言語情報と映像情報の統合による教示発話の構造解析

柴田知秀 立木将人 河原大輔 岡本雅史 黒橋禎夫 西田豊明
東京大学大学院情報理工学系研究科

{shibata,tachiki,kawahara,okamoto,kuro,nishida}@kc.t.u-tokyo.ac.jp

1 はじめに

デジタル放送やブロードバンドの普及により、テキストだけでなく、大量の映像が配信されるようになった。しかし、大量の映像から見たいシーンを効率的に検索し、容易にアクセスできる技術はまだ実用化されていない。映像の中でも、料理、工作、スポーツなどといった作業を教示するような映像の需要は高く、作業について方法、注意点、コツなどを対話的・映像的に教示するシステムが求められている。本研究では、そのようなシステムの構築を目指す。

一番単純なシステムとして、音声認識結果やクローズドキャプション (図 1) などといったテキスト化された教示発話とユーザの検索キーワードとのマッチングによって、該当する部分の映像を提示するものが考えられる。このようなシステムの問題点として、以下のものがあげられる。

- 教示発話は話し言葉であるため、省略が多い。また、作業の説明だけでなくコツや雑談などが含まれる。そのため、検索キーワードと正確にマッチングを行なうことができない。
- 映像は時間軸を持ったメディアであり、一覧性が悪い。従って、早送りや巻き戻しができるとはいえ、映像全体を概観することができない。

これらの問題を解決するためには、発話を深く解析し、構造化を行なう必要がある。教示発話には、図 2 に示すような作業の階層構造があり、これを自動抽出することにより、ユーザが映像に対して容易にアクセスできる方法を提供できる。このような構造を捉えるには、発話の談話構造だけでなく、ショットなどといった映像の構造を考慮に入れることにより、より高精度に認識できると考えられる。

階層構造を抽出するためには、まず、作業のまとまりを認識しなければならない。本稿では、料理教示発話を解析し、言語情報と映像情報を統合することにより、作業のまとまりを認識する手法について述べる (作業のまとまりの開始の発話を作業の切れ目と呼ぶことにする)。

[192] 9:32:45 香りが立つまでいためたらここで水を入れます。
[193] 9:32:55 このままだしを取り沸くのを待ちます。
[194] 9:33:01 私はいろんなめんを取り寄せて食べます。

図 1: クローズドキャプションの例

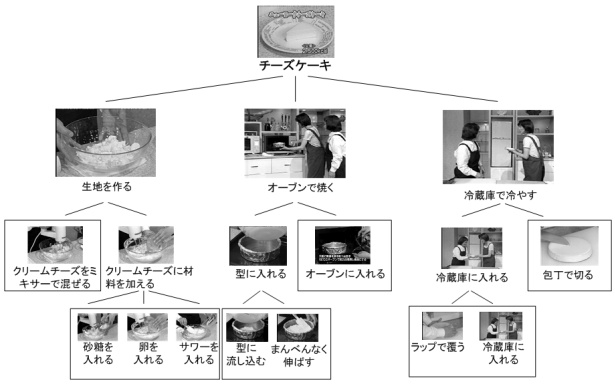

図 2: 料理における階層構造

2 章で発話の言語解析、3 章で映像解析について述べる。そして、4 章で言語情報と映像情報を統合し、作業の切れ目を認識する手法について述べる。

2 料理教示発話の言語解析

発話を解析し、図 3 に示すような談話構造を求める。言語解析は、省略・照応解析、発話タイプ付与、談話構造解析からなる [2]。それぞれの処理について順に述べる。

2.1 省略・照応解析

日本語の文章では格要素が指示詞となったり、省略されることが頻繁に起こる。本研究で対象としているテキストは話し言葉であるため、特にその傾向が強い。そこで、Web より収集した料理テキストから格フレーム辞書を自動構築し、これを利用して省略・照応解析を行う [3]。

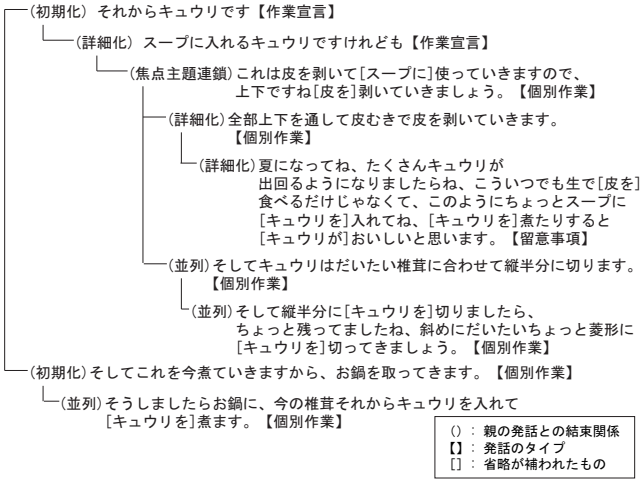

図 3: 料理教示発話の談話構造

2.2 発話タイプ付与

料理教示発話には、実際の作業に関する発話だけでなく、コツ、留意事項、雑談といった発話がある。[1] の分類を参考にし、料理教示発話を、以下の 7 つに分類した。

作業宣言 (ひとまとまりの作業を始める宣言)

・さ、では、ステーキの材料にかかります。

個別作業 (具体的な作業)

・お鍋にお水を入れます。

料理状態

・ニンジンの水分がなくなりました。

食品・道具提示

・材料は、牛ひき肉、百五十グラムです。

代替可 (別の食品、手法などでもよいこと)

・もし半個ぐらいでしたら、手で搾って頂いても結構です。

留意事項 (料理のコツ、注意すべき事項など)

・こうしますと味の染みが良くなります。

その他

・こんばんは。

作業宣言、食品・道具提示、代替可、留意事項、その他については発話の文末表現のパターンを記述することで認識できる。そのパターンは、形態素を単位として記述し、各形態素について語、品詞、活用形、シソーラスの意味素性をチェックできる能力をもつものである。

個別作業、料理状態については、料理ドメインに対してすべての述語を列挙するというアプローチも考えられるが、他のドメインへの適用の可能性を考え、自動詞、形容詞＋「なる」などを料理状態、それ以外を個別作業とする一般的な規則を用いている。

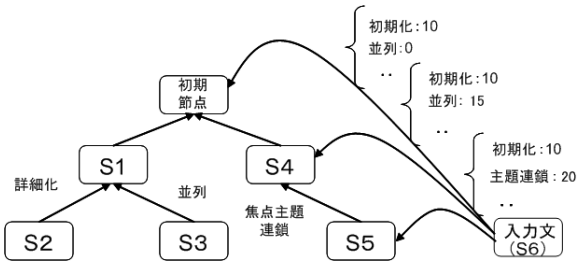

図 4: 談話構造のモデル

表 1: 談話構造解析のルール

結束関係	スコア	適用範囲	接続可能文パターン	入力文パターン
並列	5	1	～	そして～
並列	40	*	[並列]	～さらに～
対比	30	1	～	むしろ～
詳細化	30	1	～	すなわち～
詳細化	15	1	～	<留意事項>
焦点主題連鎖	25	1	<食品提示>	<個別作業>
理由	30	1	～	～からだ

2.3 談話構造解析

2.3.1 談話構造のモデル

談話構造のモデルとして、各発話を一つのノードとし、関係する発話がリンクされたグラフ構造を考える。談話構造のモデルを図 4 に示す。談話構造の初期状態として初期節点を考え、初期節点に接続することは、その発話から新しい話題が始まることを意味し、この時の関係を “初期化” とする。

2.3.2 談話構造解析手順

談話構造解析は、種々の表層的手がかりをもとに、各入力発話に対して、関係をもつ以前の発話 (接続文) とその間の結束関係を逐次的に求める [4]。新しい話題が導入された後に古くなった話題に接続することはないという仮定をおき、入力文は談話構造の右端の発話にのみ接続可能と考える。

以下のような観点から、さまざまな接続可能文との間のさまざまな結束関係を考慮し、最終的に最も高い合計点を得た関係を採用する。

談話構造解析ルール

表 1 のような談話構造解析ルールを作成した。表 1 において、適用範囲とはどれだけ離れた発話との関係まで考えるか、接続可能文パターン、入力文パターンは、それぞれに対する表層表現、発話の結束関係 ([ ] で括られたもの)、発話タイプ (<> で括られたもの) などのパターンである。ルールが一致した場合には、結束関係欄の関係に対して、スコア欄の点数が与えられる。

語連鎖

任意の 2 文間において、主題と主題、焦点と主題で語の連鎖があれば、それぞれ、“主題連鎖”、“焦点主題連鎖” にスコアを与える。

3 映像解析

映像には以下のような種々のレイヤがある。

フレーム 最小の単位、1 秒あたり 30 フレーム

ショット 単一のカメラから撮影されたフレームの集合

シーン 意味的につながりのあるショットの集合

クリップ シーンの集合

料理番組は編集済の映像であり、カメラワークには編集者の意図があると考えることができる。このような編集者の意図を利用することにより、構造解析の補助を行なう。

料理番組では主に、顔が映っているショットと手元のアップのショットからなり、それらは基本的に交互にあらわれる。顔が映っているショットでは、主に手順の説明を行ない、手元のアップのショットでは、実際の作業を視覚的に伝えている。教示者の顔が映っているショットには次のような場合がある。

1. 次に行う作業の手順の説明を行う
2. 今行なっている作業についてコツ・留意事項などを説明したり、雑談をしている
3. 長い作業の途中で短い顔のショットが挿入される
4. 教示者が複数の場合に発話している人物のショットになる

このうちの 1 は、作業のまとまりを認識する際の有効な手がかりとなる。

本研究では、映像解析によりショットの切れ目であるカット点と顔の検出を行なった。隣接する 2 フレームのカラーヒストグラムの差が閾値以上のものをカット点とし、また、ニューラルネットワークを用いた手法で顔検出を行なった。

4 言語情報と映像情報の統合

2 章で述べた言語解析結果と、3 章で述べた映像解析結果を統合することにより、作業の切れ目を抽出する。

2 章で述べたように、言語では、結束関係が “初期化” である発話は、前のいずれの発話との関係が捉えられなかったものであり、作業の切れ目の候補となる。一方、3 章で述べたように、映像では、顔のショットである時の発話が、作業の切れ目の候補となる。それらの関係は図 5 のようになっている。

したがって、言語で “初期化” であり、かつ、顔のショットである発話 (図 5 の B+E) を作業の切れ目とすることで一方の情報のみを利用した場合よりも精度が向上すると考えられる。

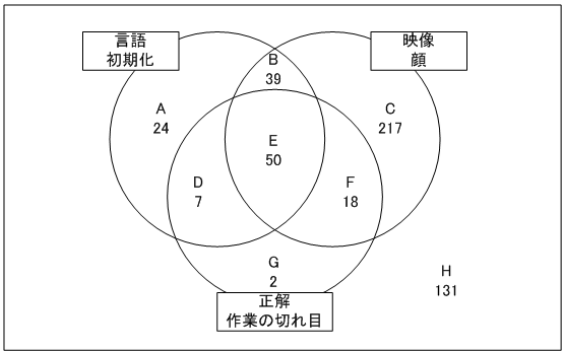

図 5: 言語の “初期化” と映像の “顔” の関係

さらに、発話の周辺の情報を考慮するために以下にあげるような素性を用いて、各発話が作業の切れ目となるかを機械学習する。図 6 で、発話 U2 が作業の切れ目となるかを学習する際に用いる素性を説明する。

言語
- 対象発話 (U2) の結束関係
- 対象発話とその前後の発話 (U1,U3) の発話タイプ
- 切れ目となる手がかり表現（そうしましたら、それでは次に、など）が存在するかどうか
- 対象発話の前後 8 文のブロック間のコサイン類似度

映像
- 対象発話の先頭から最も近い顔のショット (S2) までの時間 T0
- 前後のカット点 (S1,S3) までの時間 T1,T2
- 前後の手元のショット (S1 と S3) の類似度

音声
- 発話の前後 3 秒間における無音の時間 Q1,Q2

5 実験結果

「きょうの料理」の 5 番組を用い、各発話に対して、作業の切れ目であるかの正解を付け、実験を行なった。

機械学習には、TinySVM[1]を用い、5-fold の交差検定を行った。図 5 に発話の数を示し、精度は表 2 のようになった。表中の言語、映像、言語＋映像は、それぞれ、言語解析で結束関係が “初期化” である発話、顔が検出された発話、結束関係が “初期化” であり、かつ、顔が検出された発話を正解とした結果であり、言語+映像+機械学習は言語解析の結果と映像解析の結果を素性として機械学習を行なった結果である。言語と映像を統合すると言語に対して精度が F 値で 0.032、

[1] http://cl-aist-nara.ac.jp/~taku-ku/software/TinySVM/

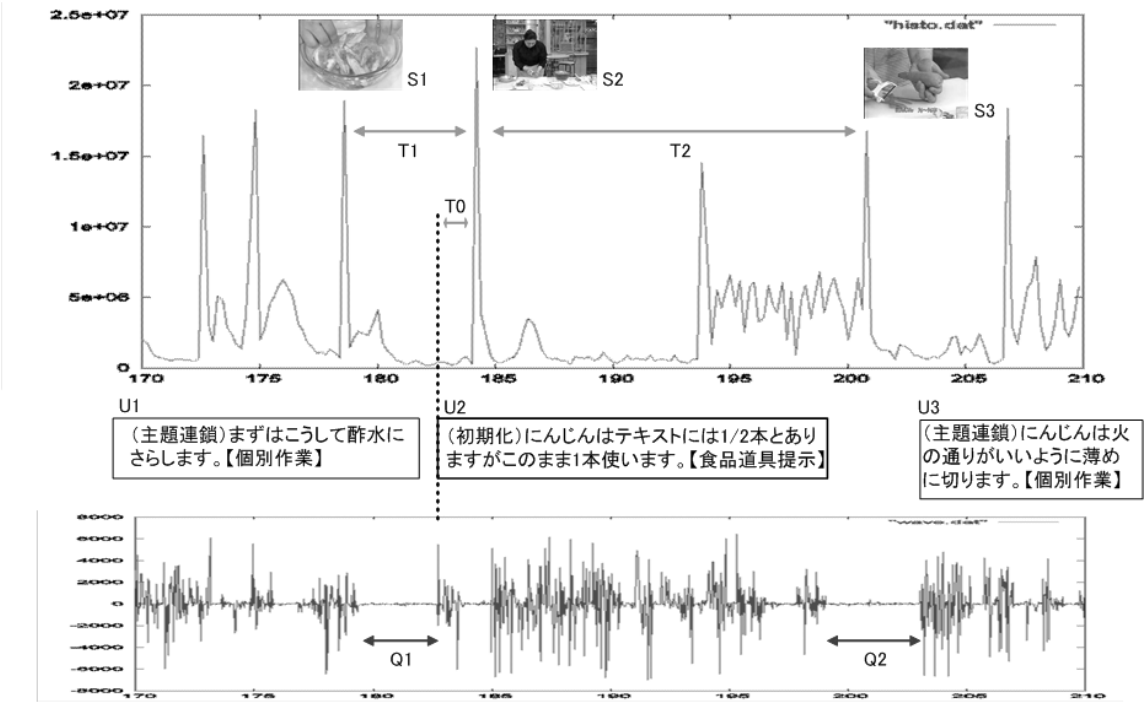

図 6: 機械学習で用いた素性

表 2: 作業の切れ目の検出結果

	Precision	Recall	F
言語	47.5%	57.4%	0.579
映像	21.1%	89.6%	0.342
言語+映像	56.7%	66.2%	0.611
言語+映像+機械学習	67.2%	58.4%	0.625

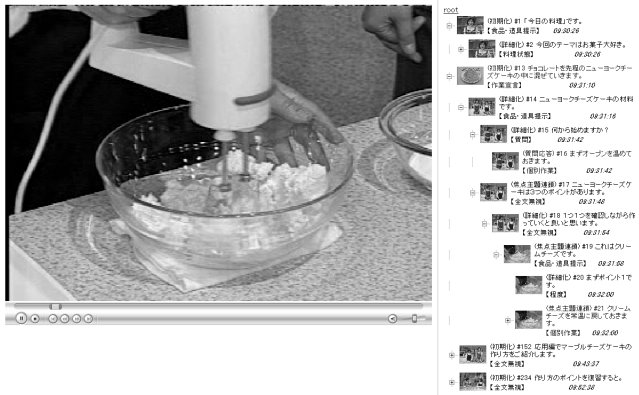

図 7: 料理映像検索システム

さらに機械学習を行なうと 0.014 あがっていることがわかる。

6 結論

本稿では言語情報と映像情報を統合することにより、料理教示発話の構造解析を行なう手法を述べた。実験結果から、言語と映像の統合の有効性が示された。現在、料理映像検索システム (図 7) を構築しており、検索・要約の観点から構造解析を再度検討する予定である。

参考文献

[1] Hidekatsu IZUNO, Yuichi NAKAMURA, and Yuichi OHTA. Quevico: A framework for video-based interactive media. In *Working Notes WS-5 International Workshop on Intelligent Media Technology for Communicative Reality, PRICAI-02 (Seventh Pacific Rim International Conference on Artificial Intelligence)*, pp. 6–11, August 2002.

[2] Tomohide Shibata, Daisuke Kawahara, Masashi Okamoto, Sadao Kurohashi, and Toyoaki Nishida. Structural analysis of instruction utterances. In *Proceedings of Seventh International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES2003)*, pp. 1054–1061, 9 2003.

[3] 河原大輔, 黒橋禎夫. 自動構築した格フレーム辞書に基づく省略解析の大規模評価. 言語処理学会 第 9 回年次大会, pp. 589–592, 3 2003.

[4] 黒橋禎夫, 長尾眞. 表層表現中の情報に基づく文章構造の自動抽出. 自然言語処理, Vol. 1, No. 1, pp. 3–20, 10 1994.