レシピ分野における動作対象の状態変化を考慮したデータセットの構築と検索モデルの提案

白井 圭佑 [1], 橋本 敦史 [2], 牛久 祥孝 [2], 栗田 修平 [3], 亀甲 博貴 [4], 森 信介 [4]
[1] 京都大学大学院情報学研究科, [2] オムロンサイニックエックス株式会社,
[3] 理化学研究所 AIP, [4] 京都大学学術情報メディアセンター
shirai.keisuke.64x@st.kyoto-u.ac.jp,
{atsushi.hashimoto,yoshitaka.ushiku}@sinicx.com,
shuhei.kurita@riken.jp, {kameko,forest}@i.kyoto-u.ac.jp

概要

実世界で言語的な指示を理解し行動するエージェントは, その行動によって環境が現在の状態からどのような状態に変化するかを理解出来る必要がある. 本研究では, 料理ドメインにおいてこの問題を調査する為に, goal state identification by retrieval (GSIR) タスクを提案する. このタスクでは, 視覚的な現在の状態と言語的な対象物と動作の情報を入力とし, 動作後の視覚的な状態の検索を行う. 実験の為に, ウェブから日本語の 200 レシピを収集し, これを基に Recipe-GSIR データセットを作成した. 本論文では, 本タスクにおけるベースラインモデルとその発展モデルも提案し, それらの実験結果を報告する.

1 はじめに

言語的な指示を理解し, それに基づいた行動を取る自律エージェントの構築は自然言語処理における一つの目標である [1]. 手順書に記述された一連の指示を遂行する場合には, 各指示文を理解し, それに対応した行動を実行する必要がある. しかし, これには各動作によってその対象物の状態がどのように変化するかを理解する能力が求められる. [1)].

本研究では, この問題に対処し, goal state identification by retrieval (GSIR) タスクを提案する. ここでは, 動作後の状態が視覚的な情報として存在すると仮定し, それを検索によって候補の中から特定することを目指す. 入力としては, 視覚的な現在の状態の情報と言語的な動作とその対象物の情報が利用可能であるとする. 本タスクは, 動作による対象物の状態の変化を考慮しなければならないという点で困難であるといえる. 本研究では料理ドメインに限定し, ウェブから収集した 200 レシピを基に, Recipe-GSIR データセットを作成した. このデータセットでは, レシピから抽出した動作とその対象物の単語列に対し, 付属の調理動画からサンプルした画像を用いることで, 動作前後の視覚的な状態の情報の付与を行った. また, 言語側のアノテーションはレシピ固有表現とレシピフローグラフ [2] を基に行った. さらに, 本研究では, 対象とする料理をサラダに限定してデータセットを構築した. 本論文では, 図 1 に示すように, 共有埋め込み空間の学習によって GSIR タスクを解く. 同時に, ベースラインモデルとレシピフローを利用した発展的なモデルも提案し, 最後にそれらの定量的な評価を行う.

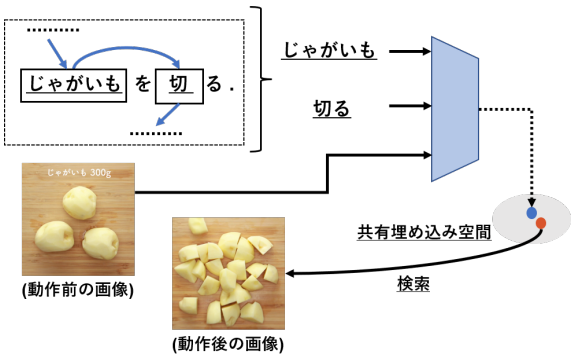

図 1 GSIR タスクの概要図.

2 データセット

Recipe-GSIR データセットの構築の為に, まずクラシル[2)]からサラダ 200 レシピの収集を行った. ここで, 収集した各レシピは, それぞれ材料リスト, 調理手順書, 調理動画から成る. 表 1 に収集した材料リス

1) 例えば, "じゃがいも"を"切る"ことによって, ”じゃがいも”は切られた状態に遷移することをモデルが理解する必要がある.

2) https://www.kurashiru.com (2021/12/14).

表 1 材料リストと調理手順書の統計情報.

データの種類	単語数	単語の種類数	材料数/ステップ数	材料数/ステップ数 (平均)
材料リスト	10,294	448	1,701	8.51
調理手順書	23,106	681	1,077	5.38

トと調理手順書の統計情報を示す[3]. 収集したデータに対して, (i) レシピ固有表現 (r-NE) の付与, (ii) レシピフローグラフ (r-FG) の付与, (iii) r-FG から抽出した調理動作と対象物への動作前後の画像の付与を行うことでデータセットの構築を行った. 以降では, 各アノテーション手順について説明する.

2.1 レシピ固有表現 (r-NE)

まず, 材料リストと調理手順書上の単語に r-NE の付与を行った. r-NE タグには, 森ら [2] の提案した 8 タグを用いた. この中で, Ac, F, T の 3 タグは本研究において重要であるため, 簡潔に説明する. Ac は調理者による動作のことを指し, 食材による動作 (Af) と区別する. F は食材を指し, 材料, 中間生成物, 最終生成物はこれに含まれる. T は調理に用いられる道具を指し, 包丁や電子レンジ等がこれに該当する.

2.2 レシピフローグラフ (r-FG)

次に, r-NE に対して r-FG の付与を行った. ここで, r-FG [2, 4] とは, r-NE を頂点とし, それらの関係を辺とする有向非循環グラフである. r-FG ラベルには, 森ら [2] の提案した 13 ラベルを用いた. この中で, Targ と Dest の 2 ラベルは本研究において重要であるため, 簡潔に説明する. Targ は基本的に F を始点とし, Ac を終点とすることで, 食材 (F) に対する動作 (Ac) を表現する. Targ は Ac を始点に取ることも可能であり, この場合は始点側の Ac はその動作によって生成された中間生成物を表す. 一方で, Dest は F や T を始点とし, Ac を終点とすることで, 動作 (Ac) が行われる方向又は場所としての食材 (F) や道具 (T) を表現する. Targ と同様に, Dest も Ac を始点に取ることが可能であり, この場合も Ac は中間生成物を表す.

ここまでに付与した r-NE と r-FG を基に, 調理動作とそれに紐づく対象物 (食材) と場所の情報を抽出する. これは, Targ の終点の単語列を調理動作, 始点の単語列を食材として抽出することで実現出来る. 同様に, Ac を終点に持つ Dest の始点の単語列を場所として抽出を行う. 以降では, これによって得られた調理動作, 食材, 場所からなる 3 項組を GSIR タスクにおける言語的な情報として用いる.

表 2 画像のアノテーション数. ✓ は付与された要素を, 同行の数はそれに対応する 3 項組の総数を, それぞれ示す.

画像		アノテーション数
動作前	動作後	
		597
✓		72
	✓	484
✓	✓	2,552
合計		3,705

本研究では, さらにこの 3 項組を細粒度なものに変換することを考える. ここで, 細粒度とは 3 項組内の 1 動作が同時に 1 食材のみを取る場合を指すものとする. 例えば, "レタスときゅうりを切る"という指示文では, 1 つの動作"切る"の対象として 2 つの食材"レタス"と"きゅうり"があるが, これは (切る, レタス), (切る, きゅうり) のように食材毎に動作を分割することで細粒度な組が得られる. しかし, 文書全体を通して r-FG からの細粒度な 3 項組の自動抽出を行うには, 各動作が生成する中間生成物の個数を把握する必要がある. 例えば, 動作"切る"は n 個の食材に対して n 個の切られた食材を生成するが, 動作"盛り付ける"は対象とする食材の数に関わらず 1 つの盛り付けられた食材を生成する筈である. 従って, 本研究では前者と後者の動作を区別し, 収集したレシピから手動で作成したそれぞれのリストを用いることで, 細粒度な 3 項組の自動抽出を行った. これによって, 合計 3,705 組の細粒度な 3 項組が得られた.

2.3 細粒度な画像アノテーション

最後に, これらの 3 項組に対して動作前後の画像の付与を行った. アノテーション候補の画像としては, 同一のレシピの調理動画から 3FPS でサンプルした画像のみを用いた. アノテーションの際は, より対象の状態の変化がわかるように画像を選択した[4] また, 対象の食材が 75%以上見えない場合, 又は対象の動作が動画内に存在しない場合には, 画像を付与す

3) これらの単語分割には KyTea [3] を用いた.

4) 例えば, "レタスを切る"という動作に対しては, 動作前の画像としては切られる前のレタスの画像が, 動作後の画像としては切られた後の状態のレタスのものが, それぞれ付与される.

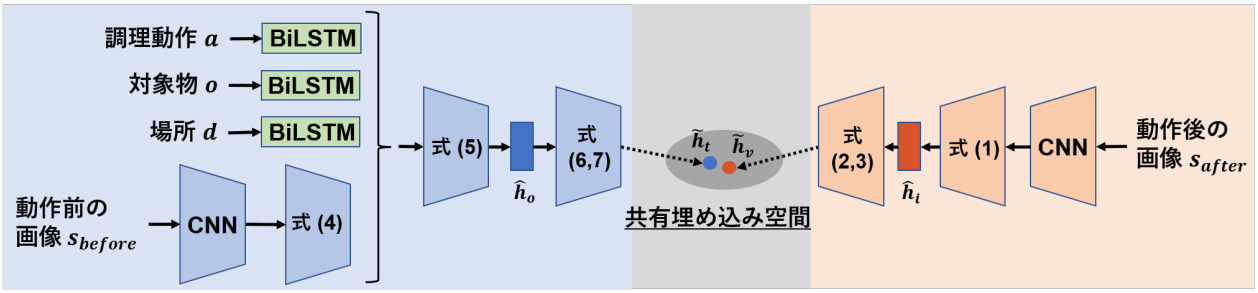

図 2　共有埋め込みモデルの概要図.

る代わりに欠損値を記録した. 表 2 に本ステップによって得られたアノテーション数を示す.

3　実験

3.1　GSIR タスク

GSIR タスクでは, 言語的な情報として調理動作 a と K 個の属性 $\{t_1, \cdots, t_K\}$ が与えられた時, 動作後の視覚的な状態 s_{after} を候補となる N 個の画像の集合 $\{s_1, \cdots, s_N\}$ から特定することが目的である. ここで, 候補の N 個の画像の中には重複がないものとする. また, a と $\{t_1, \cdots, t_K\}$ に加え, 動作前の視覚的な情報 s_{before} も入力として用いることが可能であるとする. 本研究では, 2.2 節の細粒度な 3 項組に含まれる調理動作を a とし, 対象物 o と場所 d をそれぞれ $(t_1, t_2) = (o, d)$ とする. また, 2.3 節で付与した動作前後の画像をそれぞれ s_{before}, s_{after} として用いる.

3.2　共有埋め込みモデル

本研究では, 図 2 のように, 共有埋め込みモデル [5] を学習することで GSIR タスクを解く. まず, a, o, d と s_{before} を基に動作後の対象物 o の状態を表す $\hat{h}_o$ を計算し, それを共有埋め込み空間上へ埋め込む. 同時に, 検索候補 $\{s_1, \cdots, s_N\}$ も同空間へ埋め込み, 対応する動作後の画像 s_{after} の検索を行う. 以下では, これらの具体的な計算について説明する.

s_{after} はまず, 畳み込みニューラルネットワーク (CNN) を用いて特徴量 $h_i^{after} \in \mathbb{R}^{d_i}$ の抽出を行い, $\hat{h}_i$ を以下のように計算する.

$$\hat{h}_i = W_2^I(\text{ReLU}(W_1^I h_i^{after} + b_1^I)) + b_2^I, \tag{1}$$

ここで, $W_1^I, W_2^I \in \mathbb{R}^{d_i \times d_i}$, $b_1^I, b_2^I \in \mathbb{R}^{d_i}$ は学習可能な重みである. 次に, Miech ら [6] に従い, 共有埋め込み空間上の表現 $\hat{h}_i$ を以下のように計算する.

$$h_v = (W_3^I \hat{h}_i + b_3^I) \circ \sigma(W_4^I(W_3^I \hat{h}_i + b_3^I) + b_4^I), \tag{2}$$

$$\tilde{h}_v = \frac{h_v}{||h_v||_2}, \tag{3}$$

ここで, σ はシグモイド関数, $\circ$ は要素ごとの積であり, $W_3^I \in \mathbb{R}^{d_e \times d_i}$, $W_4^I \in \mathbb{R}^{d_e \times d_e}$, $b_3^I, b_4^I \in \mathbb{R}^{d_e}$ は学習可能な重みである.

3 項組 a, o, d は埋め込み行列を用いて d_v 次元のベクトルに変換し, それらを双方向長短期記憶 (BiLSTM) ネットワーク [7] を用いることで d_t 次元のベクトル h_a, h_o, h_d にそれぞれ変換する. 動作前の視覚的な状態 s_{before} は CNN を用いて特徴量 $h_i^{before} \in \mathbb{R}^{d_i}$ を抽出し, 以下のように $\hat{h}_i^{before} \in \mathbb{R}^{d_t}$ へ変換する.

$$\hat{h}_i^{before} = W_1^T h_i^{before} + b_1^T, \tag{4}$$

ここで, $W_1^T \in \mathbb{R}^{d_t \times d_i}$ と $b_1^T \in \mathbb{R}^{d_t}$ は学習可能な重みである. これらから, 動作後の対象物の状態 $\hat{h}_o$ を以下のように計算する.

$$\hat{h}_o = W_3^T(\text{ReLU}(W_2^T[h_a; h_o; h_d; h_i^{before}] + b_2^T)) + b_3^T, \tag{5}$$

ここで, ; は結合であり, $W_2^T \in \mathbb{R}^{4d_t \times 4d_t}$, $W_3^T \in \mathbb{R}^{d_t \times 4d_t}$, $b_2^T \in \mathbb{R}^{4d_t}$, $b_3^T \in \mathbb{R}^{d_t}$ は学習可能である. 最後に, 共有埋め込み空間上の表現 $\hat{h}_o$ を以下のように計算する.

$$h_t = (W_4^T \hat{h}_o + b_4^T) \circ \sigma(W_5^T(W_4^T \hat{h}_o + b_4^T) + b_5^T), \tag{6}$$

$$\tilde{h}_t = \frac{h_t}{||h_t||_2}, \tag{7}$$

ここで, $W_4^T \in \mathbb{R}^{d_e \times d_t}$, $W_5^T \in \mathbb{R}^{d_e \times d_e}$, $b_4^T, b_5^T \in \mathbb{R}^{d_e}$ は学習可能なパラメータである.

共有埋め込み空間上における $\tilde{h}_t$ と $\tilde{h}_v$ の距離は以下のように計算する.

$$D(\tilde{h}_t, \tilde{h}_v) = ||\tilde{h}_t - \tilde{h}_v||_2. \tag{8}$$

これを基に, n 個の組 $((\tilde{h}_{t,1}, \tilde{h}_{v,1}), \cdots, (\tilde{h}_{t,n}, \tilde{h}_{v,n}))$ に対して以下の Triplet Loss [8] を最小化するように重みの調整を行う.

$$\mathcal{L} = \sum_{i=1}^{n} \{\max(D_{i,i} - D_{i,j} + \delta, 0) + \max(D_{i,i} - D_{k,i} + \delta, 0)\}. \tag{9}$$

表 3 Recipe-GSIR データセットにおける実験結果. ✓ は入力に用いた要素を意味する. また, 1 行目はランダムサーチにおける結果である.

入力に用いた要素		レシピフロー	R@1 (↑)	R@5 (↑)	R@10 (↑)	MedR (↓)
3 項組	動作前の画像					
			0.76 (± 0.42)	2.37 (± 0.94)	3.73 (± 1.05)	149.00 (± 5.76)
	✓		5.78 (± 0.86)	15.09 (± 1.70)	23.73 (± 2.34)	28.80 (± 3.22)
✓			10.08 (± 1.29)	21.98 (± 2.02)	31.79 (± 2.42)	23.60 (± 2.01)
✓		✓	10.73 (± 0.74)	24.52 (± 1.52)	36.01 (± 1.97)	18.90 (± 1.37)
✓	✓		11.74 (± 1.70)	26.45 (± 2.72)	38.85 (± 3.48)	16.40 (± 1.85)
✓	✓	✓	**12.26** (± 1.88)	**28.32** (± 2.47)	**41.66** (± 3.34)	**15.10** (± 2.51)

ここで, $D_{i,j} = D(\tilde{h}_{t,i}, \tilde{h}_{v,j})$ であり, δ はマージンを指す. 式 (9) において, $D_{i,i}$ は正例の組における距離であり, $D_{i,j}$, $D_{k,i}$ はそれぞれ負例の 3 項組と動作後の画像に対する距離である. 本研究では, 負例はミニバッチ内からのサンプリングで獲得する.

レシピフローを考慮したモデル. 2.2 節で説明した通り, r-FG から抽出した対象物 o や場所 d は調理動作 (Ac) を表す単語列となる場合がある. この場合には具体的な対象物や場所の情報が与えられなくなるため, 検索がより困難になるという問題がある. これに対処するために, r-FG のフローを利用し, 過去にその Ac タグを動作 a として用いた際に得られた動作後の状態 $\hat{h}_o$ を, 中間生成物として h_o や h_d に用いることが考えられる. このレシピフローを利用したモデルの精度も 3.4 節で報告する.

3.3 実験設定

モデルパラメータ. BiLSTM の隠れ層の次元は 256 とし, 1 層のものを用いた. 他の次元サイズはそれぞれ $(d_v, d_t, d_i, d_e) = (496, 512, 2048, 128)$ とした. CNN には事前学習済みの ResNet-152 [9] を, 重みを固定して用いた. 学習可能な重みは AdamW [10] を, 初期学習率 1.0×10^{-5} として用いた. また, エポック数は 100 に設定した他, 無作為に抽出した 4 レシピに含まれるサンプルからミニバッチを構成した. 式 (9) の δ は実験的に 0.1 に設定した.

評価. 学習と評価には, 動作後の画像が付与されている 3,036 例を用いた. ここで, 3,036 例の内 484 例は動作前の画像が付与されていないが, それらには h_i^{before} の代わりに零ベクトルを用いることで欠損値を表現した. モデルの評価は 10-分割交差検定により行った. この際, データセット全体の 90%を学習データに, 残りの 10%をテストデータとして分割した. 評価指標には Recall@k(R@k) と Median rank(MedR) を用いた.

3.4 実験結果

表 3 に実験結果を示す. まず, 動作前の画像のみ (2 行目) からでも, ある程度の検索が可能であることがわかる. しかし, 3 行目の結果との比較から, これよりは調理動作の 3 項組を用いた方がより高い精度を実現するとわかる. これは, 動作前の画像のみを用いる場合には, 同様の背景を持つ画像を対象に絞ることが可能であるが, どれが動作後のものとして適当かまでは判別出来ないことが原因だと考えられる. 次に, 3,4 行目の比較から, r-FG のフローを用いることでさらなる精度向上の実現が可能だとわかる. これは, 過去に推定した $\hat{h}_o$ を h_o や h_d として扱うことが効果的であることを示唆している. これに加えて, 5 行目の結果から, フローを用いるか動作前の画像を用いるかでは後者の方が精度が高いことがわかる. これは, 前者ではフローから推定した現在の対象物の状態の情報を用いているのに対し, 後者では直に現在の視覚的な状態を利用出来るためだと思われる. 最後に, 6 行目の結果から, 入力として利用可能な全ての情報を用いた場合に最も高い検索精度が実現出来ることがわかる. これに関しては, 動作前の画像が付与されていない例に関しては, フローから推定した情報がより効果的に働く為だと考えられる.

4 おわりに

本研究では, GSIR タスクを提案し, 料理ドメインにおいて Recipe-GSIR データセットの構築を行った. また, ベースラインモデルの提案に加え, レシピフローを用いた発展的なモデルも提案し, それらの検索精度を定量的に評価した. 今後の方向としては, フローグラフから抽出した 3 項組だけでなく, BERT 等を用いて文書上のコンテキストを考慮した表現も利用した精度向上が考えられる. また, 別の方向として, 候補からの検索ではなく動作後の状態を表す画像を直接生成する方向も考えられる.

謝辞

本研究は JSPS 科研費 21H04910 および JST さきがけ JPMJPR20C2 の支援を受けたものである．

参考文献

[1] Terry Winograd. Understanding natural language. **Cognitive psychology**, Vol. 3, No. 1, pp. 1–191, 1972.

[2] Shinsuke Mori, Hirokuni Maeta, Yoko Yamakata, and Tetsuro Sasada. Flow graph corpus from recipe texts. In **Proceedings of the Ninth International Conference on Language Resources and Evaluation**, pp. 2370–2377, 2014.

[3] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise prediction for robust, adaptable japanese morphological analysis. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pp. 529–533, 2011.

[4] Hirokuni Maeta, Tetsuro Sasada, and Shinsuke Mori. A framework for procedural text understanding. In **Proceedings of the 14th International Conference on Parsing Technologies**, pp. 50–60, 2015.

[5] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 2630–2640, 2019.

[6] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. **arXiv preprint arXiv:1804.02516**, 2018.

[7] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. **Neural networks**, Vol. 18, No. 5-6, pp. 602–610, 2005.

[8] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In **Proceedings of the British Machine Vision Conference**, pp. 119.1–119.11. BMVA Press, September 2016.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 770–778, 2016.

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **Proceedings of the 7th International Conference on Learning Representations**, 2019.

[11] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 260–270, 2016.

[12] Yoko Yamakata, Shinsuke Mori, and John A Carroll. English recipe flow graph corpus. In **Proceedings of the 12th Language Resources and Evaluation Conference**, pp. 5187–5194, 2020.

表 4　r-NE タグと各タグのアノテーション数.

タグ	意味	アノテーション数
F	食材	5,098
T	道具	758
D	継続時間	129
Q	分量	1,778
Ac	調理者の動作	2,532
Af	食材の動作	353
Sf	食材の状態	971
St	道具の状態	67
合計	—	11,686

表 5　r-NE タグの推定結果.

精度	再現率	F 値
98.54 (± 0.77)	98.82 (± 0.66)	98.68 (± 0.70)

A　r-NE/r-FG アノテーション

表 4 に r-NE タグの一覧と各タグのアノテーション数を示す. 表において, 食材や分量のアノテーション数が特に多い理由としては, それらが材料リストに常に含まれれている為である. r-NE のアノテーション結果がどの程度自動的に推定出来るかを調査するために, r-NE の推定を固有表現認識 (NER) タスクとして捉え, BiLSTM-CRF モデル [11] を用いて学習と評価を行った. この際, 評価指標には精度, 再現率, F 値を用いた. 表 5 に 10-分割交差検定による結果を示す. 結果から, 学習後の NER モデルは非常に高い精度で r-NE タグを推定出来ていることがわかる. これは, 対象のデータが料理ドメインの中でも特にサラダに限定している為, 複数のレシピに出現する特定の動作 (Ac) や食材 (F) の推定が比較的容易であったからだと考えられる.

表 6 に r-FG ラベルの一覧と各ラベルのアノテーション数を示す. 表において, Targ ラベルのアノテーション数が最も多い結果となったがこれは先行研究 [4] と同じ傾向である. r-FG ラベルのアノテーション結果がどの程度自動的に推定出来るかを調査するために, r-FG の推定を最大全域木の推定問題として捉え, 先行研究 [4] と同様に Chu-Liu/Edmonds アルゴリズムを用いて推定した. この際, 前田ら [4] と同様の特徴量を用いて学習と評価を行った. この際, 評価指標には精度, 再現率, F 値を用いた. 表 7 に 10-分割交差検定による結果を示す. この結果は, 先行研究 [4, 12] と比較して高い精度であるといえる. 考えられる理由としては, 調理手順書のみを対象としている先行研究 [4, 12] と異なり, 材料リスト上の表現から調理手順書上の表現にも r-FG ラベルのアノテーションを行っている為, それらの一部のラベルの推定が比較的容易であったからだと思われる.

表 6　r-FG ラベルと各ラベルのアノテーション数.

ラベル	意味	アノテーション数
Agent	動作と主語	330
Targ	動作と対象	2,961
Dest	動作と方向や場所	1,025
T-comp	動作と手段 (道具)	157
F-comp	動作と手段 (食材)	20
F-eq	同一の食材	2,397
F-part-of	食材の一部	330
F-set	食材の集合	987
T-eq	同一の道具	4
T-part-of	道具の一部	0
A-eq	同一の動作	1
V-tm	動作のタイミング	112
other-mod	その他の修飾語句	2,967
合計	—	11,291

表 7　r-FG ラベルの推定結果.

精度	再現率	F 値
90.78 (± 1.30)	90.60 (± 1.43)	90.70 (± 1.43)

表 8　アノテーション一致率.

アノテーションの種類	精度	再現率	F 値
r-NE (2.1 節)	97.93	98.88	98.40
r-FG (2.2 節)	86.18	86.04	86.11
画像 (2.3 節)	75.13	70.60	72.80

B　アノテーション一致率

本研究では, 1 人のアノテーターによって 2 節の全アノテーションが行われた. アノテーションの一貫性を調査するために, 別のアノテーターに無作為に選択した 10 レシピのアノテーションを依頼し, 既にアノテーション済みのものを正解ラベルとして捉えることで, 精度, 再現率, F 値を用いて一致率を計算した. 表 8 にその結果を示す. 表から, r-NE のアノテーションは非常に高い割合で一致していることがわかる. これは, r-NE に関しては, アノテーション時にまず NER モデルによる推定結果を提示し, 誤りを訂正するようにアノテーターに依頼していたため, 元々の NER モデルによる推定が高精度であったことが理由だと考えられる. 次に, r-FG のアノテーションでは少し一致率が低下しているが, これはアノテーション時に頂点としての材料リストと調理手順書に含まれる全ての NE と辺としてのラベルが候補となることを考慮すると, 一致率として高いといえる. 最後に, 画像のアノテーション結果では, r-FG の時からさらに一致率が低下していることがわかる. しかし, これも同様に, 調理動画からサンプリングした全ての画像が候補となること, 動作前後の画像として適当なものが一つとは限らない[5] ことを考慮すると, この一致率は十分に高いといえる.

5)　複数の画像が同時に基準を満たす場合がある.