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

柴田 知秀　　黒橋 禎夫
東京大学大学院情報理工学系研究科

{shibata,kuro}@kc.t.u-tokyo.ac.jp

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。本稿では、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) をラベリングする手法を提案する。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。ラベリング結果は要約の生成や後述する物体モデルの学習に利用する。

料理映像のような作業教示映像の場合、トピックの推定に、作業に関する発話 (切る、洗う、火をつけるなど) を利用することができる。一つのアプローチとして、「切る」は「下ごしらえ」、「火をつける」は「炒める」といった正解データを作り、そこから学習するといった方法が考えられる。しかし、作業に関する表現が多数あることや、他のドメインへの移植性を考えると現実的な方法であるとは言えない。そこで本研究では、作業に関する発話を時系列データと考え、それらが、隠れ状態であるトピックから生成されるという隠れマルコフモデル (HMM) でモデル化し、モデルを教師なし学習するというアプローチをとる。さらに、言語情報だけでなく映像情報・音声情報も利用することによりロバストに行なう。

2　関連研究

Barzilay らは、地震などの 5 つのドメインを対象として、HMM を用いて生コーパスからトピックの遷移モデルを構築し、文の並び順の決定と要約の生成の 2 つのタスクに利用している [2]。トピック (震度の情報、被害の情報など) が遷移しながら、そのトピックが出力する言語モデルを元に文が生成されるというモデルをとっている。

また映像解析の分野において、Nguyen らは、野球放送からハイライトシーンを抽出することを目的として、ビデオデータをインデキシングするための統計的なフレームワークを提案している [1]。各フレームの主成分分析による特徴量、テクスチャ情報であるフラクタル特徴量、移動物体の情報を反映した差分特徴量の 3 つの特徴量を用いて、マルチストリーム HMM でモデル化している。

我々は、料理映像を対象とし、発話から作業に関するものを抽出し、HMM を用いてトピック推定を行なっている [5]。本研究では、これに加えて、映像情報や種々の言語手がかり・音声情報も利用し、精度が向上することを示す。

3　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。以下に利用する特徴量をまとめる。

言語
- 用言 (後述する格フレームを利用)
- 手がかり表現 (では〜、次に〜など)

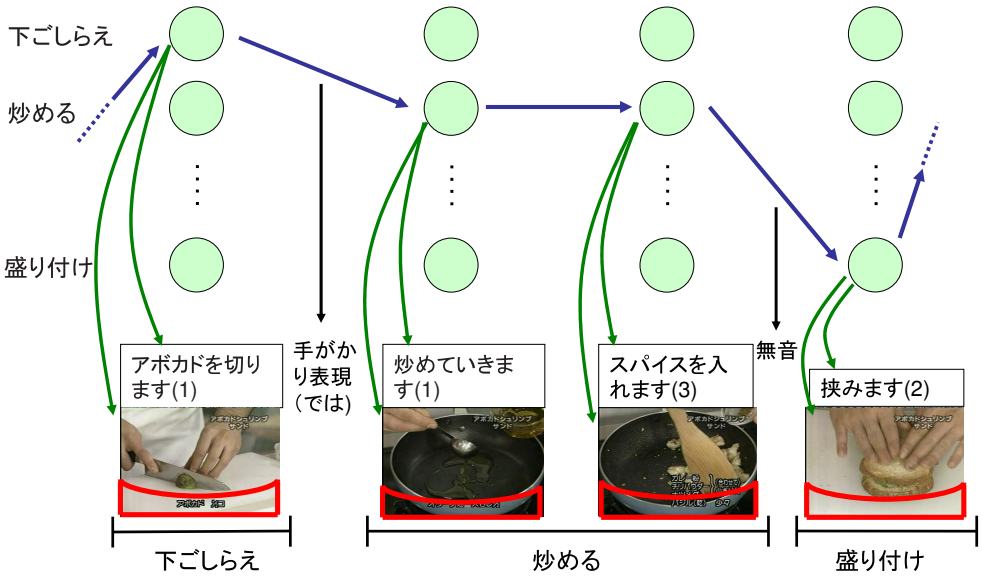

図 1: 隠れマルコフモデルによるトピック推定 (発話末尾の括弧内に格フレーム番号を示す)

- 語連鎖
- 用言の一致

画像
- 背景画像

音声
- 無音

以下、各特徴量について順に詳しく述べる。

3.1 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。

3.1.1 発話タイプ認識による作業に関する格フレーム抽出

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。表 1 に発話タイプを分類したものを示す。このうち、[作業:大]、[作業:中]、[作業:小] の発話のみを抽出する。ただし、以下の例のように、理由節・条件節は前後のトピックに言及している場合があるので、抽出しない。

(1) a. 切りましたら(条件) 炒めていきます。
b. プチトマトは油で 揚げるので(理由)、切り込みを入れます。

表 2: 用言格フレームの例

用言	格	用例
切る (1)	ガ	【主体】
	ヲ	豚肉、大根、 こんにゃく、 ···
	ニ	正方形、櫛形、三角形、 ···
切る (2)	ガ	【主体】
	ヲ	水気、水分、汁気、 ···
	ノ	なす、豆腐、肉、 ···
入れる (1)	ガ	【主体】
	ヲ	塩、油、野菜、 ···
	ニ	鍋、ボール、容器、フライパン、 ···
入れる (2)	ガ	【主体】
	ヲ	包丁 ···
	ニ	魚、腹、付け根、 ···

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、用言の表記を抽出するのではなく、格・省略解析の際に対応付けられた、意味ごとに分けられた格フレームを抽出する。用言格フレームは料理 Web テキストから自動構築した。例を表 2 に示す。

3.1.2 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。

3.1.3 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。以下の例では、「おろす」、「上げる」

表 1: 発話タイプの分類

[**作業:大**] ・さ，では，ステーキの材料にかかります.	[**留意事項**] ・最初に肉をパラパラに炒める事がポイントです.
[**作業:中**] ・強火で油を温めましょう. ・じゃあ炒めていきましょう.	[**代替可**] ・もし半個ぐらいでしたら，手で搾って頂いても結構です.
[**作業:小**] ・お鍋にお水を入れます.	[**食品・道具提示**] ・材料は，牛ひき肉，百五十グラムです.
[**料理状態**] ・ニンジンの水分がなくなりました.	[**雑談**] ・暑くなってきましたね.
	[**効果**] ・そうすると最初の方はアクが出てきます.

ともに同一の食材「かぶら」に対して行なわれているので、これらの 2 発話は同一のトピック (この例では「下ごしらえ」) である可能性が高いといえる。

(2) a. かぶらをおろし金でおろしていきます。
b. おろしたかぶらをざるに上げます。

しかし、本研究で扱っている話し言葉の場合、省略が頻繁におきるので、表層的な文字列の照合では語連鎖を検出することができないことが多い。そこで、用言 (例 (3)) と名詞 (例 (4)) の省略解析結果[1]も利用し、語連鎖の検出を行なう。

(3) a. 小松菜を切ります。
b. 一度 [小松菜を] 洗います。

(4) a. にんじんを大体４ｃｍくらい切ります。
b. [にんじんの] 皮をぐるっとむきます。

3.1.4 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。

- 格要素が同じ場合： 確認または内容を詳細にするため、同じ内容の発話を繰り返している。

  (5) a. ごぼうにはアクがあるので酢水にさらします。
  b. まずはこうして酢水にさらします。

- 格要素が異なる場合： 異なる食材に対して同一の作業を行なっている。

  (6) a. とうがらしを入れて下さい。
  b. 鶏手羽を入れます。

3.2 画像情報

現在の画像処理技術では、人手による強い作り込みなどを行なわなければ、映像中から何が映っているのか、またはどのような動作が行なわれているかといった情報を抽出することは難しい。したがって、浜田ら [6] 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。

3.3 音声情報

Galley ら [3] などが指摘しているように、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。

4 HMM によるトピック推定

隠れ状態がトピックにあたり、3 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[2]。

  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。

[1] [ ] は省略が補われたことを示す。

[2] これらのトピックは「1 から始める料理の基本」http://www.recipe.nestle.co.jp/from1/sitemap.htm を参考にして設定した。

表 3: 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%

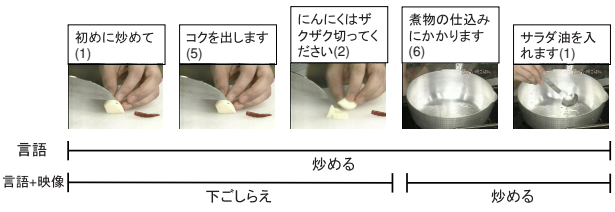

図 2: 成功例

- 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。ただし、発話のタイプが [作業:大/中] で、用言の原形がトピック名と一致する場合、または、「〜の下ごしらえです」「〜の準備です」はトピックを下ごしらえとするなどのルールにマッチする場合、その発話のトピックを固定する。図 1 の例では、「炒めていきます」の発話タイプが [作業:中] で、用言の原形がトピック「炒める」に一致するので、この時のトピックは「炒める」に固定する。

5 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」の映像を用いて実験を行なった。この番組は 10 分番組であり、約 70 日分のデータを利用した。

5.1 トピック推定

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 3 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。

画像情報を加えることにより成功した例を図 2 に示す。言語情報だけの場合は、「炒める」という発話からトピックを「炒める」と解析しているが、実際はこの発話は後に行なう作業について述べている。背景が白色の情報を利用することにより、トピックが下ごしらえと解析することができている。

誤り原因としては、「野菜を切る」と「油を切る」がどちらも「切る (1)」の格フレームになっているため、誤った学習が行なわれてしまっているなどがあった。

表 4: 物体モデルの自動学習の実験結果

トピック推定	精度
なし	40 / 73 (.548)
あり	50 / 73 (.685)

5.2 物体モデルの自動学習への利用

我々は大量の映像から、アップ画像とキーワードのペアを収集し、物体のモデル (色情報) を自動学習している [4]。料理の場合、調理されると変形・変色することから、「炒める」や「盛り付け」からはよい学習データを得ることができず、トピックが下ごしらえのところから学習データを収集することにより精度が向上することが考えられる。

そこで、本稿で述べた手法でトピックを推定し、トピックが下ごしらえのところのみから学習データを集めた。学習された 73 食材のモデルが妥当かどうかを評価したところ、表 4 のように精度が 13.7%向上した。

6 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] Nguyen Huu Bach, 篠田浩一, 古井貞熙. 隠れマルコフモデルを用いた野球放送の自動的インデキシング. 電子情報通信学会 技術研究報告 PRMU2004-107, pp. 13–19, 2004.

[2] Regina Barzilay and Lillian Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the NAACL/HLT*, pp. 113–120, 2004.

[3] Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 562–569, 7 2003.

[4] 加藤紀雄, 柴田知秀, 黒橋禎夫. 言語情報と映像情報の統合による物体のモデル学習と認識. 言語処理学会 第 11 回年次大会, 3 2005.

[5] 柴田知秀, 黒橋禎夫. 隠れマルコフモデルによるトピックの遷移を捉えた談話構造解析. 言語処理学会 第 11 回年次大会, 3 2005.

[6] 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.