社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－HI－119 (13)
2006／7／7

非言語情報を利用した会話シーンの抽出と意味的インデキシング

角 康之, 熊谷 賢, 瀬戸口 久雄, 西田 豊明
sumi@i.kyoto-u.ac.jp
京都大学 情報学研究科

話し手と聞き手がポスターを参照しながら会話する状況を想定し、会話シーンを適切なサイズに切り出し、意味的なインデキシングを自動的に行うための手法を提案する。提案手法は、音声認識などによる発話内容の言語的理解を利用せず、身振り手振り、ポスターの指さし、視線といった非言語情報を利用して会話シーンを抽出し、さらに、それらのシーンがどのトピックについてなされた会話なのか、また、盛り上がった会話だったのか、といった意味づけを推測する。

Extraction and Semantic Indexing of Conversational Scenes Using Nonverbal Information

Yasuyuki Sumi, Ken Kumagai, Hisao Setoguchi, Toyoaki Nishida
sumi@i.kyoto-u.ac.jp
Graduate School of Informatics, Kyoto University

In this paper, we show our attempts to extract and semantically index tractable-sized conversational scenes from continuous conversations between two persons along with a poster as a reference object. Our aproach trys to extract conversational scenes and infer topics and types of the conversation by nonverbal information such as gesture, pointing the poster, and gazing, not verbally understanding by speech recognition.

1 はじめに

会話は、知識伝達及び創造の最も日常的で効果的な手段のひとつである。ミーティング、小規模レクチャ、オフィスの片隅や廊下での何気ないおしゃべりといったように、我々は日常的に会話を行っている。従って我々は、知識の伝達と創造を強化するための一アプローチとして、会話を効率良く記録し、再利用性を高めるための技術を構築したいと考えている。

そのための重要な課題は、一連の会話から適切なサイズのシーンを切り出すこととと、それらに意味的なインデクスを与えることであり、これが本稿で紹介する研究の目的となる。話している内容は何なのか、何について話しているか、といったことを音声理解、映像理解によって内容的に解釈できることが望ましいが、現在の技術は、それを日常的な会話の現場で利用できるほど成熟していない。そこで、本稿では、会話の周辺で発生している非言語的な情報に着目して、会話シーンの抽出や意味的なインデキシングを行う手法を提案する。

これまでにも、発話内容の意味解釈に立ち入らずに、発話の周辺状況、特に、視線（注目対象）のセンシングによりミーティングのハイライトを推定しようとする試みなどがあった [1]。我々も、展示見学中の注視対象と発話行為の組み合わせから基本的なインタラクションの意味づけを解釈し、それによって、ある程度、展示見学者のハイライトシーンを見つけることが可能であることを示してきた [2]。本稿は、会話シーンの抽出や意味的なインデキシングについて、視線以外の非言語情報についても理解を深めることが目的である。

図 1 に、本稿で想定した会話場のモデルを示す。会話場は、会話者以外に、参照オブジェクト、つまり手持ちのノート、ホワイトボード、プレゼンテーションのスライド、話題の対象となる物等によって

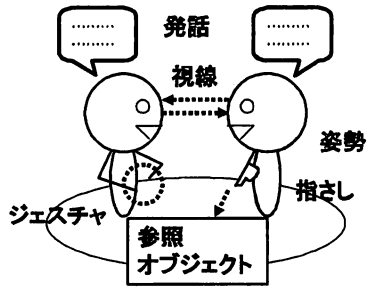

図 1: 会話者と参照オブジェクトで構成される会話場のモデル

構成されることが多い。会話を記録・再利用・創造するメディアを開発することを考えると、参照オブジェクトは、そのもの自体が重要なコンテンツになりうるし、また、参照オブジェクトへの参照行為が重要なインデクスになることが多い。したがって、我々は、図1に示したような、会話者と参照オブジェクトで構成された会話場に着目する。

図1には、本稿で扱いたい非言語情報を書き込んである。つまり、発話行為周辺で起きている、注視、姿勢の変化、指さし行為、手で表現されるジェスチャに注目する。ジェスチャについては、喜多らの分類 [3] で言うところの表象的 (representational) ジェスチャを扱うこととする。表象的ジェスチャとは、大きく分けて、描写的なジェスチャと指さし行為を含む。

- 描写的 (depicting) ジェスチャ
  - 映像的 (iconic) ジェスチャ：空間的、運動的な対象の類似的表現
  - 暗喩的 (metaphoric) ジェスチャ：上記の比喩的適用
- 直示的 (deictic) ジェスチャ：指さしなど

本稿では、こういった非言語情報を利用した会話シーンの抽出と意味的なインデキシングについて、我々の2つの試みを紹介する。一つ目は、ポスター発表を対象として設定し、参照オブジェクトとしてのポスターへのタッチ行為のみを扱うことで、会話シーンの切り出しや意味的分類を行った試みである。そこでは、タッチパネルという特殊なデバイスを利用することで状況を特殊化する替わりに、簡単な仕組みで効果的に会話シーンを切り出したり、それを再利用できることを示す。二つ目は、一般的な会話場を想定し、会話者の視線や指さし、描写的ジェスチャを網羅的にデータ化し、それらによってどれだけ会話の質的評価に迫れるかを分析した試みである。そこでは、発話量、ジェスチャの発生、参照オブジェクトに対する共同注視をセンシングすることで、ある程度、会話のタイプを識別できる可能性を示す。

2 ポスタータッチを手掛かりにした会話シーンの切り出しと分類

2.1 分身プレゼンテーションシステム

最初に紹介するのは、分身プレゼンテーションシステムである [4, 5]。ここで想定しているのは、展示や学会におけるポスタープレゼンテーションである。つまり、参照オブジェクトとしてのポスターをはさんで、説明員と来訪者が会話をするような状況である。そう言った場合、説明員は、入れ替わり立ち替わり訪れる来訪者に対して、基本的には同じトピックスについて繰り返し説明を行う。したがって、何人かの来訪者との会話を記録し、効率的にインデキシングすれば、その会話シーンを再利用することで、説明員の分身エージェントを構築することができるのではないか、という発想である。

図2にシステムの概念図を示す。まず、記録された体験データから、説明員と来訪者の間の会話シーンに構造を与える。その際、発話交代の情報や展示パネルのタッチ情報を利用して、会話内容の詳細な分割とその内容の推定を行う。分割された会話シーンは、大きく分けて、説明員による説明シーンと、説明員と来訪者の間でなされた議論シーンに分類される。そして、説明員が不在中に訪れた新しいユーザに対して、説明員と来訪者のキャラクタアイコンがポスター画面上に現れ、自動で説明シーンの音声を提示する。ユーザは、興味のあるエリアをタッチすることで緩く分身エージェントとインタラクションすることができる。すなわち、タッチされたエリアに連想づけられた議論シーンを再現することで、システムユーザは過去の来訪者と説明員との間でなされた会話を、仮想的に追体験することが可能になる。

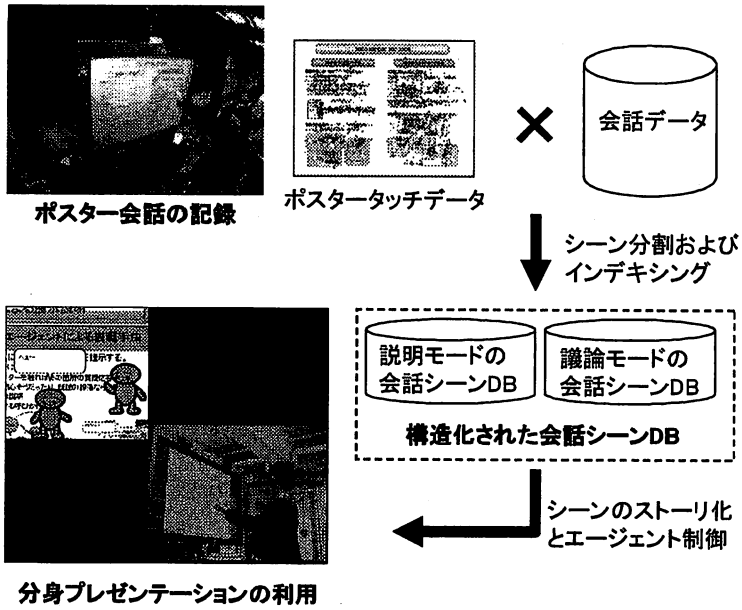

図 2: 分身プレゼンテーションシステムの概念図

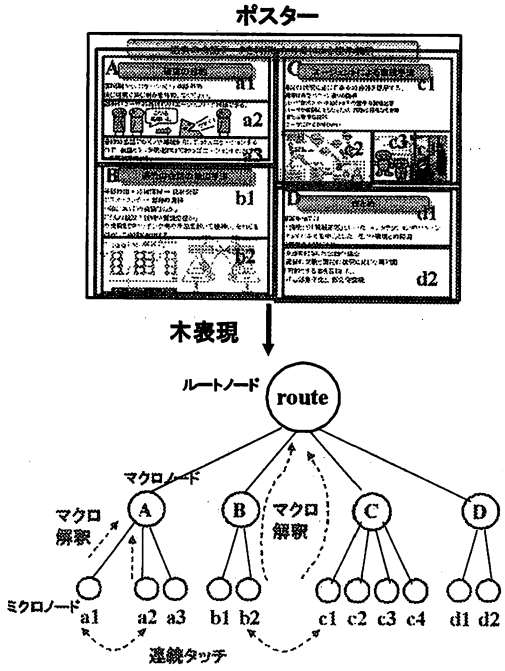

図 3: ポスター構造の木表現

2.2 背景知識としてのポスター構造

本システムの特徴は、一連の発話群が何に対する会話だったのかを推定する手掛かりとして、ポスタータッチの場所を利用することである。つまり、必ずしも発話内容を言語的に理解しなくても、ポスタータッチのデータを利用し、「発話周辺で指さしされたエリアについて会話されているであろう」と推定するのである。ポスタータッチの場所は、タッチパネルを利用すれば容易に取得することができる。また、ポスターデータは電子的なデータなのだから、あらかじめ緩い構造や内容を取得しておくことは容易である。今回我々は、あまり内容に立ち入らず、空間的な構造のみでどこまで分身プレゼンテーションが可能になるか、挑戦した。すなわち、ポスターのエリアの階層的な構造に着目し、それをあらかじめ木構造で表現しておくこととした。

図 3 はポスターの意味構造を木表現に変換している例である。PowerPoint などにおけるオブジェクトのグルーピング情報などを利用すれば、このような木表現は自動的に取得できると考えられるが、今回は手作業で以下のようなエリア分割を行った。

ミクロ領域 単一の意味的な塊からなる領域

マクロ領域 互いに関連性のある複数のミクロ領域

で構成される領域

ルート領域 互いに関連性のある複数のマクロ領域で構成される領域

次にポスターの意味的な階層構造を木で表現し、木のノードと各領域を対応付ける。木表現に基づいてタッチデータを最小木解釈し、解釈されたタッチデータに対応するノードと会話シーンを結びつけることで、会話シーンに階層的なインデクスを付与することが可能となる。

2.3 会話シーンの分割とポスターエリアとの対応付け

次に、ポスター指差し行為データを用いて、ポスター会話を適当なサイズのシーンに切り出し、それぞれをポスター領域と関連づける方法について述べる。その際、我々は以下のような仮説を用いた。

仮説：ポスター会話は、対応する指差し行為データとの時間間隔が離れるほど、その会話内容と指差し行為データで示された内容との結びつきが弱くなる

以下、図4に基づいて、ポスター会話を分割し、ポスター領域と関連づける過程を説明していく。始めに、発話データに対し、クラスタリング、ノイズ除去を行い、発話クラスタを生成する。発話クラスタには、開始時間、終了時間の属性が存在し、各発話クラスタを以下の3つの場合に分けて処理する。

場合1 発話クラスタの時間内に、指差し行為データが一つ、もしくは一種類ならば、発話クラスタをその指差しエリアと対応づける。

場合2 発話クラスタの時間内に、指差し行為データが二種類以上存在する場合、2つ目以降に発生した指差し行為データごとに発話クラスタを分割する。分割後、先頭の発話クラスタは時間内の指差しエリアに対応付け、それ以外の発話クラスタは、時間的に直前の指差しエリアに対応付ける。

場合3 発話クラスタの時間内に、指差し行為データが存在しない場合、発話クラスタの開始時間前、終了時間後で、最も時間的近傍にある指差しエリアと対応付ける。

場合2は、展示員が説明を行なっているシーンに多い場合で、この場合、比較的時間幅の大きい発話クラスタが形成される。したがって適切なサイズに切り分けるために、指差し行為データに注目することで、1つの指差し行為データ毎に対応したポスターエリアに対応付けられるように発話クラスタを分割する（図4の場合2)。なお、そうした説明シーンにおいては、多くの場合、指差し行為を起点として会話の対象が切り替わるため、時間的に直前の指差し行為データに各クラスタを結びつけている。

場合3は、会話と同時にポスター指差し行為が行なわれていない時に生じる現象で、この場合、時間的近傍の指差しエリアを発話クラスタに対応付ける（図4の場合3)。しかし、前方と後方それぞれの時間幅に同じ重みを与えることが適切かどうかは自明ではなく、s,t 間の関係に対して、今後さらに考察が必要である。

以上のような準備を経て、蓄積された会話シーンが、ポスターの各エリアに連想づけられる。また、ポスターの各エリアが図3に示したような階層構造を持っていることを考慮して、会話シーンの間に緩い構造を想定することが可能になる。システムユーザがシステム操作をしない限り、システムは図3にあるノードを左側から走査し、それらに対応づけられた説明シーンを再生する。ユーザは任意のタイミングでポスタータッチをすることで、興味のあるエリアを指し示すことができる。すると、説明モードに割り込みが入り、タッチされた点を包含するエリアに対応づけられた議論シーンを再生し、それが終わったら、割り込み時点の説明モードに戻る。したがって、システムユーザは任意の質問をできるわけではないが、少なくとも興味のあるエリアについての質疑応答を聞くことができ、過去の他人の会話から、潜在的な興味を満たすことが可能になると期待している。

3 非言語情報を利用した会話シーンの意味的インデキシング

3.1 非言語情報による会話の内容と質の推定

前節で紹介した分身プレゼンテーションシステムは、様々な非言語情報のうち、特に直示的ジェスチャ（指さし）にのみ着目し、それを明示的に観測するためのポスタータッチデータを利用して、会話シー

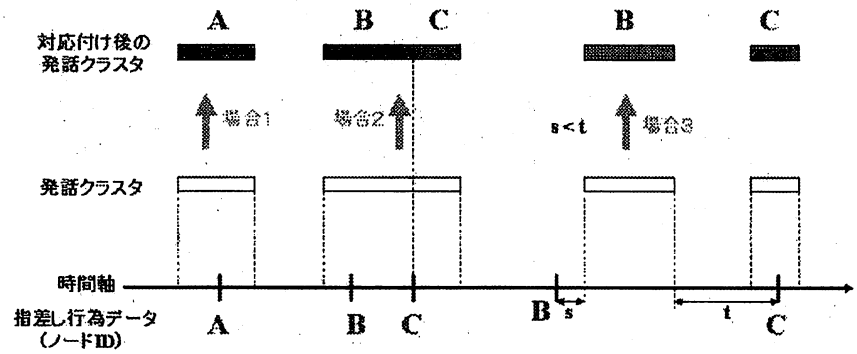

図 4: 会話シーンの分割

ンの切り出しと分類を試みたものであった。試作したシステムは、単純な仕組みであるにも関わらず、会話シーンを適切なサイズに切り出したり、それをポスター上のエリアに連想的に貼り付けることを可能にし、ポスターの前でなされた会話を仮想的に追体験するための枠組みとしては、興味深い方向を示すことができた。そこで我々は、新たな課題として以下の 2 つに興味を持った。

- 会話周辺で発生している状況情報について、ポスタータッチデータだけでなく、もっと多様な非言語情報を扱いたい。
- 会話データが増えてくると、ポスターの同一エリアに対して複数の会話シーンが連想づけられるので、会話再生の際にシーンを選択するための、会話シーンの「質」や「タイプ」を表す指標が欲しい。

これらの課題に応えるために、我々は、会話の周辺で発生している身振り手振り、視線を網羅的に記録したデータを用い、それら非言語情報が会話シーンの内容や質を推定するのにどの程度利用できるのかを、分析してみることとした。

ここで言う「会話シーンの内容や質」というのは、例えば、盛り上がった会話だった；本来の内容から脱線した会話だった；聞き手にとっては退屈な会話だった、といったタイプ分けを想定している。もしも、言語処理などによる発話内容に踏み込んだ処理をせずに、身振り手振りなどの非言語情報から上記のような会話のタイプが推定できるとすると、記録された会話シーンの利用価値は格段に向上する。例えば、前節で紹介した分身プレゼンテーションの例で言えば、短くて簡潔なシーンを選択すべきか、それとも、少々長めでも盛り上がったシーンを選択して提示するべきかを、ユーザの好みに合わせて判定する、といった細やかなサービスが可能になる。

3.2 非言語情報を含んだ会話データの準備

本分析で利用したデータは、ATR メディア情報科学研究所における実験 [6] で取得されたデータである。そのデータは、「説明者」と「聞き手」とにあらかじめ役割分担された二人の実験参加者が、実験室内に設置された展示室に掲示された仙台の観光情報についてのポスターを共に参照しながら会話を行う様子を記録したものである。実験は、説明者がポスターに書かれているテーマに沿って聞き手に観光情報を説明していくというものであり、会話に参加している二人の身体動作や位置はモーションキャプチャシステムにより記録され、視線情報はアイマークレコーダによる記録された。

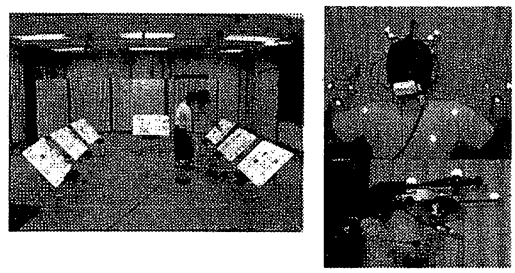

図 5: ポスター会話における身振り手振りと視線の記録

図 5 に実験の様子を示す。参加者およびポスターの三次元位置の記録には、Vicon Motion Capture

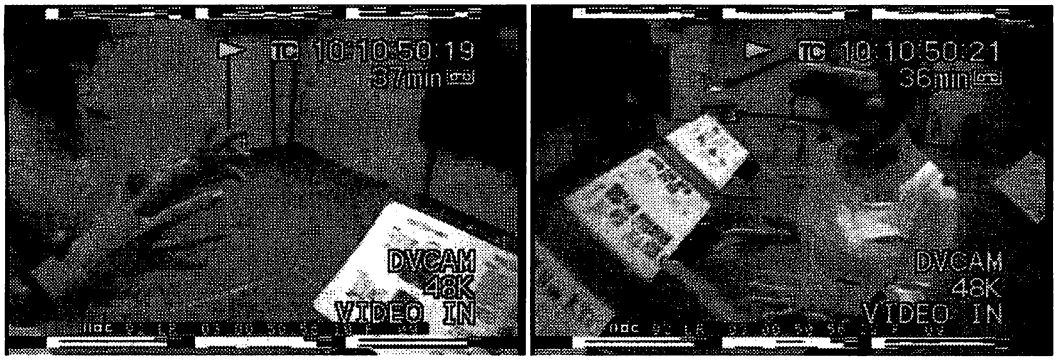

図 6: 描写的ジェスチャが同調している例

System を用いた。身体姿勢や動作計測用のマーカは、人物の頭部 4 点、肩部 3 点、腕部 3 点（両腕）の、計 13 点に装着した。Vicon は、各所に取り付けたマーカを 60Hz の時間分解能、約 1mm の空間分解能で記録する。視線記録にはナックイメージテクノロジー社製の EMR-8B を用い、30Hz の時間分解能と約 0.15 度の分解能で記録した。

説明員はガイドを職業とする同一人物が担当し、学生アルバイトの被験者 19 人と順々にペアを作り、観光案内を繰り返してもらった。実際には図 5 にあるように複数のポスターがあり、順々にポスターを移動しながら観光案内の会話をしてもらったが、本稿では 1 枚のポスターの前での会話のみを分析対象とした。

ここで我々が興味を持っているのは、非言語情報（身振り手振り、視線、発話量）と会話内容の関係を見いだすことである。ここでは、会話シーンは既に一つのトピックについて適切なサイズで切り出されていることを前提とし、その会話シーンのタイプを、非言語情報から推定することができるかどうかを確認したい。したがって、会話シーンのビデオを閲覧しながら手作業で切り出しを行った。非言語情報（身振り、手振り、視線、発話量）はできる限りセンサデータから自動的に抽出することを試みたが、それでもエラーが少なくなかったので、最終的にはビデオを見ながら確認した。

今回注目した非言語情報は、発話量、視線、ジェスチャである。ジェスチャは、直示的ジェスチャ（指さし）と描写的ジェスチャ（ものの形状や動きを手振りで示すジェスチャ）に着目した。発話量は各自がつけている接話マイクのパワー検出を利用したので、比較的容易に定量化できる。自動抽出が難しいのは視線とジェスチャである。我々が欲しいのは、視線や指さしが向かう対象が何なのか、また、描写的ジェスチャはどういったタイミングで発生するのか、といったことである。

図 6 は、今回我々が最も注目している描写的ジェスチャが発生しているシーンである。聞き手が、ガイド情報に出てきた「南部鉄器」の大きさを手振りを交えながら尋ねているのに対し（右）、説明者が形や大きさを表現している（左）。この例では、2 人の間で同一のジェスチャが同調しており、一般的にこういう状況は、聞き手が会話に積極的に参加しており、再利用性の高い（記録に残すに値する）会話シーンであると考えられる。

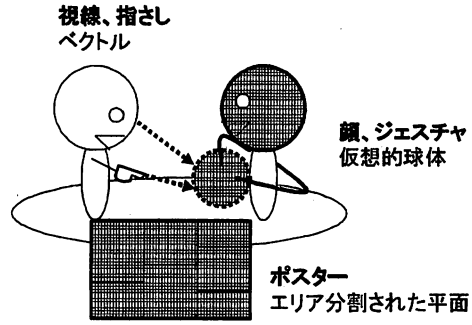

図 7: 視線・ジェスチャ認識のための近似モデル

非言語行為発生に関する定量的データの構築にあたって、最終的にはビデオを見ながら確認するとは言え、大量のビデオを見ながらすべてを手作業で行うことは現実的ではない。したがって、モーションキャプチャシステムとアイマークレコーダデータのセンサデータを利用して、非言語行為が発生するタイミングを自動的・網羅的に見つけることとした。そ

のために、以下のような近似モデルを導入した（図7参照）。

- 視線と指さしをベクトルで表現した。
- 顔は球体で近似した。
- ポスターは、トピックに対応した複数のエリアに分割された平面として表現した。
- 描写的ジェスチャは、両手先を含む仮想的な球体で近似した。

このモデルに基づき、視線や指さしが、顔、ポスター、相手のジェスチャに向いた瞬間を網羅的に数え上げることとした。ただし、指さしの発生は、腕がある閾値以上の角度で体から離れたときのみ発生するものと仮定し、描写的ジェスチャは、ある閾値以上に肘が曲がり、かつ、両手先がある閾値以内に接近した場合にのみ発生しているものと仮定した。

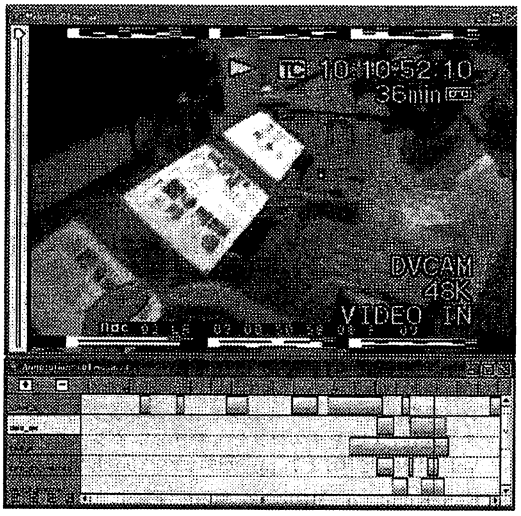

図 8: 描写的ジェスチャが共起している部分を Anvil によって表示した例

次に、自動抽出された非言語行為のデータをビデオと照合しながら確認し、間違いのある部分を手作業で修正した。その作業には、ビデオ・アノテーションのツールである Anvil[7] を利用した。Anvil は、図8にあるように、ビデオを閲覧しながら、時間幅のあるアノテーションを付与・編集することができるツールである。我々は、前述の方法で自動抽出された非言語行為データをアノテーションとして Anvil に読み込み、ビデオで確認しながら、間違いを修正し、データを精緻化した。精緻化されたアノテーションデータは、再び数値データとして出力され、それを非言語情報の分析に用いた。

3.3 非言語情報による会話シーンの特徴付け

分析については、まだ試行錯誤中であり、確定的なことは言えないが、ここでは予備的な分析において見ることができたことを紹介する。我々が現時点で持っているデータは、会話シーンごと、二人の会話参加者ごとに、

- 発話量
- 描写的ジェスチャの発生回数
- 視線（が何かに向かった瞬間）と指さしが発生した回数と、それが向けられた対象（ポスターか、相手の顔か、相手の描写的ジェスチャ）

で構成された多変量データである。我々が注目したのは、二人の視線やジェスチャ（指さしと描写的ジェスチャ）の発生の共起性である。なぜなら、我々が興味があるのは、二人によって構成される会話場の質を知ることなので、個人個人の発話行為、非言語行為の単独の発生よりも、二人によるものの共起性が重要であると考えた。

そこで、二人それぞれについて、向けられた対象ごとの視線と指さしのすべての組み合わせについて共起回数を数え上げ、二次的なデータを作成した。つまり、1ペアごとに複数の会話シーンがあるわけだが、それぞれについて、例えば、二人が同時にポスターを見た回数；説明員が作った描写的ジェスチャに聞き手が目を向けた回数；説明員が指さしたポスターに聞き手が目を向けた回数、といった数値が埋まった多変量データを作成した。そして、そのデータを主成分分析したところ、第3主成分までで累計寄与率が80%を超え、それぞれの寄与率は、第一主成分：約45%、第二主成分：約25%、第三主成分：約10%であった。

これらの主成分が、複数の会話シーンのタイプを特徴づける指標となる。そこで、各主成分について、因子負荷の高い属性を見てみたところ、以下の通りとなった。まず、第一主成分については、ポスターに対する二人の視線の共起（共同注視）、説明員による描写的ジェスチャの発生、説明員による発話量、

聞き手による発話量といった属性の影響が大きかった。このことはつまり、第一主成分の値が大きい会話シーンは、二人が参照オブジェクトやジェスチャを活用しながら、バランス良く発話していることを示しているので、第一主成分の意味は、「会話の盛り上がり」を示す指標であると解釈できそうである。

次に第二主成分を見てみると、ポスターへの共同注視の影響は大変強いが、説明員と聞き手の発話量と描写的ジェスチャの発生については比較的強い負の影響が確認された。このことはつまり、第二主成分の意味は、「説明の簡潔さ」を示す指標であると解釈できそうである。

第三主成分については、共同注視の影響は全く無く、説明員による発話や描写的ジェスチャ発生回数と比較的強い負の影響が確認された。しかし、それだけでは第三主成分の意味を解釈することはできなかった。したがって、第三主成分の点数の高い会話シーンを実際に閲覧してみたところ、説明員が用意している標準的な話題以外の話題、例えば、南部鉄器のサイズや利用される場面に関する話題が加えられているシーンが多いことに気づいた。そこで、第三主成分を「新情報の追加」と意味解釈してみたい。

ここで得られた知見は、比較的我々の直感に沿ったものである。そして本研究の価値は、そういった直感を、非言語情報の具体的な数値データに帰属させる可能性を示したことである。つまり、将来、こういった非言語情報を安価で日常的なデバイスでセンシングできるようになったときには、内容に踏み込んだ処理を解することなく、会話シーンの意味理解を近似することが可能になる。

4 おわりに

話し手と聞き手がポスターを参照しながら会話する状況を想定し、会話シーンを適切なサイズに切り出し、意味的なインデキシングを自動的に行うための手法を提案した。提案手法の特徴は、音声認識などによる発話内容の言語的理解を利用せず、身振り手振り、ポスターの指さし、視線といった非言語情報を利用して会話シーンを抽出し、さらに、それらのシーンがどのトピックについてなされた会話なのか、また、盛り上がった会話だったのか、といった意味づけを推測することである。本稿では、詳細な身体動作や視線情報のセンサデータを利用して、非言語情報によってどの程度、会話の内容を推し量れるか、検討した。現状ではその試みは始まったばかりであるが、このアプローチで十分な知見を得られれば、より単純で日常的なセンサ（例えば、加速度センサ数個とかマイクといったセンサ類）で非言語情報獲得を近似し、会話情報を構造化することが可能になることが期待される。

謝辞

本稿の前半で紹介した分身プレゼンテーションシステムの初期バージョンはATRメディア情報科学研究所で研究開発し、通信・放送機構の民間基盤技術研究促進制度の補助を受けた。後半で紹介した非言語情報による会話内容分析の研究で使用したデータは、ATRメディア情報科学研究所において伊藤禎宣氏、岩澤昭一郎氏らによって採取されたものをお借りしたものであり、その分析においては山本哲史氏にご助力頂いた。これらの諸氏に深く感謝する。

参考文献

[1] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Modeling focus of attention for meeting indexing. In *ACM Multimedia '99*, pp. 3–10. ACM, 1999.

[2] 角康之, 伊藤禎宣, 松口哲也, Sidney Fels, 間瀬健二. 協調的なインタラクションの記録と解釈. 情報処理学会論文誌, Vol. 44, No. 11, pp. 2628–2637, 2003.

[3] 喜多壮太郎. ジェスチャー：考えるからだ. 金子書房, 2002.

[4] 川口洋平, 角康之, 西田豊明, 間瀬健二. 展示会場における過去の対話データを利用した分身プレゼンテーション. 情処研報（モバイルコンピューティングとユビキタス通信/ユビキタスコンピューティングシステム）, Vol. MBL-32/UBI-7, pp. 225–232, March 2005.

[5] Ken Kumagai, Yasuyuki Sumi, Kenji Mase, and Toyoaki Nishida. Detecting microstructures of conversations by using physical references: Case study of poster presentations. In Satoshi Tojo Takashi Washio, Akito Sakurai and Makoto Yokoo, editors, *New Frontiers in Artificial Intelligence: Proceedings of the 19th Annual Conferences of the Japanese Society for Artificial Intelligence*, Lecture Notes in Artificial Intelligence. Springer, in press.

[6] 伊藤禎宣, 岩澤昭一郎, 馬田一郎, 鳥山朋二, 土川仁, 角康之, 間瀬健二, 小暮潔, 萩田紀博, 片桐恭弘. 外部観測可能な非言語活動による興味傾向判別の提案. ヒューマンインタフェース学会論文誌, Vol. 8, No. 1, pp. 9–22, 2006.

[7] Michael Kipp. ANVIL: A generic annotation tool for multimodal dialogue. In *Proc. Eurospeech 2001*, pp. 1367–1370, 2001.