レシピテキストと調理映像からの実世界理解に向けて

森 信介　　舩冨 卓哉
京都大学学術情報メディアセンター

1　はじめに

言語の基本的な機能として、実世界の事象の記述がある。これまで、自然言語解析として解析結果とテストコーパスの正解との比較が行われてきた。しかし、言語だけに閉じた取り組みがコンピューターによる言語の理解に本当に向かっているのか甚だ疑問である。本報告では、調理手順指示文書 (レシピ) とそれを実施している映像を対象とし、各言語表現と映像中の物体との対応を推定するシステムについて述べる。

レシピを題材とする理由は、言語処理の観点からは主観や時制などの問題がほとんどないなど文が比較的単純であり、また映像処理の観点からは制御下にある同じ場所で動作が進行していき、収録と処理が容易であることである。

レシピテキストと調理映像の対応付けの研究はすでにいくつかある [6, 16, 12]。しかしながら、これらの研究では、レシピテキストの処理が既存のツールの単純利用であったり、映像による物体や行動の認識もそもそも行われていないなど、初期の取り組みの域を出ない。

本研究では、まず、レシピテキストに対する言語処理と調理映像の映像処理の双方において、学習データのアノテーション方法を工夫することで、一般性を保ちながら高い精度を実現する。実際、単語分割と固有表現認識と係り受け解析に関しては、既に十分な大きさのテストコーパスと小規模な学習コーパスを準備し、これを実現済みである。本稿では、その効果を実験結果を交えて報告する。さらに、これらのレシピコーパスに対して、述語項構造を付与する。また、学習のための部分的アノテーションコーパスを準備する。さらに得られた述語項構造から、固有表現をノードとする有向グラフ (図 1 参照) を出力するシステムを構築する。

映像処理においも同様に、映像中の一部の物体にのみ領域が指定され、レシピの固有表現へのリンクが付与されたデータ (図 2 参照) を作成し、物体や動作の認識精度を高める。さらに、映像が含む時間情報から推定される物体の把持の順序や外見の変化を推定し、食材と動作をノードとする有向グラフを自動構築する。

最後に、言語処理から得られた有向グラフと映像処理から得られた有向グラフの最大マッチを計算することにより、固有表現 (食材、道具、動作) と映像の領域の対応をとる。まずは、双方で独立に最適な有向グラフを構築してマッチングを試みる。次いで、一方または両方の出力を確率とともに複数列挙し、確率値も考慮してマッチすることで、言語と映像の認識精度が相補的に高まることを示す。例えば、映像処理では動作の認識が容易ではないが、食材映像の認識結果とレシピテキストに書かれた動詞から映像中での動作の認識精度が向上することなどが想定される。

本研究を通じて、レシピテキストと対応する調理映像にアノテーションを行う。また、言語・映像のそれぞれの処理の高精度化のために学習コーパスを準備する。レシピテキストには著作性がないので、本文の再配布に問題がない。したがって、本文とアノテーションの両方を公開する。調理映像は、独自に実施・収集し、レシピとの対応も含めて公開する。これにより、我々を含めた様々な研究者が、各処理において学習コーパスの量や手法が異なる様々なシステムを構築することができる。それらを組み合わせて全体の精度を比較することで、どの処理が問題なのか、それをアノテーションで解決すべきか手法の改良で解決すべきかなどの定量的議論が可能になる。このように、本研究の取り組みは言語と映像による実世界理解のモデルケースとなり、コーパスと併せて今後の研究に資する。

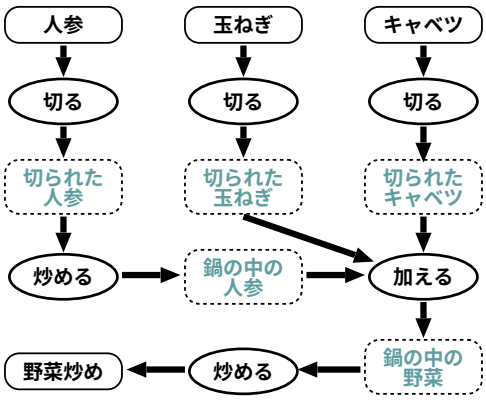

図 1: レシピからの有向グラフ

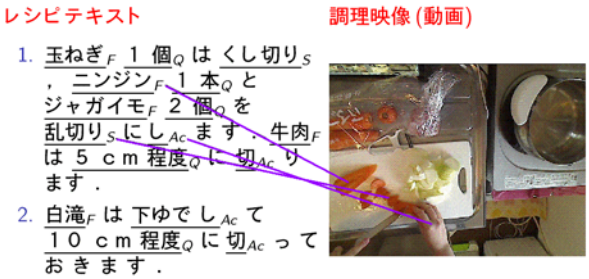

図 2: 映像へのアノテーション

2 レシピに対する言語処理

レシピに対する言語処理の入力は文章であり、出力は食材と動作をノードとする有向グラフである。この節では、すでに実施済みの単語分割と固有表現認識と係り受け解析について説明する。

2.1 レシピの言語的性質

レシピは、ある料理の食材のリストと調理動作の指示文（テキスト）からなる。言語処理は、テキストを対象とする。

指示文は、比較的短いいくつかの文からなり、調理者が取るべき動作や食材の状態変化などが記述されている。1 文が 2 つ以上の動作や状態変化を含むこともある。言語処理の観点からは、テキストは、主観や時制などの問題がほとんどなく、比較的単純である。

一方で、用いられる単語や表現が独特であり、多くの言語処理ツールが学習に用いている新聞記事などとは大きく異なる。したがって、既存のツールの単純な適用では精度が低く、分野適応が必須である [9]。

2.2 単語分割

最初の処理は、文中の単語の確定である。単語の定義は、国立国語研究所の短単位 [7] とした。ただし、活用語は語尾と語幹を分割している。これにより、活用形の異なる活用語の同一性の判定を単なる文字列比較によって実現できる[1]。

単語分割の入力は次のような文である。

> 水４００ｃｃを鍋で煮立て、沸騰したら中華スープの素を加えてよく溶かす。

出力は、単語列である。

> 水|４−０−０|ｃ−ｃ|を|鍋|で|煮−立−て|、|沸−騰|し|た−ら|中−華|ス−−−プ|の|素|を|加−え|て|よ−く|溶−か|す|。

ここで、文字間の記号「|」は、文字間に単語境界があることを示し、「−」はないことを示す。

レシピテキストの単語分割を高い精度で実現するために、点予測による方法 (KyTea) [10] を採用する。この理由は、部分的アノテーションコーパスが利用可能で、分野適応が容易であることである。すなわち、以下の例が示すように、レシピに特有の単語や表現の周辺にのみ単語境界情報を付与したコーパスから単語分割器を学習できることである。

> 水␣４␣０␣０␣ｃ␣ｃ␣を|鍋|で|煮−立−て|る

ここで、「␣」は、単語境界情報が付与されていないことを意味する。

KyTea では、単語分割を 2 値分類問題として定式化し、線形 SVM を用いて各文字間の単語境界の有無を判定する。このとき、周辺の文字 n-gram と文字種 n-gram、および周辺の文字列が辞書に含まれるかを素性として参照する。

2.3 固有表現認識

実世界で 1 つの物体や動作となる表現 (固有表現) は、1 単語であるとは限らない。したがって、単語分割に次いで、これら複数単語からなる固有表現を認識する必要がある。固有表現の種類は完全に分野依存である。レシピテキストに対して、次の固有表現の種類を設定した。

> 食材 (F), 量 (Q), 道具 (T), 継続時間 (D),
> 食材の状態 (S), 調理者の動作 (Ac),
> 食材の動作 (Af)

既存の固有表現認識では、IOB2 タグ体系を用いて各単語のラベル推定の問題として定式化する。すなわち、各固有表現タグには、開始を表す B と継続もしくは終了を表す I が付加される。さらに、固有表現の一部ではない単語のラベルとして O を導入する。以上から、タグセットは $\mathcal{T} = \{\mathsf{F}, \mathsf{Q}, \mathsf{T}, \mathsf{D}, \mathsf{S}, \mathsf{Ac}, \mathsf{Af}\} \times \{\mathsf{B}, \mathsf{I}\} \cup \{\mathsf{O}\}$ となる。例えば、以下のタグアノテーションは、単語「水」が食材で、単語列「４００ ｃｃ」が量であり、単語「を」は固有表現ではないことを表す。

> 水/F-B ４００/Q-B ｃｃ/Q-I を/O

固有表現認識についても、部分的アノテーションからの学習を目的として、点予測による方法を採用する。

[1] 異なる動詞が語幹の文字列を共有する例が少数ながらある。例えば、「行う」と「行く」の過去形は「行った」である。

		単語 w				
$P(y\|w)$		水	４００	ｃｃ	を	$\cdots$
	F-B	0.62	0.00	0.00	0.00	$\cdots$
	F-I	0.37	0.00	0.00	0.00	$\cdots$
タ	Q-B	0.00	0.82	0.01	0.00	$\cdots$
グ	Q-I	0.00	0.17	0.99	0.00	$\cdots$
y	T-B	0.00	0.00	0.00	0.00	$\cdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$
	O	0.01	0.01	0.00	1.00	

図 3: 固有表現認識における最適パス探索のためのヴィテルビ (Viterbi) テーブル

まず、BIO2 タグ形式のコーパスから単語ごとの固有表現タグを推定するロジスティック回帰を構築しておく。解析時には、入力の単語列の各単語に対して可能な全ての固有表現タグとその確率を出力し、ヴィテルビ (Viterbi) テーブル (図 3 参照) を得る。一部のタグ列は、元の固有表現列に復元できない[2]。したがって、固有表現列に復元し得るタグ列の中で、確率の積が最大となるタグ列を探索し出力する。

2.4 係り受け解析

固有表現間の統語的関係を明らかにするために、単語単位の係り受け解析を行う。係り受け解析においても、部分的アノテーションによる分野適応を実現するために、点予測による方法 (EDA) [1] を採用する。

EDA は、最大全域木による係り受け解析 [3] の一種である。既存の最大全域木による係り受け解析との違いは、単語間の係り受けスコアの推定に際して、周辺の係り受けを参照しないことである。これにより、文中の一部の単語にのみ係り先が付与された部分的アノテーションコーパスから学習することが可能になる。

i 番目の単語の係り先が j 番目の単語である場合のスコアは、以下の式で計算される。

$$\sigma(i \to j|\boldsymbol{w}, \boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\theta} \cdot \boldsymbol{\phi}(\boldsymbol{w}, j)\right)}{\sum_{j' \in \mathcal{J}} \exp\left(\boldsymbol{\theta} \cdot \boldsymbol{\phi}(\boldsymbol{w}, j')\right)}$$

ここで、$\boldsymbol{w}$ は入力の単語列であり、$\boldsymbol{\theta}$ は重みベクトルである。また、$\boldsymbol{\phi}$ は素性ベクトルである。素性ベクトルは、係り元と係り先の単語の距離と係り元の前後の単語の表記や品詞、および係り先の前後の単語の表記や品詞からなる。

全ての単語の組に対して係り受けスコアを列挙した後、以下のように、スコアの合計が最大となる係り受けの列を計算し、係り受け解析の結果として出力する。

$$\hat{\boldsymbol{d}} = \underset{\boldsymbol{d}=(d_1,d_2,\cdots,d_h)}{\mathbf{argmax}} \sum_{i=1}^{h} \sigma(i \to d_i|\boldsymbol{w}, \boldsymbol{\theta})$$

ここで、h は入力文の単語数である。

2.5 述語項構造解析

以上で説明した 3 つの処理の結果、入力文は単語の係り受けとなっており、いくつかの単語列は固有表現としてまとめられている。このような係り受け木から述語項構造を以下の規則にって抽出する。

1. タグが Ac または Af の固有表現を見つける。

   煮立て/Ac

2. 係り受け木を辿り、その述語に係る固有表現を項とする。多くは、格助詞を介して接続している。

   /水/F /４００ ｃｃ/Q を
   /鍋/T で

3. 以上で確定された述語と項から述語項構造を構成する。その際、意味役割を明示するために、格助詞がある場合はそれを項に含める。

   煮立て (を:水-４００-ｃｃ, で:鍋)

上記の規則では、間接的に項を示すゼロ照応や使役や関係詞節などの現象に対応できない。これは、今後の研究において実現する。

3 評価

前節で述べた単語分割と固有表現認識と係り受け解析については、テストコーパスと少量ながら部分的アノテーションコーパスを準備し、評価実験を行った。この節では、それぞれの言語処理の個別の精度について報告する。また、全体の評価として、述語項構造抽出の精度についても報告する。

3.1 実験条件

一般分野に関しては、前節で述べた言語処理のための言語資源はあるが、レシピテキストの学習コーパスはない。そこで、一般分野のコーパスにレシピ分野の学習コーパスを加えることで精度向上を図ることとした。なお、固有表現認識は、クラス設定がタスクに大きく依存するので、新たに 2.3 項で述べた固有表現クラスを設定し、コーパスが全くない状態から開始した。

[2]例えば、「F-B S-I」が、不正な BIO2 タグ列である。

表 1: フルアノテーションコーパス

用途	出典	文数	単語数	文字数	固有表現数	係り受け数
学習	BCCWJ	53,899	1,275,135	1,834,784	–	–
	レシピ	242	4,704	7,023	1,523	–
	辞書の例文	11,700	147,809	197,941	–	136,109
	新聞記事	9,023	263,425	398,569	–	254,402
テスト	レシピ	724	13,150	19,966	3,797	12,426

BCCWJ: 現代日本語書き言葉均衡コーパス [5]

煮立て (頻度=1,497)
中␣火␣で|煮-立-て|、␣(␣1␣)␣の␣ほ␣う␣れ␣ん ···
A␣を|煮-立-て|、␣(␣1␣)␣の␣し␣い␣た␣け␣・ ···
鍋␣に␣B␣を␣加␣え|煮-立-て|る␣。

図 4: 未知語候補の KWIC のチェックによって得られる部分的単語分割コーパス (文字間の記号については 2.2 項参照)

係り受け解析では、一般分野として辞書の例文と日経新聞の記事を用いた。これらの文には、単語境界情報と単語間の係り受けが付与されている。

表 1 は、フルアノテーションコーパスの諸元である。これらに加えて、レシピ分野の部分的アノテーションコーパスを準備した。詳細は、以下のそれぞれの処理の評価において述べる。テストコーパスは、クックパッドのサイトから無作為抽出した 100 のレシピである。

3.2 単語分割

単語分割には、日本語テキスト処理ツール KyTea[3] [10] を用いた。一般分野の言語資源として、現代日本語書き言葉均衡コーパス (BCCWJ) [5] と辞書の例文と新聞記事を用いた。一般分野のモデルは、このコーパスと UniDic から構築した[4]。さらに、アラビア数字や姓名や記号の辞書も用いた。辞書の総見出し語数は 423,489 であった。

分野適応のために、まず、大規模なレシピの生テキストから分布分析 [11] によって未知語候補を抽出した。次に、抽出された単語候補の KWIC (図 4 参照) を作業者に提示し、作業者はこれらがその文脈で単語であるか判断し、必要に応じて単語境界情報を修正した。総作業時間は 8 時間であった。各 1 時間の作業の後に単語分割精度を測定した。

[3] http://www.phontron.com/kytea/ から利用可能 (2012 年 6 月にアクセス)。
[4] http://www.tokuteicorpus.jp/dist/ から利用可能 (2012 年 5 月にアクセス)。

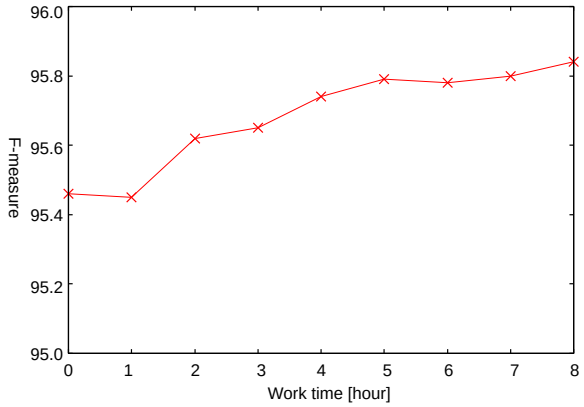

図 5: 単語分割の学習曲線

評価基準は、適合率と再現率の調和平均である F 値である [8]。

図 5 は、自動単語分割の分野適応の学習曲線である。一般分野のモデルによるレシピに対する単語分割精度は、一般分野のテストデータに対する精度 (98.13% [10]) よりも低い。未知語のチェック作業によって得られる部分的アノテーションコーパスを学習に加えることで精度が上昇していくことが分かる。グラフから、まだ上昇は飽和しておらず、学習コーパスの追加が依然として効果的であることが分かる。

3.3 固有表現認識

固有表現認識では、2.3 項で述べた固有表現クラスを独自に設定し、タグ付きコーパスを作成した (表 1 参照)。作業時間は 5 時間であった。その上で、学習コーパスのサイズを 1/10 から 10/10 に変化させて精度を測定した。

評価尺度は F 値 (適合率と再現率の調和平均) である。適合率は、システムの出力に対する適切に認識された固有表現の割合である。再現率は、コーパスに付与された正解に対する適切に認識された固有表現の割合である。

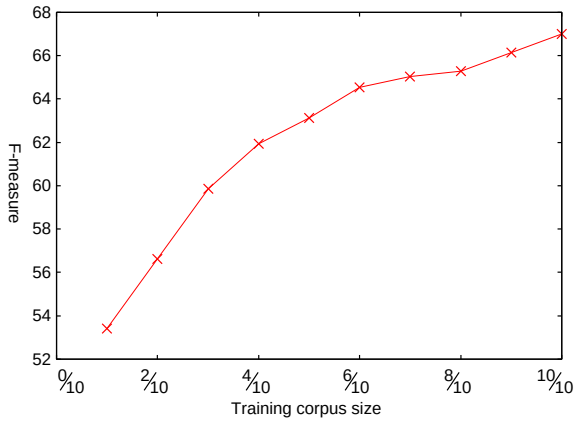

図 6: 固有表現認識の学習曲線

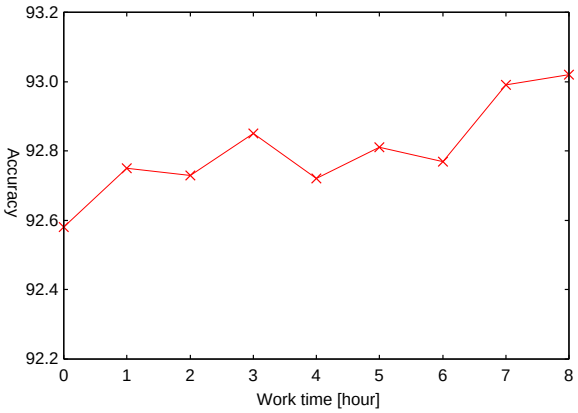

図 7: 係り受け解析の学習曲線

図 6 は、固有表現認識の学習曲線である。学習コーパスの 1/10 のみを使う場合の精度は非常に低いが、学習コーパスの増加に従って精度が大幅に向上していく。しかしながら、全ての学習コーパスを使っても一般分野において報告されている 80 程度の F 値 [4] には及ばない。この報告での学習コーパスのサイズは約 12,000 文であり、直近の課題は、手法の改善ではなくコーパスの増量であるといえる。

3.4 係り受け解析

係り受け解析には、EDA[5][1] を用いた。EDA は、各単語に品詞が付与されていることを前提とするが、学習コーパスには品詞が付与されていない。したがって、BCCWJ から推定された KyTea を用いて品詞を推定した。ベースの解析器は、このようにして品詞が付与された辞書の例文と新聞記事から構築した。

レシピテキストへの分野適応には、一般分野の学習コーパスにない名詞と助詞からなる列に対してのみ係り受けを付与することで得られる部分的アノテーションコーパスを用いた。総作業時間は 8 時間であった。各 1 時間の作業の後に係り受け解析の精度を測定した。

評価基準は、係り先が適切に推定された単語の割合である。ただし、文末の単語は評価から除外している。

図 7 は、係り受け解析の学習曲線である。一般分野のコーパスのみから学習したモデルによる結果 (最も左の点) は、同一分野に対する精度 (約 97%) [1] よりも有意に低い。この理由は、学習コーパスとテストコーパスのドメインの違いであろう。学習曲線から、部分的アノテーションコーパスの追加により精度が上昇していくことが分かる。アノテーション作業を継続することで容易に精度向上が実現できると言える。

[5] http://plata.ar.media.kyoto-u.ac.jp/tool/EDA/ から利用可能 (2012 年 6 月アクセス)。

3.5 述語項構造抽出

最後に、全体の評価として述語と項の組の抽出精度を評価した。この実験の入力は文である。入力文は、自動で単語に分割され、固有表現にタグが付与され、単語単位の係り受け情報が付与される。固有表現認識と同様に、評価基準は F 値である。唯一の違いは認識すべき単位で、ここでは、述語と項の組である。例えば 2 節の最後に示した例文では、以下の 2 つが認識すべき対象である。

⟨煮立て, を:水-４００-ｃｃ⟩, ⟨煮立て, で:鍋⟩

認識した述語と項の組が正解であるのは、述語と項のそれぞれが正解で、さらに格助詞の有無とある場合は、その表記も一致している必要がある。

分野適応以前の各処理をつなげたシステムの F 値は 42.01 であった。各処理を分野適応することにより、F 値は 58.27 となった。分野適応以前のシステムの精度は低いが、コーパスアノテーションによる分野適応の結果、約 28.0%の誤りが解消された。しかし、分野適応後の精度は、後段の処理にとって依然として不十分であろう。本論文での合計の作業時間は 21 時間 (8+5+8) とまだまだ少ない。各処理の作業追加することでさらなる精度向上を図ることができる。

本論文で採用している各処理の枠組みにより、レシピ分野に特有の表現や単語にアノテーション作業を集中することができ、効率的に全体のシステムの精度向上を図ることができる。各処理の学習曲線 (図 5, 6, 7) から、当面は固有表現認識のコーパス作成に注力するのがよい戦略であることが分かる。このように、各処理のためのアノテーションの効率化のみならず、各時点でどの処理のための学習コーパスの作成に注力すべきかを見極めることも重要である。

4　言語処理の研究計画

前節まで、言語処理の枠組みと学習コーパスの作成の進捗を報告した。本節では、固有表現認識の改善と有向グラフを出力するための残りの言語処理について説明する。

4.1　固有表現認識の改善

現在の固有表現認識は、各単語のタグをロジスティック回帰を用いて推定し、確率の積が最大になるタグ系列を動的計画法によって算出している (図 3 参照)。

先行研究により、条件付き確率場を用いる方法がより高い精度となることが分かっているので、動的計画法による解探索を条件付き確率場を用いる方法に置き換える。このとき、既存手法では部分的アノテーションコーパスからの学習が困難になるので、まず部分的アノテーションコーパスを含む全ての学習コーパスから推定したロジスティック回帰で各単語のタグを推定し、それを素性としフルアノテーションコーパスのみから推定した条件付き確率場を用いて最適なタグ系列を算出することとする。

4.2　機械学習による述語項構造抽出

現在の述語項構造抽出は、簡単な規則によって実現している。これを機械学習による方法で実現する。その際に、ゼロ代名詞などの係り受け関係にない項も抽出できるようにする。モデル構築のために、固有表現認識のためのコーパス作成作業で動作タグ (Af と Ac) が付与された動詞に対して、さらに項をアノテーションする。機械学習による述語項構造の抽出でも、迅速かつ安価な分野適応を実現するために、一部の動詞にのみアノテーションしてある部分的アノテーションコーパスが活用できるように設計する。

4.3　有向グラフへの変換

述語項構造は、基本的に物体を表す幾つかのノード (図 1 の四角) と動作を表すノード (図 1 の楕円) が接続された部分グラフに対応する。ある動作の結果得られる物体はテキストに明示されないことが多い。言語処理ではこれを指示対象として認識し、動作の先にある物体のノード (図 1 の破線の四角) とする。このようにして得られる部分グラフから、全体の有向グラフを生成する。

4.4　各処理の学習コーパスの充実

ここまでで述べたように、各処理は機械学習に基づいており、さらに学習コーパスとして部分的アノテーションコーパスが利用可能である。各処理に対し一般分野と同等の精度となる量の学習コーパスを準備する。これを用いて、全体の精度向上のためにどの処理に注力すべきかというアノテーション戦略を論じる。

さらに、様々な既存手法などを実装し、各処理に複数の方法を用意する。それらと各処理の学習コーパスを用いて、ある処理の手法の改善による効果と学習コーパスの増量による効果を定量的に比較できるようにする。

5　映像処理の研究計画

我々はこれまで、映像からの物体や動作・認識の研究の事例として、調理映像の収録とその認識の研究を行ってきた。図 2 の右の写真はその画像の例である。本節では、まずこれまでに開発した基礎となる技術要素について説明する。次に、それらを活かし、調理作業の映像において物体と動作に着目した処理を行うことで、有向グラフを出力する研究の計画について述べる。

5.1　調理映像に対するアノテーションの予備検討

これまでに我々は、映像中で観測される食材とそれが受ける調理動作の履歴に着目し、調理行動のモデル化を目的としたアノテーションを試みた [14]。この試みでは、食材・調理器具といった映像中に登場する物体とそれに対する調理動作にアノテーションを行った。しかしながら、全てのアノテーションを手作業で行ったため構築できたデータの量は限られたが、大量データ処理に向けてアノテーションを半自動化しようとした場合に起こる課題が明らかになった。

有向グラフの半自動生成における最大の問題は、レシピに現れる食材・動作ノードと映像に現れる食材・動作ノードの粒度が一致しないことである。例えば映像中において、1 種類の物体が 1 つの連結領域として現れるとは限らない。近接して置かれた異種の食材が 1 つの領域になることや、切削により分割された破片が複数の領域になることが頻発する。通常の画像処理によって映像中に現れた 1 つの連結領域を 1 つのノードとして生成すると、レシピから生成された 1 つのノードが映像側の複数のノードに対応したり、逆に映像側

の 1 つのノードがレシピ側の複数のノードに対応したりするため、マッチが困難になる。[14] では厳密な領域指定は行わず、また完全に人手でアノテーションを行ったためにこのような問題は起こらなかったが、この処理を自動化するためには従来の画像処理では不十分なところが多い。

5.2 物体と動作へのアノテーション

本研究では、図 2 のように物体と動作にアノテーションを行う。映像中にはレシピや調理とは関係のない物体や動作が多数含まれる。そのためにまず重要箇所の領域の自動抽出が必要となる。領域がある程度自動抽出されれば、作業者はレシピの固有表現とのリンクをアノテーションするだけでよくなる (図 2 参照)。

この項ではその実現計画、および認識結果からの部分グラフの構築について述べる。

5.2.1 調理映像からの物体ノードの抽出

まず、調理を行っている様子を観測した場合には、調理者の動作に伴い光源環境も変動する。そのような状況でも撮影された映像から安定して物体領域を抽出するため、照明変動にロバストな物体領域抽出処理を行う [13]。

調理においてはさまざまな道具や食材が物体として出現するが、一般的に調理者は少数の必要な道具や食材を把持して調理手順を遂行していくため、他の道具、食材については触れられず置いておかれる。したがって、固定カメラの映像から得られる時間的継続性を考慮することで、仮に既にあった物体に近接して別の物体が置かれたとしても、これらを別々の画像領域として検出することができる。これにより、複数の食材が近接して置かれた際の問題に対処する。また、新たに物体が現れたり、物体が消滅したことを検出することにより、調理動作のために調理者が物体を把持したことを検出したり、解放された物体の領域と対応付けたりする [2]。これにより、調理者の手によって加工・移動された食材領域を追跡し、把持の前後で映像から生成された物体ノード間にエッジを張ることができると考えられる。これにより、1 種類の食材が複数のノードに分かれたものを、1 まとまりとして扱うための手がかりが得られる。

5.2.2 調理映像からの動作ノードの抽出

調理加工のために調理者に把持された食材は、手によって遮蔽されているため直接観測できない上に、切削などの加工を受けることでその外見が変化する。通常の映像処理技術では、このような状況下では加工の前後で安定して物体を追跡することは困難であるが、先と同様の仮定を設けることで、ある程度の精度で追跡を実現することができる。この際、外見が大きく変化したか否かの判定を行い、変化が起こった場合には何らかの加工動作が行われたと推定して、動作ノードを生成する。

また、複数の食材を混合したり加熱を行ったりするときに見られる「かき混ぜる」などの同定が困難な行動の認識にも取り組んでおり [15]、このような処理を併用することで、動作ノードの生成や複数種類の食材ノード間に対するエッジの生成も試みる。

5.2.3 有向グラフの出力

以上のように、これまで開発してきた技術を用いることで映像から道具や食材の領域、動作が起こった区間をノードとして抽出し、また調理者による食材の把持・移動に応じてノード間にエッジを張ることができる。

領域抽出、物体追跡、動作認識のそれぞれの技術は 100%の精度を達成しているわけではない。言語処理から得られる情報とのマッチングにより精度向上が実現できれば、非常に有意義である。

6 言語処理と映像処理の統合

言語処理と映像処理の結果、同じ調理作業を記述する有向グラフが得られる。これらは、大域的構造において類似しているはずであるが、それぞれにのみ含まれる情報がある。

テキストに対する言語処理の結果にのみ含まれる情報として主に以下が挙げられる。

- 調理動作のラベル (動詞): 映像中の動作は認識できるので、なんらかの調理動作が行われたことは分かるが、動作名を推定するのは非常に困難である。一方、レシピテキストには動作名が端的に表現されている。
- 最終的にできあがる料理の名前: 個々の食材や道具は、名前とともに学習コーパスにアノテーションされているが、料理映像は多種多様でカバー率は非常に低く、料理の名前の推定は困難である。
- 量: 映像中の面積や持続時間から絶対的な量は多少推定できるが、言語表現では相対的であり、適切な表現が推定できない。ただし、今回の提案で

は、レシピテキスト中の量を表す固有表現にアノテーションをするものの、映像とのマッチングは余力がある場合にのみ行う。

また、調理映像に対する処理の結果にのみ含まれる情報として主に以下が挙げられる。

- レシピに明示されない食材の中間状態: レシピでは、各調理動作後に得られる結果は、ほとんど外界照応として参照される。日本語ではゼロ代名詞になることが多く、この場合にはラベルが推定できない (図 1 の破線の四角)。
- レシピに明示されない物体: まな板など、切る動作が当然要求する物体は、レシピテキストには含まれない。このように、映像中にはレシピテキストにはないにもかかわらず出現する物体が多数含まれる。
- 随伴動作: 野菜を「洗う」などの準備のための動作やへたを「捨てる」などの片付けの動作などは、テキストに含まれない。

上述の相違に加えて、1 つの固有表現で表される動作が、映像上では 2 つ以上の動作からなるなどの相違がある。例として、人参の入った鍋にキャベツと玉ねぎを加える場合が挙げられる。映像中では、まずキャベツを加え、次いで玉ねぎを加えるという複数の動作で実施されるということが十分考えられる。

本研究計画では、以上のような差異を考慮に入れて、双方の有向グラフの対応をとる。既存研究 [16] では、編集距離を用いるなど単純なモデル化が提案されているが、以上のような点を考慮に入れた目的関数を再構築する必要がある。また、レシピと調理映像に特化しないことにも留意する。

既存手法では、双方の最適解をマッチの対象としているが、本研究では、両方の出力を確率的にすることも考える。すなわち、双方の出力を有向グラフとその生成確率の組の列とし、全体の確率を考慮したマッチを行う。これにより最適解ではない有向グラフが選択されるということが起こる。すなわち、言語情報が映像処理を補助したり、またその逆がなされるということを実現できる。

7 おわりに

本稿では、レシピテキストと調理映像を題材とした実世界理解に向けた研究の進捗と計画について報告した。

今後、言語処理では、一般分野と同等の精度を迅速かつ安価に実現するために、部分的アノテーションコーパスを構築していく。映像処理においては、調理映像を収録し、我々のこれまでの研究により自動提示する領域とレシピの固有表現とのリンクをアノテーションする。これらのデータから学習した双方の処理システムが出力する有向グラフをマッチすることで、それぞれの認識精度が相補的に向上することを示す。

参考文献

[1] Daniel Flannery, Yusuke Miyao, Graham Neubig, and Shinsuke Mori. A pointwise approach to training dependency parsers from partially annotated corpora. *Journal of Natural Language Processing*, Vol. 19, No. 3, 2012.

[2] Atsushi Hashimoto, Naoyuki Mori, Takuya Funatomi, Masayuki Mukunoki, Koh Kakusho, and Michihiko Minoh. Tracking food materials with changing their appearance in food preparing. In *Proc. of the CEA2010*, 2010.

[3] Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-projective dependency parsing using spanning tree algorithms. In *Proc. of the EMNLP*, pp. 523–530, 2005.

[4] Kiyotaka Uchimoto, Qing Ma, Masaki Murata, Hiromi Ozaku, and Hitoshi Isahara. Named entity extraction based on a maximum entropy model and transformation rules. In *Proc. of the ACL00*, pp. 326–335, 2000.

[5] 前川喜久雄. 代表性を有する大規模日本語書き言葉コーパスの構築. 人知誌, Vol. 24, No. 5, pp. 616–622, 2009.

[6] 三浦宏一, 高野求, 浜田玲子, 井手一郎, 坂井修一, 田中英彦. 料理映像の構造解析による調理手順との対応付け. 信学論, Vol. J86-DII, No. 11, pp. 1647–1656, 2003.

[7] 伝康晴. 多様な目的に適した形態素解析システム用電子化辞書. 人知誌, Vol. 24, No. 5, pp. 640–646, 2009.

[8] 永田昌明. 統計的言語モデルと n-best 探索を用いた日本語形態素解析法. 情処論, Vol. 40, No. 9, pp. 3420–3431, 1999.

[9] 森信介. 自然言語処理における分野適応. 人知誌, Vol. 27, No. 4, 2012.

[10] 森信介, Neubig Graham, 坪井祐太. 点予測による単語分割. 情処論, Vol. 52, No. 10, pp. 2944–2952, 2011.

[11] 森信介, 長尾眞. n グラム統計によるコーパスからの未知語抽出. 情処論, Vol. 39, No. 7, pp. 2093–2100, 1998.

[12] 柴田知秀, 加藤紀雄, 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. 情処論, Vol. 49, No. 3, pp. 1451–1464, 2008.

[13] 橋本敦史, 舩冨卓哉, 中村和晃, 椋木雅之, 美濃導彦. TexCut: GraphCut を用いたテクスチャの比較による背景差分. 信学論, Vol. J94-D, No. 6, pp. 1007–1016, 6 2011.

[14] 橋本敦史, 大岩美野, 舩冨卓哉, 上田真由美, 角所考, 美濃導彦. 調理行動モデル化のための調理観測映像へのアノテーション. 第 1 回データ工学と情報マネジメントに関するフォーラム (DEIM2009), 2009.

[15] 宮澤飛鳥, 中村和晃, 橋本敦史, 舩冨卓哉, 美濃導彦. 調理者の手と容器の位置関係を利用した「かき混ぜる」行動の認識. 信学技報 データ工学研究会, Vol. 112, No. 75, pp. 25–30, 6 2012.

[16] 山肩洋子, 角所考, 美濃導彦. 調理コンテンツの自動作成のためのレシピテキストと調理観測映像の対応付け. 信学論, Vol. J90-DII, No. 10, pp. 2817–2829, 2007.