言語情報と映像情報の統合による作業教示映像の構造的理解

柴田 知秀　　黒橋 禎夫

東京大学大学院 〒113-8656 東京都文京区本郷 7-3-1

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

あらまし　実世界情報、映像情報などの高度な利用のためには、その内容の構造的理解が必要であり、そのためには、話し言葉を現場を含めた広い文脈の中で正確に解釈することが重要になる。本稿では、作業教示映像、具体的には料理番組映像を対象として、まず、その発話（クローズドキャプション）の言語解析を行ない、作業構造の抽出を行なった。次に、言語情報と映像情報の統合的処理によるトピックの推定と物体のモデル学習を行なった。

キーワード　自動インデキシング、談話構造、トピック推定、物体モデルの自動学習

Structural Analysis of Instruction Videos by Integrating Linguistic and Visual Information

Tomohide SHIBATA and Sadao KUROHASHI

The University of Tokyo　7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656 Japan

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

Abstract　In realizing flexible utilization/access of real-word information or video contents, the crucial point is the structural analysis of their contents, which requires the interpretation of utterances based on wider contexts including the scene. This paper describes a method for analyzing cooking instruction utterances (closed caption texts) and extracting action structure. By integrating linguitic and image analysis, we perform an unsupervised topic identification and acquire an object model automatically.

Keyword　Automatic Indexing, Discourse Structure, Topic Identification, Automatic Object-model Acquisition

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(インデキシング)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。

このような背景のもと、映像解析の分野ではショットやカメラワーク、顔の検出などが行なわれてきたが、意味的インデキシングにまでは踏み込めていない。意味内容を考慮したインデキシングを行なうには、映像中の登場人物・ナレーターの発話や、スポーツ映像における実況コメントなどといった言語情報を正確に解釈することが必要となる。

近年、テキストを中心とする自然言語処理研究は大規模コーパスの利用などにより急速に進展している。これに伴い、これまで新聞記事などの書き言葉を対象としてきたが、Web テキストや話し言葉などといったくだけたテキストの解析も可能となってきた。

このような問題意識から、我々は映像情報中の話し言葉の解析を行っている。本稿では、再利用価値の高い、料理、園芸、工作などを説明する作業教示映像、特に料理教示映像を扱う。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。

まず、作業教示発話(クローズドキャプション)の言語解析について述べる[1]。次に、言語情報と映像情報を統合することにより、トピック(下ごしらえ、炒める、盛り付けなど)を推定する手法について述べる。そして、言語処理と映像処理を統合し、物体のモデルを教師なし学習する手法を述べる。

2. 作業教示発話の言語解析

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、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、話し言葉では頻繁に省略がおこり、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。

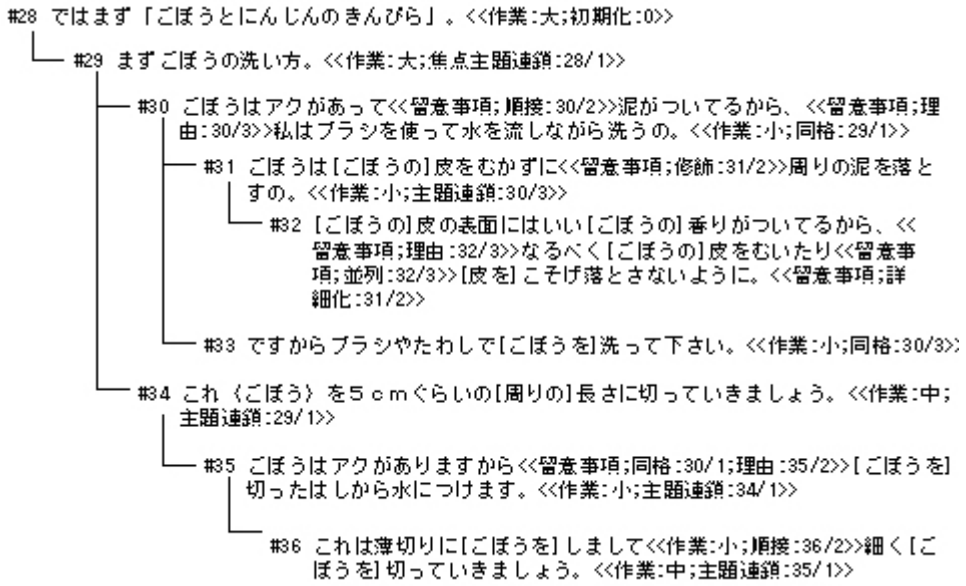

図 1　料理教示発話の言語解析

このような話し言葉の情報に対して、単純に単語マッチングのような検索を行っても、意図する映像を正確に取り出すことは難しい。

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。これは、作業教示発話においては、作業の構造を反映したものとなっており、映像の要約や、検索システムのための構造化されたインデックスとして利用可能なものとなる。

2.1. 格フレームの自動構築と関係解析

言語表現中には省略されている関係が多数存在する。図 1 の例では「皮」は「ごぼうの皮」であり「切る」は「ごぼうを切る」である。このような省略は、書き言葉にもあるが､話し言葉ではより頻繁におこる。このような関係を計算機によって検出するためには、「皮」というものは｢動物や植物の皮｣であり、｢切る｣ときには「何を切る」ということといった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。この格フレームを人手で与えることはとてもできないので、国語辞典やコーパス（大規模なテキスト集合）からうまくクラスタリングを行なうことにより格フレームを自動構築した[2, 3]。以下に自動構築した動詞格フレームの例を示す。ここでは、コーパスとして Web から「料理」などのキーワードによって収集した 1200 万文を用いた。

切る(1): {[主体]}が{豚肉,大根,こんにゃく‥}を{正方形,櫛形,三角形}に切る

切る(2): {[主体]}が｛なす,豆腐,肉‥｝の{水気,水分,汁気‥}を切る

入れる(1): {[主体]}が{塩,油,野菜‥}を{鍋,ボール,容器‥}に入れる

入れる(2): {[主体]}が{魚,腹,付け根‥}に{包丁}を入れる

この例が示すように、クラスタリングすることにより、動詞の多義性が解消されている。

このような格フレームを用いて、まず、必須的な要素が欠けていることを認識し、次に、欠けている要素と類似した語を文脈中から探し出す。語と語の類似性については、シソーラスの木構造における語と語の距離から計算することができる。

2.2. 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。これらのタイプを正確に認識しておくことは談話構造の解析のためにも、また検索のためのインデキシング、映像との対応付けの際に重要となる。

発話のタイプは節ごとに考える。以下に発話のタイプと例を示す。

- **作業:大**
  さ、では、ステーキの材料にかかります。
- **作業:中**
  強火で油を温めましょう。
- **作業:小**
  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。
- **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。

- **代替可**
  手で搾って頂いても結構です。
- **食品・道具提示**
  材料は、牛ひき肉、百五十グラムです。
- **雑談**
  暑くなってきましたね。

作業:大、作業:中、食品・道具提示、代替可、留意事項、雑談については節末の表現のパターンを記述することで認識することができる。例えば、代替可については「～しても結構です／構いません／よい」など、留意事項については「～できます」「～しやすいです」「～を目安にしてください」などのパターンである。作業:小、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方法も考えられるが、他のドメインへの移植性を考え、自動詞、形容詞＋「なる」などを料理状態、それ以外を作業:小とする一般的な規則を用いている。

2.3. 談話構造解析

談話構造のモデルとして、前節で述べた節を一つのノードとし、関係するノードがリンクされたグラフ構造を考える。節間の関係として以下に示すものを考える。

- 一文内における節の係り受け関係
- 主節(「～と思う」などといった節を除いた、一文で最後の節)間の関係

まず、一文内で係り受け関係にある節間の結束関係を決定する。付与する関係は、順接 (～て、(連用形))、並列(構文解析器 KNP で並列構造とされたもの)、理由(～から、～ので)、条件 (～と、～たら)などであり、それぞれ括弧内に示すような表層パターンにより決定される。

次に、主節間の関係を求める。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。初期化以外の主節間の結束関係としては、並列、対比、理由、条件、主題連鎖、焦点主題連鎖、詳細化、理由、原因結果、例示、質問応答などの関係を考える。主節間の関係は、種々の表層的手がかりをもとに、各入力文に対して、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]。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、最終的に最も高い確信度を得た関係を採用する。確信度は表 1 に示すようなルールにより決定される。

表 1　談話構造解析のルール

接続可能文	入力文	適用範囲	結束関係	スコア
～	それでは～	1	初期化	10
～	そして～	1	並列	5
Xは～	X'は～	*	対比	20
～	〈料理状態〉	1	詳細化	15

表 1 において、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、発話タイプ(<>で括られたもの) などのパターン、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。

談話構造解析の結果の具体例は図 1 に示したものである。

次節から言語情報と映像情報の統合処理について述べる。

3. 隠れマルコフモデルによるトピック推定

はじめに述べたとおり、大規模映像に対して検索・要約を行なうには、自動でインデキシングする技術が必要となる。ここでは、料理映像に対して、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について述べる。例えば図 2 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次節で述べる物体モデルの学習に利用する。

作業教示映像の場合、トピックを推定する際に「野菜を切る」、「火をつける」「のせる」といった作業に関する発話が有用である。すなわち、「野菜を切る」から”下ごしらえ”、「火をつける」から”炒める”といったトピックを推定することができる。本研究では、この情報に加えて、映像情報も利用することにより、トピックの推定を頑健に行なう。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。

これらの特徴量を利用し、トピック推定を隠れマルコフモデル(HMM)を用いてモデル化を行なう。

3.1. 利用する特徴量

トピックの推定に以下の特徴量を利用する。

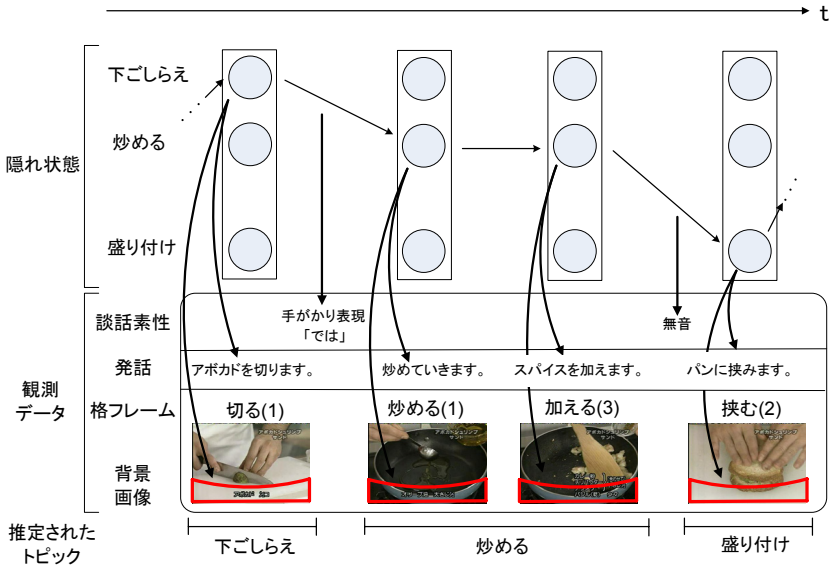

図 2　隠れマルコフモデルによるトピック推定

・　言語情報

- 作業に関する発話(格フレーム): 2.2 節で述べた発話タイプのうち、トピック推定には作業に関するものが有用であり、ほかの発話はノイズとなってしまう。発話タイプが作業と認識された節から、2.1 節で述べた、格・省略解析時に対応付けられた格フレームを抽出する。これは汎化を行ないつつ、多義性解消を行なうためである。
- 手がかり表現: 「では」「次は」「そうしましたら」などといった表現は、トピックが変化したことを示す手がかりとする。
- 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。
- 用言の一致: 用言の原型が一致する場合、トピックが同一である可能性が高い。

・　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[5]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 2 に示すように、画面下部の RGB の重心を特徴量とする。

・　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。

3.2. HMM によるトピック推定

隠れ状態がトピックにあたり、前節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 2)。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。本研究では下ごしらえ、蒸す、ゆでる、揚げる、煮る、炒める、盛り付け、その他の 8 種類を考える。HMM のパラメータを以下にあげる。

・　初期状態確率　pi_i

・　状態遷移確率　a_{ij}　: 状態 i から状態 j への遷移確率。

- 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。
- 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。

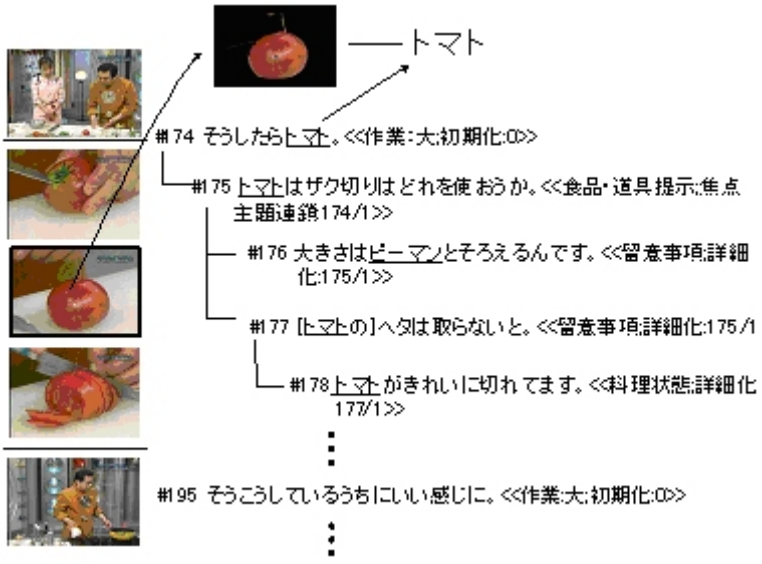

図 3　アップ画像とキーワードのペアの収集

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、その発話のトピックを固定する。図 2 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。

3.3. 実験

2 つの料理番組、NHK の「きょうの料理」と 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。2 番組においてモデルを学習し、学習されたモデルをそれぞれの番組 5 日分に適用して実験を行ない、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、評価した。表 2 に実験結果を示す。

表 2 トピック推定の実験結果

素性				きょうの料理	キューピー3分クッキング
格フレーム	背景画像	談話素性	無音区間		
○				61.7%	66.4%
	○			56.8%	72.9%
○	○			69.9%	77.1%
○	○	○		**70.5%**	**82.9%**
○	○	○	○	70.5%	82.9%

言語情報(格フレーム)に加えて映像情報(背景画像)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談話素性(手がかり表現、語連鎖、用言の一致)を利用することにより精度が向上した。無音区間を加えても精度が向上しなかったのは、クローズドキャプションと映像にはずれがあるためであると考えられ、音声認識技術が向上すれば音声認識結果を利用することを考えており、そうすれば解決すると思われる。

4. 物体モデルの自動学習

2 章で述べた関係解析や談話構造解析といった言語処理の誤りは映像中の情報を利用することにより解消される可能性が高いと考えられる。しかし、映像中からそこに映っている物体を取り出すことは簡単ではない。限定された物体に対してモデルを作り込めばできるだろうが、それはスケーラビリティのある方法ではない。

そこで、物体の認識を行うために、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モデルを自動学習する問題を考える。その第一段階として、ここではまず物体の色情報(RGB)を学習する。

モデルを学習するために、物体が大写しになっている画像とキーワードのペアを収集する。以下では、大写しになっている画像をアップ画像、アップ画像で最も焦点のあたっている領域を注目領域と呼ぶ。

4.1. アップ画像とキーワードのペアの収集

単純に動画像列から画像を切り出し、その時刻に近い発話から抜き出した名詞を対応付けてもよい学習データを得ることはできない。そこで、画像処理と、発話を談話構造解析し重要な単語を抽出する処理を行なうことにより、アップ画像とキーワードのペアだけを収集する。解析の概要を図 3 に示す。

画像とキーワードの対応付けは、ある瞬間の画像とその時の発話で行なうのではなく、ショットと談話構造木といったある程度広い範囲同志で行なう。それは、省略が多いことや、発話には作業の説明だけでなくコツや雑談などといった様々なタイプがあるため、きちんと対応がとれないためである。例えば、図 3 の例で、

トマトの画像に一番近いものを対応付けると、トマトの画像と「ピーマン」を対応付けてしまうことになる。

4.1.1. エッジ処理によるアップ画像の判定

以下の画像処理により、アップ画像を抽出し、そこから注目領域を抽出する。

1. ショットに分割

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であるところをカット点とし、ショット単位に分割する。

2. エッジ抽出によるアップ画像の抽出

すべての画像に対して、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)を計算する。ショット内で最もエッジ率の小さいものを選び、エッジ率が閾値(0.5)を下回った場合、その画像をショットの代表とする。この処理により、食材が複数映っている画像や、人が映っている画像を除外することができる。

3. RGB 空間への写像と極大点の探索

エッジ処理によりアップと判定された画像について、各画素を RGB3 次元空間に写像する。その後、平滑化を行い、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。

4. 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、極大値の大きさを下式のように重み付けし、最もスコアの小さいものを注目領域として選ぶ。

4.1.2. キーワードの抽出

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、前節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る。

次に、トピックが下ごしらえと推定された談話構造木から、そこで最も重要な単語をキーワードとする。シソーラスを用いて食材タグのふられた名詞に対して、談話構造解析結果に基づき、スコア付けを行ない、最もスコアの高いものを選ぶ。このスコア付けは、作業の説明を行なっている発話には重要な食材名がくる可能性が高いことや、談話構造木の最初の方の発話が重要であるなどといったことを反映したものである。

以上の処理によって得られたアップ画像とキーワードのペアにおいて、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 を物体モデルとする。

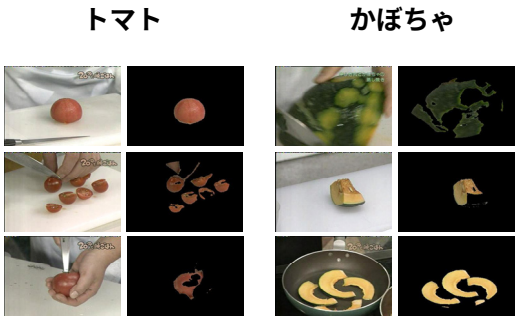

図 4 アップ画像とキーワードのペア

4.2. 実験

NHK の「きょうの料理」の映像約 2 年分を用いて実験を行なった。物体モデルの学習結果を表 3 に示す。

表 3 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)

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、精度は 68.5%であった。このドメインで頻繁に出現する食材に対してはモデルが学習されており、提案手法の有効性を確認できた。なお、トピックの推定を行わなかった場合は精度が 54.8%であり、このタスクでトピック推定が有効であることが示された。

5. おわりに

本稿では、まず、料理教示発話の言語解析について述べ、次に、言語情報と映像情報を統合し、HMM を用いたトピックの推定、物体のモデル学習について述べた。今後は、学習された物体モデルを用いて物体認識を行い、その結果と関係解析、談話構造解析といった言語解析と統合する予定である。

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, 62, 1, pp.117-122(2004)

[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, 9, 1, pp.3-19(2002)

[3] 笹野遼平, 河原大輔, 黒橋禎夫：“名詞格フレーム辞書の自動構築とそれを用いた名詞句の関係解析”, 自然言語処理, 12, 3, pp.129-144(2005)

[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, 1, 1, pp.3-20(1994)

[5] 浜田玲子, 井手一郎, 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”, 信学論, J85-D-II1, 1, pp.79-89(2002)