言語情報と映像情報の統合解析: 作業教示映像の構造理解に向けて

柴田知秀

1　はじめに

映像情報を高度に利用するには、何が映っているか、何が行なわれているかなどの情報を明示的に記号として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、大変コストがかかるものとなっている。将来的に、そのような情報の自動付与を考えた場合、映像中のナレーションや登場人物の発話が大きな手がかりとなる。

本研究では、このような問題意識から、映像情報中の話し言葉の解析を行なった [3]。具体的には、作業教示映像、特に料理番組映像を対象とし、当面は音声認識の問題をさけ、番組のクローズドキャプションを利用している。話し言葉の解析で重要なことはそれを映像的な文脈に対応付けて理解することである。しかし、映像から直接的に情報を取り出すことはまだ難しいことから、まずは話し言葉単独の解析を十分に高度化し、その上で、言語情報が少しリードする形で映像情報との統合を行なう。

2　作業教示発話の解析

本研究で対象としている料理番組のクローズドキャプションの解析例を図 1 に示す。図において、文中の括弧 ([]) で示されたものは省略要素が補われたものであり、節末の括弧 (<<>>) は発話のタイプ、結束関係、親の節の文番号／節番号を示すものである。この例に示すように、話し言葉では主語、目的語などが頻繁に省略され、その一方で、説明が何度か繰り返されるという冗長性もある。また、作業の説明だけでなく、コツや注意点などの説明も含まれている。

そこで、料理分野の「常識」に相当する知識を自動構築し、それを用いて発話中の明示されない関係の検出を行なう。さらに、各発話のタイプを解析し、それらの情報を統合して発話全体の談話構造を求める。

2.1　格フレームの自動構築と関係解析

省略された関係を検出するためには、各名詞、用言に対して必須的に関係するものとしてどのようなものがあるかという常識的知識が必要となる。このような知識は格フレームと呼ばれている。本研究では、格フレームをコーパスや国語辞典から自動構築した [2]。そして、自動構築した格フレームを用いて、必須的な要素が欠けていることを認識し、次に、そこに何が補われるべきであるかを文脈中から探し出すことにより、省略の解析を行なう。

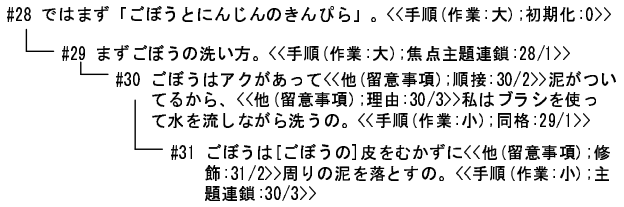

図 1: 料理番組のクローズドキャプションの解析例

2.2　用言の類義表現の自動獲得

話し言葉では何度か同じ内容の発話が繰り返されるという冗長性がある。そういった冗長性を認識するには、「弱火にする」と「火を弱める」といった用言の類義表現の知識が必要となる。このような用言の類義表現を、2 用言における前後の用言の分布の類似度に着目し、コーパスから自動獲得した。

2.3　発話タイプの解析

作業教示発話では、基本的には作業が順をおって説明されるが、中にはコツ、注意点や、雑談のような発話もある。[1] を参考にし、発話タイプを作業、料理状態、留意事項、雑談などの 9 種類に分類する。発話のタイプは節ごとに考え、節末のパターンを記述することで認識した。

2.4　談話構造解析

省略解析結果、発話タイプ、表層パターンを統合することにより、節間の関係を逐次的に求める。談話構造解析の結果、図 1 に示すような談話構造を得ることができる。

3　言語情報と映像情報の統合による物体のモデル学習

現在の省略解析・談話構造解析の誤りの多くは、映像中に何が映っているかがわかれば解消できる可能性があると考えられる。物体認識を行なうには、物体の色・形状・大きさといった知識が必要であるが、この知識をどのようにして得るかが問題となる。画像に人手でキーワードを付与したデータから学習により物体認識を行なう手法もあるが、画像にキーワードを付与するには大きなコストがかかってしまう。そこで、本研究では、映像中の発話に着目し、言語情報と映像情報を統合することにより、大量の映像から物体のモデルを自動学習する。

3.1　注目領域とキーワードのペアの収集

映像に対して画像処理と言語処理の両方を行なうことによって、モデル学習を行ないやすい、大写しの画像とキーワードのペアを収集する。

3.1.1　大写し画像と注目領域の抽出

以下のような画像処理により、物体が大写しになっている画像を選び、そこから注目領域を抽出する。

1. エッジ抽出による大写しの判定

   エッジ抽出を行ない、エッジ率 (エッジ検出された画素/ 全画素) が閾値を下回っているものを大写しと判定する。この処理により、食材が複数映っている画像や顔が映っている画像を除外する。

2. RGB 空間への写像と極大点の探索

   大写し画像において、各画素を RGB3 次元空間に写像する。3*3 のメディアンフィルタにより平滑化を行ない、山登り法で極大点を探索する。極大値が閾値を下回るものは除外する。

3. 注目領域の抽出

   抽出された極大点のうち、画像の中心と重心との距離、重心から各点までの距離の分散を考慮することにより、最も焦点が当っているような領域を抽出する。

3.1.2　キーワードの抽出

2 節で述べた談話構造解析を行ない、一つの談話構造木からキーワードを一つ選ぶ。具体的には、シソーラスで「食材」タグのふられたものに対して、発話タイプ、名詞が主節にあるか従属節にあるか、省略解析結果かどうかなどを考慮してスコア付けを行ない、最もスコアの高いものを選ぶ。

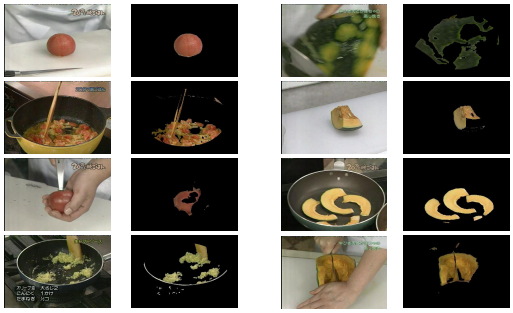

図 2: 注目領域とキーワードのペアの例

3.2　物体モデルの構築

大写しと判定された画像と、時間的に近い談話構造木を対応付け、注目領域とキーワードを対応付ける。実際に得られた注目領域とキーワードのペアの例を図 2 に示す。図の左の列は原画像、右の列はそこから抽出された注目領域を示す。食材ごとに、注目領域の RGB データを計数し、最も頻度の高い RGB (の平均) を物体モデルとする。今後、学習された物体モデルを用いて、言語解析と統合しながら物体の認識を行なう予定である。

4　おわりに

本稿では、作業教示映像を対象として、その言語解析と、言語情報と映像情報の統合による物体モデルの自動学習について述べた。今後、得られた物体モデルを用いて物体認識を行ない、映像にインデキシングし、省略解析・談話構造解析といった言語解析と統合する予定である。

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*, pp. 6–11, 8 2002.

[2] 河原大輔, 黒橋禎夫. 用言と直前の格要素の組を単位とする格フレームの自動構築. 自然言語処理, Vol. 9, No. 1, 2002.

[3] 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.