会話テキストデータからの討論番組自動生成システム

6Z−07　有安　香子　林　正樹
NHK放送技術研究所　マルチメディアサービス

1. はじめに

放送と通信の融合は、ハード面の融合のみならず、コンテンツの融合というソフトの面からも今後ますます進んでいくと考えられる。通信において主要なテキスト中心のコンテンツと、放送の映像中心のテレビコンテンツの融合の一例として、我々は、会話テキストデータからTV討論番組を自動生成するシステムを構築し実験を行った。入力テキストの表層的な特徴から、討論番組の番組制作手法の統計的調査結果を用いて適切な演出を導き出し付加することで、TV番組としての表現力を向上させることができた。本稿では演出生成手法と構築したシステムについて報告する。

2. 討論番組自動生成システム

はじめに、討論番組自動生成システムについて簡単に説明する。図1にシステムの概念図を示す。

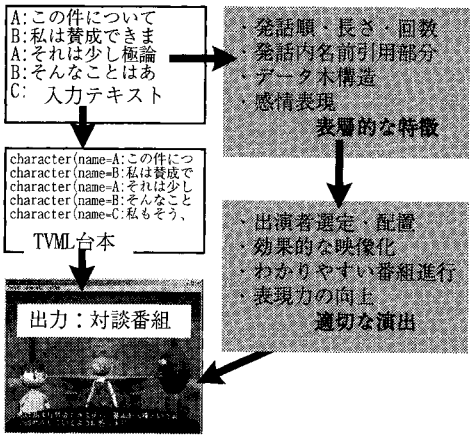

図1：システム概念図

本システムでは、TVML(TV program Making Language)[1]を用いてスクリプトからテレビ映像音声を生成する。生成の際、入力テキストの発話順・回数や名前引用、感情表現記述などの様々な特徴を抽出し、それを元に出演者の選定、カメラスイッチング、ジェスチャなどの映像演出の付加を行なう。

3. 会話テキストデータの映像化

討論番組の制作では、視聴者に討論内容の理解を促し、興味をひきつけるために、様々な演出が行なわれている。特に画面構成・ショットの選定・スイッチングなどの映像演出は、様々な演出の中でも重要な役割を担っている。

3-1. 映像演出の統計的決定法

討論番組を制作する際の番組構成の基本は、発話者のワンショットを発話順につなぐことである。発話者のワンショットを撮る際、目線処理や会話軸の考慮など画面構成の工夫を行なうことにより、話者同士の位置関係がわかり易くなり、視聴者の混乱を防止することができる[2]。更に、参話者間の関係を示唆し、画面に動きと変化を与え視聴者の興味を引きつけるため、間に周辺ロングショットやジェスチャフォローショットをはさみ、討論番組が作られていく。通常、これらの演出は経験則に基づき行なわれており、映像を作る上で最も重要な要素のひとつである。

そこで、本システムでは、実際に放送された討論番組（42対談、30時間、9000カット）を様々な角度から分析し、映像演出の統計的算出を行なった。

3-1-1. ショットの決定

先に述べたように番組構成の基本は発話者のワンショットなので、当然話者交代時には発話者のワンショットが映される確率が高く、その後挿入されていくショットの決定は様々な要素と密接に絡み合っているが、特に直前のショットとの関係が高い。表1に直前のショットの種類と直後のショット種類の関係を示す。

後＼前	発話開始	話者1S	話者周辺	ドリー	参話周辺	参話1S
話者1S	70%	11%	85%	73%	77%	72%
話者周辺	19%	32%	4%	11%	14%	6%
ドリー	5%	8%	1%	1%	2%	1%
参話周辺	1%	15%	2%	4%	3%	2%
参話1S	5%	35%	8%	11%	4%	19%

表1：ショット種類の直前ショットとの関係

A system of automatic TV debate generation from BBS-data

本システム内では、各ショットの出現頻度がこれらの確率に近くなるよう、ショットを切り替えるタイミングごとに乱数で決定していく。

3-1-2. スイッチングタイミングの決定

次に、ショット切替えのタイミングの算出方法について述べる。ショットを切り替えるタイミングをそのショット種類だけで決めると、例えば、話者ワンショットの場合、平均16秒標準偏差12秒となり、予測精度は高々20%にしかならない。これは全てのショットについて同様である。そこで、ショットを切り替えるタイミングを決める主な要因を洗い出した。

A) 参話者がジェスチャを起したとき（ジェスチャフォローショット）
B) 発話中に参話者の名前が引用されたとき
C) ひとつの発言が長く画面に変化をつけるほうが演出上望ましいとき
D) 番組構成上スーパーインポーズや説明フリップ（本システム実装上では参考URLなどの映像）が映された時
E) 他の参話者の発言を引用したとき

などが主な要素として挙げられた。

これらの要因とショットの種類（話者1s・話者周辺ショット・引用話者1s・ドリーショットなど）を説明変数とし、数量化1類によりショットの継続時間（目的変数）を算出した。結果、重相関係数0．83（予測精度70%）までを予測精度を高めることができたので、これらの要因をそれぞれ

A) 後述（3-3. キャラクタジェスチャの自動付加）
B) 入力テキストと名前のパターンマッチングによる切出
C) 入力テキストの長さを標準的話速（150文字／分）で計算し発話の長さを算出
D) スーパーインポーズは各参話者の初回発話について挿入。説明フリップは入力テキスト内に参考URLが存在するとき映像を挿入
E) 入力テキスト同士のパターンマッチング

の方法で具体的にテキストから抜き出し、得られた統計値を基準に、各ケースに応じた残差の分散で幅を持たせスイッチングタイミングを決定した。

3-3.出演者の決定・配置

以上のショット演出により映像化される討論シーンをより効果的にするために必要な出演者の決定・配置を行なわせる。出演者は発話頻度の高い発話者上位7人とし、第一発話者に司会者の役割を担わせ、出演者以外の発言の代理発言や、番組進行などを行なう。また出演者の配置は、発話時間の近さや発話回数の頻度、初回の発話の順番などを元に決定する [2]。

3-3. キャラクタジェスチャの自動付加

出力番組内での会話を自然に見せるため、キャラクタへのジェスチャ付与も同時に行なう。対面討論時の参話者のジェスチャについて解析を行なった統計データ[3]をもとに、50文字程度のテキストごとに、話の終わり又は文法的な切れ目を検出し、乱数で幅をもたせ決定したタイミングによって、ジェスチャ付加を行なった。

また、顔文字や記号、語尾延ばしなど、現在通常的に使われている感情表現句をデータからパターンマッチングで抽出し、これに相当するジェスチャの付加も同時に行なった。

4. システム

以上に述べた演出を付加し、討論番組を生成するシステムをWindows PC上で実装した。入力テキストは電子掲示板などのデータをインターネットを通して自動的に収集したものとし、入力データは即座に解析され番組台本へと変換される。この台本を元に、番組を再生する際TVML のコントロールモードを用いてリアルタイムで上記の演出を生成し、出力番組に付加している。

5. まとめ

会話テキストデータからTV討論番組を自動生成するシステムを構築し実験を行った。入力テキストの表層的な特徴から、討論番組の番組制作手法の統計的調査結果を用いて演出を付加するシステムを構築し、出力番組の表現力の向上を図った。今後は、より予測制度の高い統計値を得られるようデータ解析の方法を検討しなおし、より質の高い演出を自動生成できるよう改良を行なう。

参考文献：

[1] http://www.strl.nhk.or.jp/TVML/
[2] 時系列に並んだ発話データの映像化における一考察，有安，映像情報メディア学会夏季全国大会（2001）
[3] Öestrom B"Turn-taking in English conversation" Lund,Sweden:Gleerup(1983)