり単語の集合を、以下の規則により降順にソートする。

- $den = 0$ と $den > 0$ なら、$den = 0$ の方を大とする。
- $den = 0$ 同士なら、num の値の大きい方を大とする。
- $den > 0$ 同士で、共に $num = 0$ なら、den の値の小さい方を大とする。
- $den > 0$ 同士で、少なくとも一方が $num > 0$ なら、num/den の大きい方を大とする。

指定した数の上位の単語を、特徴単語として図示する。

図示する特徴単語数を 5 として、情報通信審議会のトピックセグメンテーション結果をクラスタリングして得られたツリーの一部を、正解のツリーの一部とともに、表 5.1 に示す。各ノードは単語の羅列であるため、内容の把握は困難であるが、正解ツリーの項目に関係した単語が抽出されている傾向があることが分かる。

表 5.1

正解のツリー	話題構造抽出結果のツリー
情報通信審議会 ・国際競争力 ・通信主権 ・国の安全を損なう恐れのある外国投資の制限 ・通信主権について欧米との整合性 ・研究開発 ・研究開発体制の在り方 ・進捗についての質問 ・競争政策 ・公正で透明な市場環境の整備 ・消費者保護 ・消費者保護という言葉の問題 ・消費者保護と自立 ・消費者対応 ・消費者の混乱 ・接続事業者の件	従来, 競争, 既存, 体系, 範囲 ・競争, 競争促進, 緩和, 競争力, 外資 ・感じ, 欧米, 誤解, 風, 外資 ・事業者, 通信, 規制, 融合, 区分 ・外資, 資本, 競争, 市場, 側面 ・主権, 責務, 対処, 国家, 体制 ・研究, 基礎, 開発, 研究所, 産学 ・革命, ＩＴ, 情報, 委員会, 部会 ・資料, 目的, 我が国, 開発, 体制 ・苦情, 窓口, 保護, 相談, マニュアル ・消費者, 意味, 言葉, 要素, 環境 ・政策, 消費者, 競争, 業者, 利益 ・言葉, 意味, 政策, 所々, 保護 ・殺到, 電話, メール, 件数, 注文 ・ネット, インターネット, 電話, 接続, 情報 ・当事者, 話, 総務省, 風, 対応

6. 終わりに

本稿では、会議音声から、内容が項目にまとめられ階層的に整理された議事録を自動生成する方式について述べ、書き起こしテキストを入力とした場合の評価実験結果を報告した。今後は音声認識結果を入力とした場合の、認識誤りに頑健な話題構造抽出処理の検討を進めていく予定である。

参考文献

[1]三村正人，河原達也，堂下修司： パネル討論音声の話者と話題に関する自動インデキシングの検討，情報処理学会研究報告，Vol.SIG-SLP 011，pp.13-18(1996).
[2]秋田祐哉，河原達也： 会議音声の自動アーカイブ化システム，情報処理学会研究報告，Vol.SIG-SLP 034，pp.61-66(2000).
[3]松村真宏，加藤 優，大澤幸生，石塚 満: 議論構造の可視化による論点の発見と理解，日本ファジィ学会誌，Vol.15，No.5，pp.554-564 (2003).
[4]Kato, T., Shimada, S., Kumamoto, M. and Matsuzawa, K.: Idea-Deriving Information Retrieval System, Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp.187-193(1999).
[5]熊本 睦，島田茂夫，加藤恒昭： 概念ベースの情報検索への適用―概念ベースを用いた検索の特性評価，情報処理学会研究報告，Vol.SIG-ICS 115，pp.9-16(1999).
[6]別所克人： クラスター内変動最小アルゴリズムに基づくトピックセグメンテーション，情報処理学会研究報告，Vol.SIG-NL 154，pp.9-16(1999).
[7]宮本定明： クラスター分析入門―ファジィクラスタリングの理論と応用，森北出版(1999).