言語情報と映像情報の統合による作業教示映像の構造理解

柴田知秀

1　はじめに

近年の計算機･ネットワーク環境の発展により、大量の映像が配信・蓄積されるようになってきた。蓄積された映像を高度に利用するには、映像の各部分において何に関する映像であるかといった情報を付与する必要がある。これは現在のところほとんど人手で行なわれており、大規模映像に対して行なうには自動付与する技術が必要となる。我々は、料理映像を対象として、映像セグメントにトピック (下ごしらえ、炒める、盛り付けなど) をラベリングを行なった [2]。例えば図 1 では順に、「下ごしらえ」、「炒める」、「盛り付け」とラベリングを行なう。

2　利用する特徴量

まず、トピック推定に利用する特徴量について述べる。トピックの推定には、「切る」「火をつける」「のせる」などといった作業に関する発話が有用であるが、料理ドメインに限定しても作業に関する用言は多様であり、これだけではロバストに解析することができない。

一方、画像の情報としては、背景の色情報を利用することができる。例えば、「炒める」「煮る」といった作業はガスレンジ台で行なわれるため、背景が黒であることや、「下ごしらえ」「盛り付け」などの作業はまな板の上で行なわれるため、背景が白であるといった情報を手がかりとすることができる。

またこれらに加えて、トピックが変化したことを示す手がかり表現や無音、トピックが同一であることを示す語連鎖や用言の一致などを利用する。

2.1　言語情報

テキストとして料理番組に付随するクローズドキャプションを利用する。クローズドキャプションに対して JUMAN･KNP で形態素・構文・格・省略解析を行ない、その結果に対して以下の処理を行なう。

発話タイプ認識による作業に関する発話抽出　作業教示発話の場合、作業に関する発話が中心であるが、コツや留意事項、雑談などの発話も含まれている。トピック推定には作業に関する発話が有用であり、その他の発話はノイズになると考えられる。そこで作業に関する発話のみを抽出する。そのために、まず文を節に分割し、節末の表層パターンを用いて発話タイプの認識を行なう [1]。このうち、[作業:大]、[作業:中]、[作業:小] の発話のみを抽出する。

また一般に用言は複数の意味をもつ。例えば、「入れる」という用言は、「塩を入れる」、「包丁を入れる」において異なる意味を持ち、これらは異なるトピックで現われる。したがって、用言の表記を利用するのではなく、格･省略解析の際に、意味ごとに分けられた格フレームの選択を行なう。用言格フレームは料理 Web テキストから自動構築した。

手がかり表現　多くの研究でこれまで指摘されてきたように、トピックの変化を示す手がかり表現がある。本研究では、「では」「次は」「そうしましたら」など約 20 個を利用した。

語連鎖　ある 2 つの作業が同一の食材に対して行なわれている場合、それらのトピックは同一である可能性が高いと考えられる。

用言の一致　連続する 2 発話の用言が一致する場合、それらのトピックは同一である可能性が高いと考えられる。格要素が同じ場合と異なる場合があるが、いずれの場合も同一のトピックであると考えられる。

2.2　画像情報

浜田ら [3] の研究を参考にし、比較的安定して情報を抽出することができる背景画像に着目する。図 1 に示すように、画面下部の RGB の重心を特徴量とする。

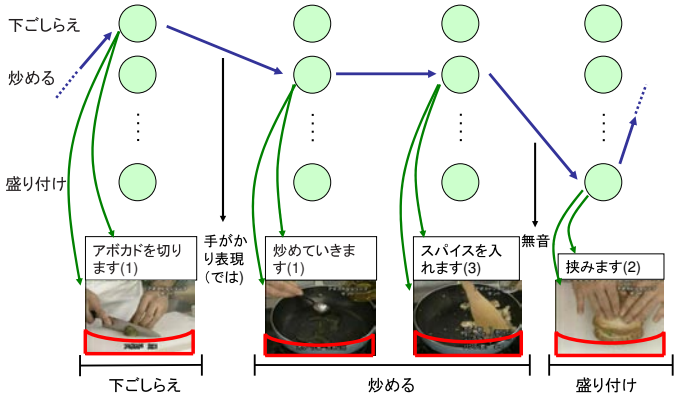

図 1: 隠れマルコフモデルによるトピック推定 (発話末尾の括弧内に格フレーム番号を示す)

2.3 音声情報

トピックが変化する時に無音がおかれることが多く、無音がトピックの変化を検出する手がかりとして利用することができる。本研究では、音声の振幅が閾値以下である部分が 1 秒以上続く時を無音とした。

3 HMM によるトピック推定

隠れ状態がトピックにあたり、2 章で説明した種々の特徴量が出力シンボルとして観測される HMM でトピックの推定を行なう (図 1)。このモデルでは、格フレームと背景画像は隠れ状態から出力され、トピックが同一/異なることを捉えた特徴量 (手がかり表現、語連鎖、用言の一致、無音) は隠れ状態を遷移する時に出力される。HMM のパラメータを以下にあげる。

- 隠れ状態 s_i : トピックにあたる。本研究では以下の 8 種類 ($N = 8$) を考える。

  下ごしらえ、蒸す、ゆでる、揚げる、
  煮る、炒める、盛り付け、その他

- 初期状態確率 π_i
- 状態遷移確率 a_{ij} : 状態 i から状態 j への遷移確率であり、トピックの遷移確率にあたる。
- 出力シンボル確率

  - 格フレーム $b_j(cf_k)$: 状態 s_j から格フレーム cf_k が出力される確率。

  - 背景画像 $b_j(R,G,B)$: 状態 s_j から背景画像の色情報 (R,G,B) が 出力される確率であり、平均 (R_j,G_j,B_j)、分散 σ_j の正規分布で出力されると考える。

表 1: トピック推定の実験結果

用いる特徴量				精度
格フレーム	背景画像	手がかり表現などの言語情報	無音	
○				59.8%
○	○			68.9%
○	○	○		75.4%
○	○	○	○	79.5%

  - 手がかり表現、語連鎖、用言の一致、無音: 状態 s_i から状態 s_j に遷移する時に各特徴量が出力される確率。状態 s_i と s_j に依存するのではなく、隣接する状態が同一か異なるかに依存すると近似し、隣接する状態が同一の場合 $(i = j)$ は p_s、異なる場合 $(i \neq j)$ は p_d とする。

これらのパラメータを、教師なし学習である Baum-welch アルゴリズムで学習する。

4 実験

提案手法の有効性を確かめるために、NTV の「キューピー 3 分クッキング」約 70 日分の映像を用いて実験を行なった。学習されたモデルを番組 5 日分に適用して実験を行ない、トピックが正しいかどうかを節単位で評価した。表 1 に実験結果を示す。言語情報に加えて映像情報を利用することにより精度が向上していることがわかる。また、それらに加えて種々の言語手がかりや音声情報も利用することにより精度が向上した。

5 おわりに

本稿では、言語情報と映像情報を統合し、HMM を用いてトピックの推定を行なった。実験を行なったところ、提案手法の有効性を示すことができた。

参考文献

[1] 柴田知秀, 黒橋禎夫. 料理教示発話の理解と作業構造の自動抽出. 情報処理学会 自然言語処理研究会, No. 2004-NL-164, pp. 117–122, 11 2004.

[2] 柴田知秀, 黒橋禎夫. 言語情報と映像情報を統合した隠れマルコフモデルに基づくトピック推定. 言語処理学会 第 12 回年次大会, 3 2006.

[3] 浜田玲子, 井出一郎, 坂井修一, 田中英彦. 料理テキスト教材における調理手順の構造化. 電子情報通信学会論文誌, Vol. J85-D-II, No. 1, pp. 79–89, 2002.