一般論文

Peer-Reviewed

調理動作後の物体の視覚的状態予測を目指した Visual Recipe Flow データセットの構築と評価

白井　圭佑[†]・橋本　敦史[††]・西村　太一[†]・亀甲　博貴[†††]・栗田　修平[††††]・森　信介[†††]

本稿では，調理レシピにおいて，調理動作後の物体の視覚的な状態の予測を目指し，Visual Recipe Flow (VRF) データセットを提案する．VRF データセットは (i) 物体の視覚的な状態遷移と (ii) レシピ全体のワークフローに対するアノテーションから成る．視覚的な状態遷移は動作前後の物体の観測を表す画像の組として，ワークフローはレシピフローグラフとして，それぞれ表現する．ここでは，データセットの構築方法，アノテーション手順について順に説明し，アノテータ間のアノテーション一致率を測ることでデータセットの品質を調査する．最後に，動作前後の画像と物体のテキスト情報を用いたマルチモーダルな情報検索の実験を行うことで，各アノテーション要素の重要性について調べる．

キーワード：アノテーション，マルチモーダル，グラフ構造

Visual Recipe Flow: A Dataset for Learning Visual State Changes of Objects with Recipe Flows

Keisuke Shirai[†], Atsushi Hashimoto[††], Taichi Nishimura[†], Hirotaka Kameko[†††], Shuhei Kurita[††††] and Shinsuke Mori[†††]

We present a new multimodal dataset called Visual Recipe Flow, which enables us to learn each cooking action result in a recipe text. The dataset consists of object state changes and the workflow of the recipe text. The state change is represented as an image pair, while the workflow is represented as a recipe flow graph (r-FG). We explain the data collection and annotation procedure and evaluate the dataset by measuring the inter-annotator agreement. Finally, we investigate the importance of each annotation component by conducting multi-modal information retrieval experiments.

Key Words: *Annotation, Multimodal, Graph Structure*

[†] 京都大学大学院情報学研究科, Graduate School of Informatics, Kyoto University
[††] オムロンサイニックエックス株式会社, OMRON SINIC X Corporation
[†††] 京都大学学術情報メディアセンター, Academic Center for Media Studies, Kyoto University
[††††] 理化学研究所 AIP, RIKEN AIP, JST PRESTO

本論文の一部は言語処理学会第 28 回年次大会 (白井 et al. 2022) および COLING 2022 (Shirai et al. 2022) で発表した内容を加筆修正したものである．

1　はじめに

予測は人間が生来備えている能力である．我々は日常生活において，様々な予測を行い，それを基に行動を行っている．例えば，調理においては，手順や調理動作によって得られる目標の状態を先に予測し，それを手掛かりに動作を実行しているといえる．これには動作や対象の物体に関する知識が必要であり，同時に作業全体のワークフローについても理解する必要がある．したがって，調理レシピを基に人間と同様に調理を行う自律エージェントを確立するためには，これらの能力の実現が必須であるといえる．図 1 に例を示す．この例では，エージェントは二つ目の調理動作に必要な食材を特定しつつ動作後に得られる観測を予測している．

この方向の先行研究として，調理手順に視覚的なアノテーションを施したものが存在する (Nishimura et al. 2020; Pan et al. 2020)．ここで，Nishimura et al. (2020) は各調理手順に対応する 1 枚の画像に対して，動作や食材の表現に対応する箇所に矩形領域のアノテーションを行った．また，Pan et al. (2020) は調理手順を文レベルに分割した後，各文に対応するフレーム列を，レシピに紐づく調理動画から抽出してアノテーションを行った．これらは調理手順に視覚的な情報を結びつけているが，調理動作の結果を予測したい場合には，まだアノテーションが不十分である．例えば，“トマトを切ってボウルに入れる．”には，1 文に「切る」と「入れる」の二つの調理動作が含まれているが，上記の先行研究のアノテーションでは，これらを個別に扱うことが出来ない．従って，調理動作結果を予測するためには，調理動作レベルのアノテーションが必要であり，率直な解決策としてはより密なアノテーションを用意することが考えられる．

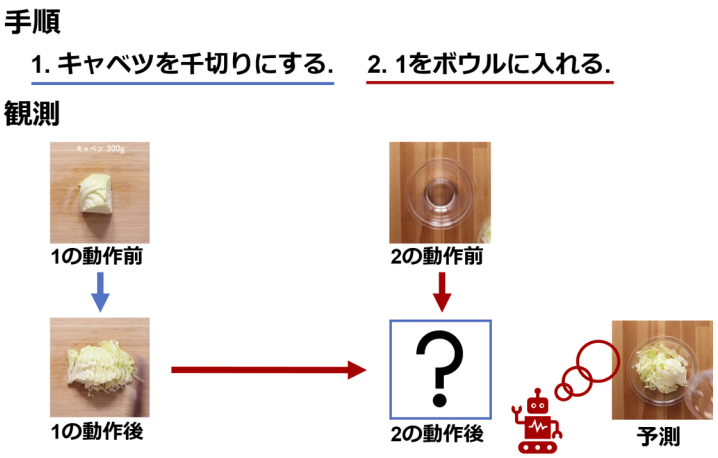

図 1　調理エージェントによる調理動作結果の予測の例．ここでは，2 つ目の調理動作（入れる）の結果を予測している．

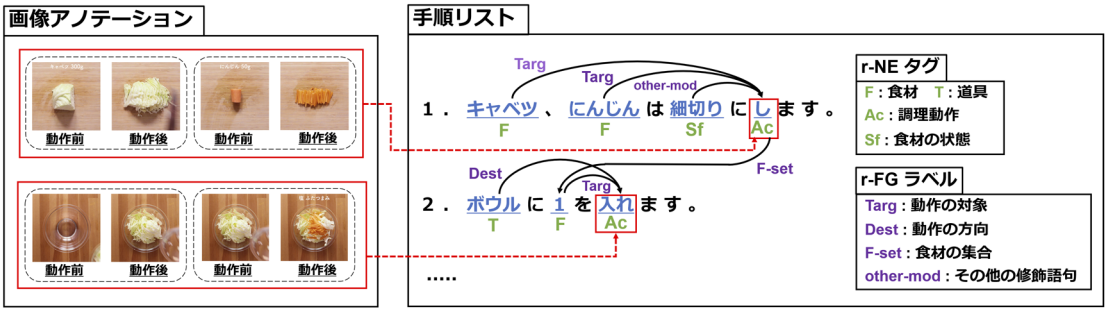

図 2　アノテーションの例．画像の組は物体毎の動作前後の観測を表しており，これらは手順リスト内の調理動作に紐づけられている．手順リスト内では，r-FG を用いて調理動作を含む表現間の依存関係を表現している．

本稿では，レシピにおける調理動作後の物体[1]の状態の予測を目指し，新たに Visual Recipe Flow (VRF) データセットを提案する．図 2 に示す通り，VRF データセットは (i) 調理動作による物体の視覚的な状態の遷移と (ii) レシピ全体のワークフローに対するアノテーションから成る．視覚的な状態遷移は動作前後に対応する観測の組として表現し，ワークフローは先行研究のレシピフローグラフ (Recipe flow graph; r-FG) (Mori et al. 2014) を用いて表現する．ここで，観測の組は r-FG 中の調理動作と紐づいており，これによって実世界とテキストのクロスモーダルな関係を考慮することが可能となる．また，ウェブサイトにてデータセット作成に用いたレシピの URL リストとダウンロードしたデータを基にデータセットを構築するスクリプトを公開している[2]．

また，本研究と関わりの深い研究として，手順書上での物体の状態遷移の追跡を行うものが存在する (Dalvi et al. 2018; Bosselut et al. 2018; Tandon et al. 2020)．これは，各手順による実世界における影響を考慮することで，手順書の理解を目指したものである．本研究はこれらの流れを汲んだものでもあり，大きな違いは状態遷移をテキストでなく画像で表現している点にある．画像は物体の外観に関する情報を与えるため (Isola et al. 2015; Zhang et al. 2021)，先行研究と比較して実世界に関するより豊富な情報を提供することが期待される．また，調理の持つ逐次的な性質を活かし，大規模言語モデルの文書理解能力の評価 (Srivastava et al. 2022) や，vision-language モデルの few-shot 設定における学習能力の評価 (Alayrac et al. 2022) に用いることも考えられる．

[1] 本稿では，物体は食材か道具のいずれかを指すものとする．
[2] https://github.com/kskshr/Visual-Recipe-Flow.

2 Visual Recipe Flow データセット

Visual Recipe Flow (VRF) データセットでは，レシピ上の物体の調理動作前後の状態を，視覚的なアノテーションを用いて表現する．ここで，物体や調理動作等の重要表現はレシピ固有表現 (Recipe namend entity; r-NE) (Mori et al. 2014) を用いて識別される．また，r-NE を基に，レシピ全体のワークフローをレシピフローグラフ (Recipe flow graph; r-FG) (Mori et al. 2014) を用いて表現する．本節では，まず r-FG について概説し，その後，本データセットにおける視覚的なアノテーションの特徴について説明する．

2.1 Recipe flow graph (r-FG)

r-FG は図 2 に示すような非巡回有向グラフである．r-FG において，点は r-NE によって識別された表現であり，辺は二つの r-NE 間の依存関係をラベルを用いて表したものである．r-FG の特徴の一つとして，動作が必要とする物体を特定出来ることが挙げられる．例えば，図 2 で動作 “入れ” が対象とする物体 “1” は手順 1 の生成物であり，細切りにしたキャベツとにんじんを指しているが，これは r-FG の辺を辿ることで特定可能である．また，本データセットでは，先行研究 (Mori et al. 2014) と異なり，食材リストから手順リストへの依存関係もアノテーションしており，原材料の情報を考慮することも可能である．

2.2 視覚的アノテーション

視覚的アノテーションは，調理動作による物体の視覚的な状態遷移を表したものであり，動作前後に対応する画像の組として表現する．ここで，各画像の組は手順リスト中の調理動作に紐づいている．また，一つの調理動作が複数の物体を対象とする場合に関しては，図 2 のように，物体毎に個別の画像の組を提供する．これによって，一動作が扱う物体が複数個ある場合にも，それらの状態遷移も捉えることが出来る．

3 アノテーション仕様

本節では，VRF データセットを構成する各アノテーションについて説明する．アノテーションは，(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)，物体の動作前後の画像の付与の 3 段階から成る．また，各レシピは日本語で記述されており，それぞれ材料リスト，手順リスト，調理動画を備えていると想定する．図 3 にアノテーションの例を示す．

3.1 r-NE

まず，材料リストと手順リスト中の表現に対してレシピ固有表現タグ (Recipe named entity;

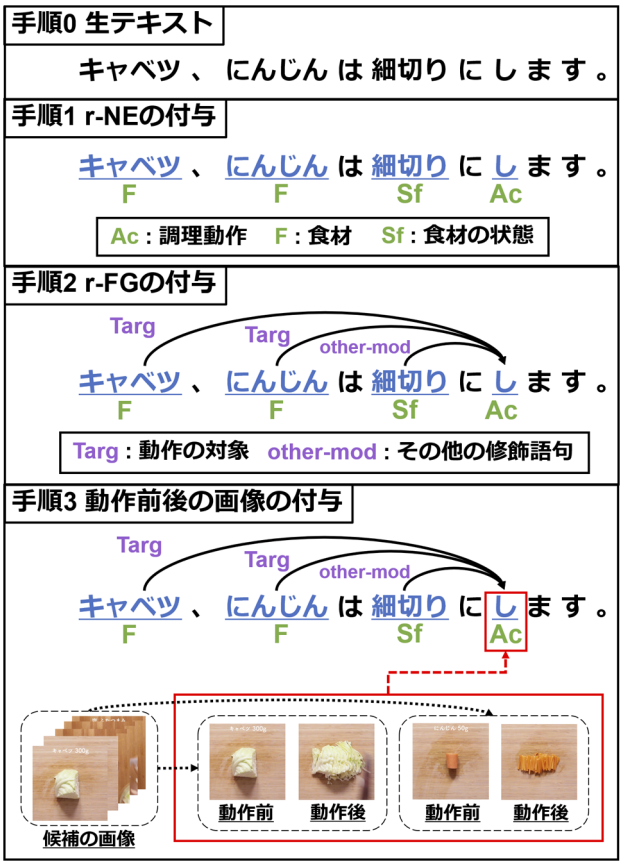

図 3　各アノテーション手順における例．アノテーションは前段階の結果をもとに行われる．

r-NE) のアノテーションを行う．文は日本語テキスト解析機 KyTea (Neubig et al. 2011) によって，事前に単語レベルに分割されているものとする．r-NE タグの定義に関しては，先行研究 (Mori et al. 2014) に従い，表 1 に示す 8 種類のものを用いる．これらのうち，Ac (Action by chef; Ac)，F (Foods; F)，T (Tools; T) の 3 タグは，調理動作に直接関わるという点で，本研究において重要な意味をもつ．また，材料リスト中における動作の表現（茹でた，スライス等）に関しては，調理作業の開始時点で既に完了しているものとし，動作 (Ac) ではなく食材の状態 Sf (State of foods; Sf) としてアノテーションする．

3.2　r-FG

次に，前段階でアノテーションした r-NE を基に，r-FG のアノテーションを行う．ラベルは先行研究 (Maeta et al. 2015) に従い，表 2 に示す 13 種類の依存関係ラベルを用いる．これらのうち，Targ （動作の対象）と Dest （動作の方向）は動作 (Ac)，食材 (F)，道具 (T) 間の関係

タグ	タグの意味	g アノテーション例	付与数
F	食材	g にんじん，根本，１，···	5,098
T	道具	g ボウル，ラップ，電子レンジ，···	758
D	継続時間	g5 分程，１０分，１分ほど，···	129
Q	分量	g 残り，大さじ１，３０ｍｌ，···	1,778
Ac	調理者による動作	g 切，入れ，切り落と，···	2,532
Af	食材による動作	g 馴染，取れ，沸騰，···	353
Sf	食材の状態	g 一口大，薄切り，半分，···	971
St	道具の状態	g ６００W，中火，５００W，···	67
Total	—	g– -	11,686

表 1　r-NE タグと各タグの付与数.

ラベル	ラベルの意味	アノテーション例（始点 → 終点）	付与数
Agent	主語	盛り付け → 完成，味 → 馴染（ませる），···	330
Targ	対象	キャベツ → 切（る），１ → 入れ（る），···	2,961
Dest	方向	器 → 盛り付け（る），耐熱ボウル → 入れ（る），···	1,025
T-comp	道具	電子レンジ → 加熱（する），フォーク → 潰（す），···	157
F-comp	食材	水 → さら（す），塩コショウ → 調え（る），···	20
F-eq	同一の食材	にんじん（材料リスト） → にんじん（手順リスト）， 切（る） → ２，···	2,397
F-part-of	食材の一部	にんじん → 皮，ミニトマト → ヘタ，···	330
F-set	食材の集合	酢 → A，ドレッシング → 材料，···	987
T-eq	同一の道具	加熱 → 耐熱ボウル， フライパン（手順 1） → フライパン（手順 3），···	4
T-part-of	道具の一部	—	0
A-eq	同一の動作	な（り） → 乳化（する）	1
V-tm	動作のタイミング	馴染（んだら） → 盛り付け（る）， しんなり → 加熱（する），···	112
other-mod	その他の修飾語句	薄切り → し，半分 → 切（る），···	2,967
Total	—	—	11,291

表 2　r-FG ラベルと各ラベルの付与数.

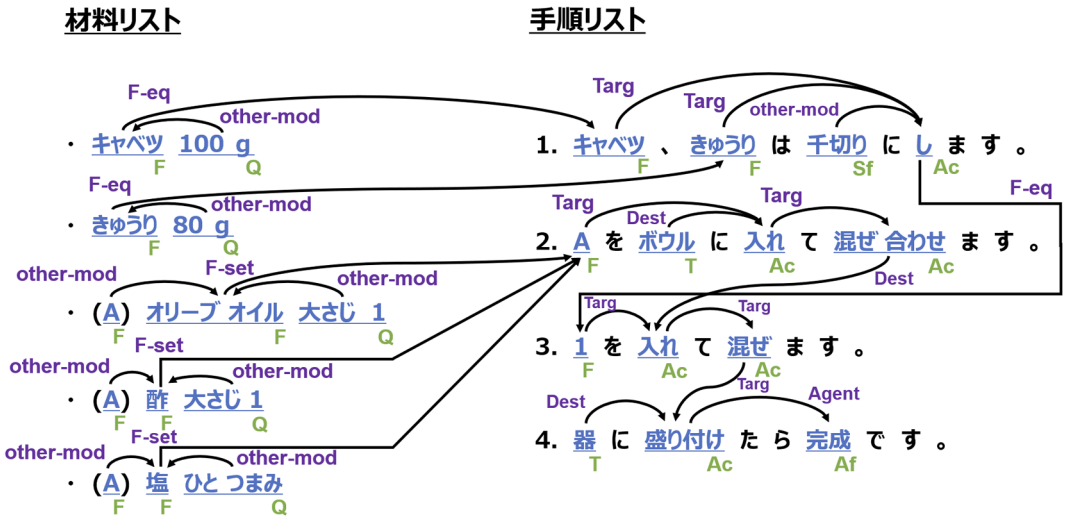

図 4　材料リスト，手順リストに対する r-NE, r-FG のアノテーション例.

を記述するため，本研究において重要な意味をもつ．図 4 に材料リスト，手順リストに対する r-NE, r-FG のアノテーション例を示す.

3.3　動作前後の画像の組

最後に，画像による物体の状態遷移のアノテーションを行う．画像は，調理動画から 3FPS で抽出したものを用いる．アノテーション候補の画像が複数枚存在する場合は，対象の物体が最もはっきりと見えるものを選択する．また，調理動作が動画中に収録されていない場合や物体の大部分が手や道具に隠れている場合は，アノテーションを行わず，代わりに欠損値を記録する．従って，全ての状態遷移に対して 2 枚の画像がアノテーションされているとは限らない点に関して，注意が必要である.

4　アノテーション結果

本節では，まずデータ収集とアノテーション手順について説明し，アノテーションから得られた統計量について説明する．次に，アノテーションの品質の調査を行い，最後にデータセット評価のための実験とその結果について述べる.

4.1 アノテーション手順

200 記事の日本語レシピをクラシル[3]から収集した．ここで，既存のフローグラフコーパス (Mori et al. 2014) を用いない理由は，レシピを基にした調理動画が必ずしも存在するとは限らないためである．本レシピに付属する調理動画では，各調理の手順が固定カメラによって収録されているため，固定視点での物体の視覚的な状態遷移のアノテーションが可能である．1 節で述べた通り，本研究の目標は調理レシピを入力として受け取り，実世界で調理動作を実行するエージェントの確立である．しかし，調理レシピによっては複雑な手順が含まれることがある．そこで，本研究では手始めとして比較的手順が容易であるサラダのカテゴリを対象とし，調理レシピの収集，アノテーションを行った．

本データセットの構築には複雑なアノテーションが要求されるだけでなく，対象とするデータの特性上，調理に関する経験が必要であると考えられる．従って，ここでは自然言語処理に関わるアノテーションと調理の経験を持つ 1 人のアノテータに依頼し，アノテーションを行った．アノテータの訓練は，既存の r-FG コーパス (Mori et al. 2014) から無作為に収集した 20 レシピを用い，それらにおける r-NE, r-FG の既存のアノテーションとの一致率が 80%を超えるまでアノテーションを繰り返すことで行った．また，画像アノテーションの訓練に関しては，アノテーション対象である全体 200 レシピのうちの 50 レシピを用いて，10, 20, 50 レシピのアノテーションが終了するごとに結果を確認し，仕様と異なる場合には修正の指示を出すことで行った．また，このとき，必要に応じてアノテーション仕様の修正を行った．

先行研究 (Mori et al. 2014) ではスプレッドシートに入力する方式で r-NE と r-FG のアノテーションを行っているが，この形式はコストが高く，手作業のために予期せぬアノテーションミスを生む可能性がある．従って，本研究では，フローグラフアノテーションのためのツールを開発し，これを用いてアノテーションを行った．なお，本アノテーションツールの実装はウェブ上にて公開済みである[4]．図 5 にツールを用いたアノテーションの例を示す．全アノテーションには，計 120 時間を要した．

4.2 統計量

収集したレシピには，合計 1,701 個の材料と 1,077 個の手順が含まれており，1 レシピ平均 8.51 個の材料と 5.31 個の手順で構成されていた．また，全体を通して 89 種類の調理動作と 275 種類の材料表現が確認できた．r-NE と r-FG のアノテーション結果を表 1，表 2 にそれぞれ示す．表から，r-NE アノテーションでは 11,686 個のタグが，r-FG アノテーションでは 11,291 個のラベルが，それぞれ得られていることがわかる．

視覚的アノテーション結果を表 3 に示す．表から，動作対象となる物体の総数は 3,705 であ

[3] https://www.kurashiru.com (2021/12/14).
[4] https://github.com/kskshr/Flow-Graph-Annotation-Tool.

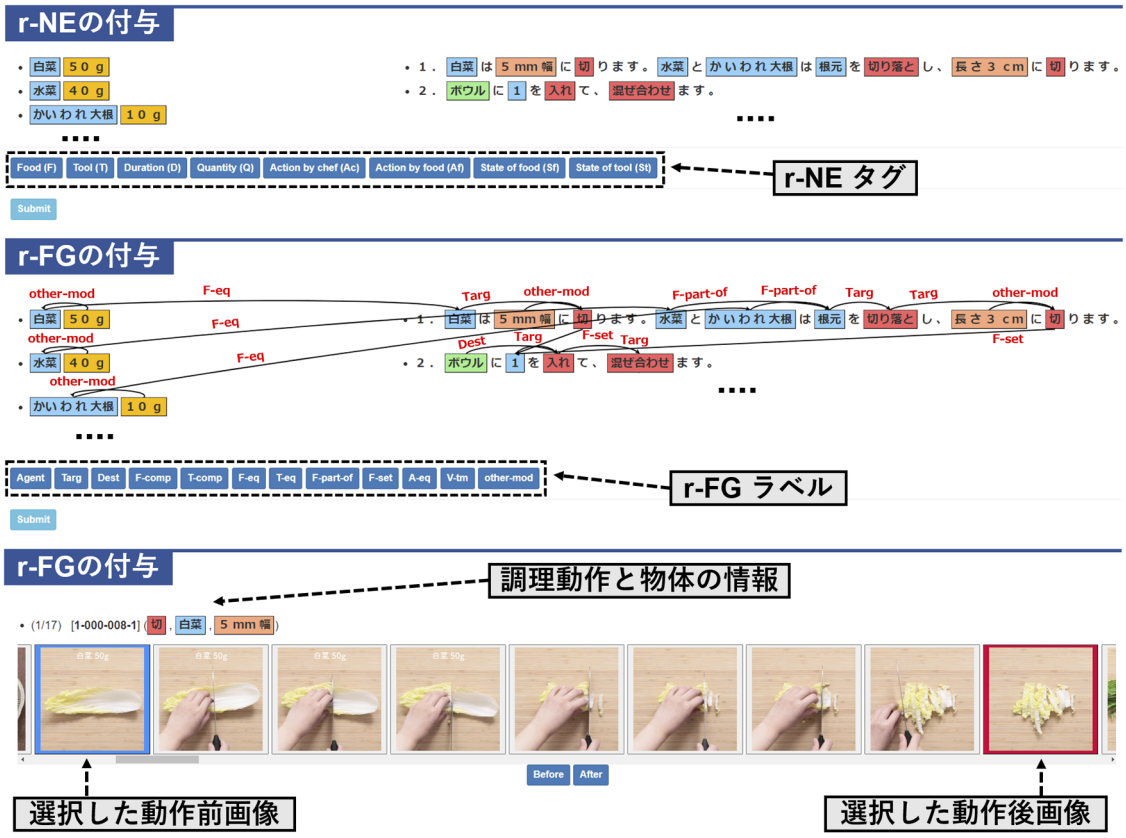

図 5　アノテーションツールを用いたアノテーションの例.

ることがわかる．このうち，2,551 例は動作前後共に，485 例は動作後のみに，72 例は動作前のみに，それぞれ画像を付与することが出来た．残りの 597 例に関しては，動作前後の双方ともに画像の付与が出来なかった．また，このアノテーションでは，合計 5,659 枚（重複なしで 3,824 枚）の画像を使用した．ここで，重複が発生しているのは，ある動作の動作後画像と次の動作の動作前画像が一致する場合があるためである．

4.3　アノテーション品質

アノテーションの品質を調査するため，4.1 節とは別の，自然言語処理と調理の経験を持つアノテータに，データセット全体から無作為に収集した 10 レシピ（全体の 5%）[5]のアノテーションを依頼し，それらの一致率を計算した．アノテータの訓練は，r-NE と r-FG に関しては 4.1 節

[5] 抽出した 10 レシピには，623 個の r-NE タグ，616 の r-FG ラベル，199 個の物体を対象とした画像の組が含まれていた．

と同様の手順で行い，画像アノテーションに関しては，4.1 節で作成した仕様を元に，アノテーション対象外の 5 レシピにおけるアノテーション結果を確認，修正した．

結果を表 4 に示す．ここでは，4.1 節で得られたアノテーションを正解データとして捉え，それらの間の精度，再現率，F 値を計算することで，一致率を算出した．r-NE, r-FG のアノテーションにおいては，それぞれ 98.40%，86.11% という非常に高い一致率が得られた．また，画像アノテーションに関しては，72.80% と前 2 段階のアノテーション結果と比較すると低い一致率となったが，前段階からのアノテーションミスの影響や，画像アノテーション時に複数の画像が候補となることを考慮すると，この値は十分に高いといえる．

4.4　実験

4.4.1　タスク設定

データセットにおける各アノテーション要素の重要性について調査するため，マルチモーダルな情報検索のタスクを考え，実験を行った．このタスクでは，調理動作を表す動詞 a と物体 o のテキスト情報，動作前画像 i^{pre} を基に，候補の画像の集合 $i_1, i_2, \cdots, i_n$ から正解の動作後画像 i^{post} を検索することを目指す．図 6 に例を示す．

4.4.2　モデル

まず，調理動作 a，物体 o のテキスト情報と動作前画像 i^{pre} から，動画後画像に対応するベクトル表現の計算を行う．ここで，a と o は BiLSTM (Lample et al. 2016) を用いて d_v 次元の

画像		物体の数
動作前	動作後	
		597
✓		72
	✓	485
✓	✓	2,551
合計		3,705

表 3　画像アノテーションの統計量．

Annotation	精度	再現率	F 値
r-NE	97.93	98.88	98.40
r-FG	86.18	86.04	86.11
Image	75.13	70.60	72.80

表 4　アノテーション一致率．

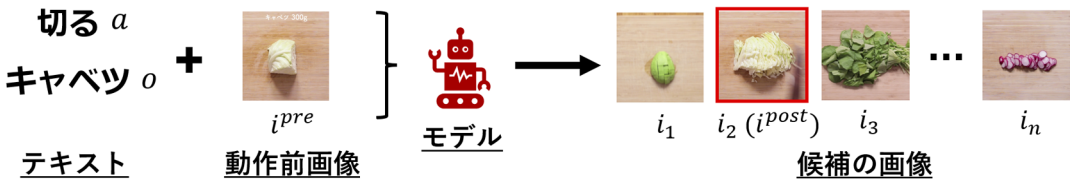

図 6　検索タスクの例．調理動作と物体に対応するテキストの情報，動作前画像をもとに対応する動作後画像の検索を行う．この例では，赤枠で囲われている画像が正しい動作後画像である．

ベクトル h_a，h_o に変換する．この h_a，h_o はレシピ全体を BiLSTM でエンコードして得られる単語の分散表現のうち，a，o に対応する表現である．a，o が複数の単語からなる場合には，それらの平均を取ることで対応する分散表現を獲得する．i^{pre} は事前学習済みの畳み込みニューラルネット (Convolutional Neural Network; CNN) を用いて d_t 次元のベクトル h_i^{pre} を以下のように計算する．

$$h_i^{\mathrm{pre}} = W_1^{\mathrm{pre}}\mathrm{CNN}(i^{\mathrm{pre}}) + b_1^{\mathrm{pre}}.$$

ここで，$W_1^{\mathrm{pre}} \in \mathbb{R}^{d_t \times d_i}$, $b_1^{\mathrm{pre}} \in \mathbb{R}^{d_t}$ は学習可能なパラメータである．これらを基に，$\tilde{h}^{\mathrm{pre}}$ を以下のように計算する．

$$\tilde{h}^{\mathrm{pre}} = W_3^{\mathrm{pre}}(\mathrm{ReLU}(W_2^{\mathrm{pre}}[h_a; h_o; h_i^{\mathrm{pre}}] + b_2^{\mathrm{pre}})) + b_3^{\mathrm{pre}}.$$

ここで ; は結合を指し，$W_2^{\mathrm{pre}} \in \mathbb{R}^{3d_t \times 3d_t}$, $W_3^{\mathrm{pre}} \in \mathbb{R}^{d_t \times 3d_t}$, $b_2^{\mathrm{pre}} \in \mathbb{R}^{3d_t}$, $b_3^{\mathrm{pre}} \in \mathbb{R}^{d_t}$ は全て学習可能である．最後に，$\tilde{h}^{\mathrm{pre}}$ を以下のように共有埋め込み空間上にマッピングする (Miech et al. 2018) ことで，動作後画像検索のための表現 $\hat{h}^{\mathrm{pre}}$ を獲得する．

$$\hat{h}^{\mathrm{pre}} = \frac{f(\tilde{h}^{\mathrm{pre}})}{||f(\tilde{h}^{\mathrm{pre}})||_2}.$$

ここで，

$$f(h) = (W_4^{\mathrm{pre}}h + b_4^{\mathrm{pre}}) \circ \sigma(W_5^{\mathrm{pre}}(W_4^{\mathrm{pre}}h + b_4^{\mathrm{pre}}) + b_5^{\mathrm{pre}})$$

である．また，σ はシグモイド関数であり，$\circ$ は要素ごとの掛け算を指す．$W_4^{\mathrm{pre}} \in \mathbb{R}^{d_e \times d_t}$, $W_5^{\mathrm{pre}} \in \mathbb{R}^{d_e \times d_e}$, $b_4^{\mathrm{pre}}, b_5^{\mathrm{pre}} \in \mathbb{R}^{d_e}$ は全て学習可能である．

動作後画像 i^{post} は動作前画像と同様に，事前学習済み CNN を用いて d_t 次元のベクトル h_i^{post} を以下のように計算する．

$$\tilde{h}_i^{\mathrm{post}} = W_2^{\mathrm{post}}(\mathrm{ReLU}(W_1^{\mathrm{post}}\mathrm{CNN}(i^{\mathrm{post}}) + b_1^{\mathrm{post}})) + b_2^{\mathrm{post}}.$$

$W_1^{\mathrm{post}}, W_2^{\mathrm{post}} \in \mathbb{R}^{d_i \times d_i}$, and $b_1^{\mathrm{post}}, b_2^{\mathrm{post}} \in \mathbb{R}^{d_i}$ は学習可能であり，同様に共有埋め込み空間上へのマッピングを行う．

$$\hat{h}^{\mathrm{post}} = \frac{g(\tilde{h}^{\mathrm{post}})}{||g(\tilde{h}^{\mathrm{post}})||_2}.$$

ここで，

$$g(h) = (W_3^{\mathrm{post}}h + b_3^{\mathrm{post}}) \circ \sigma(W_4^{\mathrm{post}}(W_3^{\mathrm{post}}h + b_3^{\mathrm{post}}) + b_4^{\mathrm{post}})$$

である．$W_3^{\mathrm{post}} \in \mathbb{R}^{d_e \times d_i}$, $W_4^{\mathrm{post}} \in \mathbb{R}^{d_e \times d_e}$, and $b_3^{\mathrm{post}}, b_4^{\mathrm{post}} \in \mathbb{R}^{d_e}$ は学習可能である．検索候補と

なる他の画像も同様に，上記の計算を経て共有埋め込み空間上へのマッピングを行う．

誤差関数. 共有埋め込み空間上での $\hat{h}^{\mathrm{pre}}$, $\hat{h}^{\mathrm{post}}$ 間の距離は以下のように計算する．

$$D(\hat{h}^{\mathrm{pre}}, \hat{h}^{\mathrm{post}}) = ||\hat{h}^{\mathrm{pre}} - \hat{h}^{\mathrm{post}}||_2.$$

これを基に，n 例のベクトルの組 $((\hat{h}_1^{\mathrm{pre}},\hat{h}_1^{\mathrm{post}}), \cdots, (\hat{h}_n^{\mathrm{pre}}, \hat{h}_n^{\mathrm{post}}))$ が与えられた時，以下の誤差関数の最小化を考える (Balntas et al. 2016):

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

ここで，$D_{i,j} = D(\hat{h}_i^{\mathrm{pre}}, \hat{h}_j^{\mathrm{post}})$ であり，δ はマージンを指す．$D_{i,i}$ は i 番目の正例のベクトル間の距離を表す．$D_{i,j}$ は h_i^{post} の負例として h_j^{post} を，$D_{k,i}$ は h_i^{pre} の負例として h_k^{pre} を，それぞれ用いた時の距離を表す．負例のサンプリングに関しては，ここでは単純にミニバッチ内から無作為に選択することを考える．

4.4.3 実験設定

データ分割. 本実験では，アノテーションで得られた動作前後の観測のうち，その双方がアノテーションされている 2,551 例のみを用いた．全体の 200 レシピを 10 分割し，そのうち 1 分割（20 レシピ）をテストデータ，残りの 9 分割を結合したもの（180 レシピ）を学習データとして割り当てた．このテストデータに対応する分割を変更することで，後述する 10 分割交差検証を実現した．ここで，学習データ，テストデータに含まれるサンプル数はそれぞれ平均で 2296.8 例，255.2 例であった．また，動作後画像の候補は，学習時はミニバッチ内の他の動作後画像を，テスト時にはテストセットに含まれる他の全ての動作後画像を，それぞれ用いた．各動作前後の画像のペアは手順リスト中の動作表現に紐付いているが，これを動詞のテキスト情報として用いた．また，物体のテキスト情報はこの動詞を終点とし，r-FG の Targ ラベルで紐付けられている始点の表現のうち，r-NE の F（食材）または Ac（調理者の動作）が振られているものを選択することで，自動的に特定し，用いた．

モデルパラメータ. 実験では 1 層で 256 次元の BiLSTM を用いてテキストのエンコードを行った．また，次元数に関しては，$(d_v, d_t, d_i, d_e) = (496, 512, 2048, 128)$ とした．事前学習済み CNN には ResNet-152 (He et al. 2016) を用い，2048 次元の特徴量を画像から抽出した．

最適化. 最適化手法として，AdamW (Loshchilov and Hutter 2019) を初期学習率 1.0×10^{-5} として用いた．学習中は，CNN のパラメータは固定した．各モデルの学習には，4 レシピからミニバッチを構成し，350 エポック分行った．式 (1) の δ は実験的に 0.1 に設定した．また，データセットのうち 90% を学習データに，10% をテストデータに分割することで，10 分割交差検証を行い，モデルの評価を行った．

評価指標. 評価指標として，Recall@5 (R@5) と Median rank (MedR) を用いた．ここで，R@5 は検索結果の上位 5 件に正解画像が含まれている割合を調査するために，MedR は正解画像の候補画像中における順位を調査するために，それぞれ用いている．また，これらの指標はクロスモーダルな検索タスクの評価において，先行研究でも用いられている (Socher et al. 2014; Salvador et al. 2017; Miech et al. 2019).

4.4.4 実験結果

動作後画像の検索に用いるテキスト情報 (a, o) と動作前画像 (i^{pre}) に関して，いずれか片方を用いる場合と両方を用いる場合を考え，それぞれ実験を行った．表 5 に実験結果を示す．1 行目はランダムベースラインを表している．2 行目と 3 行目の比較から，画像かテキストのいずれか一方を用いる場合では，画像を用いた方がより良い R@5 (33.77) と MedR (12.60) を獲得出来ていることがわかる．次に，これらと 4 行目の比較から，画像とテキストの両方を用いることで，R@5 で 3.24, MedR で 2.2 の改善が得られていることがわかる．これらの結果は，動作後画像の検索において，画像はテキスト以上に重要な情報を与えていることを示唆している．また，テキストと画像を組み合わせることでさらなる改善が得られることから，テキストは画像とは異なる手がかりをモデルに与えていることが推察できる．

4.4.5 考察

本タスクに残された課題としては，(i) 動作前後で変化が少ない例と (ii) 動作結果の画像の予測の際に，過去の動作を遡って解析する必要がある例への対処が挙げられる．(i) は例えば，「ドレッシングの材料をボウルに加える」等が対応する．この例では，加える食材に応じた動作結果の画像を検索できることが望ましいが，食材が塩や油などの場合，動作前後で見た目の変化が起こりにくいという問題があり，実際にこういった例に対しては，モデルによる検索の失敗が多く見受けられた．これの解決策としては，例えば，加える原材料の見た目や形状に関する特徴量を予め用意しておき，$\hat{h}^{\mathrm{pre}}$ の計算時に用いること等が考えられる．

入力		R@5 (↑)	MedR (↓)
テキスト	動作前画像		
		2.37	149.00
✓		21.24	26.70
	✓	33.77	12.60
✓	✓	37.01	10.40

表 5 実験結果．テキストは調理動作と物体の表現を指している．また，枠内のチェックマークは動作後画像の検索に用いた要素を表している．

(ii) に関しては，物体のテキスト情報が過去の手順の番号[6]や過去の動作[7]となる場合が挙げられる．これの解決策としては，例えば，対象の動作からフローグラフを逆に辿ることで，過去の動作やそれに含まれる食材の情報を利用することが考えられる．このとき，r-NE タグの情報を合わせて用いることで，物体の形状や原材料まで特定することが可能となる．また，Neural Process Networks (Bosselut et al. 2018; Nishimura et al. 2021) のように，食材の状態遷移を追跡する機構を用いて，実際の動作後画像の検索時に用いることも考えられる．

さらに，これらの課題に加え，本実験で用いた以外の r-NE タグ，r-FG ラベルの情報を用いてさらなる精度向上を図ることも期待できる．例えば，r-NE の Sf（食材の状態）タグや r-FG の Dest（動作の方向）は動作や物体に関するより細かい情報を提供できるため，モデルの検索精度向上に利用できる可能性がある．

5 応用

本節では，本データセットの応用先について述べる．

5.1 マルチモーダル常識推論

マルチモーダルな情報を用いた手順書の常識推論は近年注目を集めている研究の一つである (Zellers et al. 2019; Yagcioglu et al. 2018; Alikhani et al. 2019)．調理分野においては，食材の状態遷移に関する推論 (Bosselut et al. 2018; Nishimura et al. 2021) が存在する．この方向において，本データセットを用いることで，食材の視覚的な状態遷移を考慮した推論を行うことが可能となる．また，r-FG が表すワークフローを利用して，各調理動作がレシピ全体に与える影響について分析することも可能である．

5.2 手順書生成

画像や動画等，視覚的な情報からの手順書生成 (Ushiku et al. 2017; Nishimura et al. 2019) は，作業の再現性を高める上で重要なタスクであるといえる．ここで，正しく作業を再現するためには，生成された手順列が一貫性を持つことが重要である．r-FG は手順間の依存関係を表現するため，これを利用することで，より一貫した手順書の生成が行える可能性がある．

[6] 「1 をボウルに加える」の「1」等.

[7] 例えば，図 4 の手順 2 では，「入れる」と「混ぜ合わせる」の 2 つの動作が登場しているが，後者の動作に対応する物体のテキスト情報は r-FG の性質上，「入れる」となる．これは「入れる」という動作の生成物が「混ぜ合わせる」の入力となっていることを表している．

6 おわりに

本稿では，調理レシピにおいて，調理動作後の物体の視覚的状態の予測を目指し，Visual Recipe Flow データセットを提案した．データの収集方法やアノテーション手順について述べ，アノテータ間の一致率を測ることでデータセットの品質を調査した．また，マルチモーダルな情報を用いた情報検索タスクを考えることで，各アノテーション要素の重要性について調べた．最後に，本データセットの応用例として，マルチモーダル常識推論や手順書生成について述べた．

謝 辞

本研究は JSPS 科研費 JP20H04210，JP21H04910，JP22K17983 と JST，さきがけ，JPMJPR20C2 の助成を受けたものです．また，研究の過程でオムロンサイニックエックス株式会社の牛久祥孝氏にはご助言をいただきました．

本論文の一部は言語処理学会第 28 回年次大会 (白井他 2022) および The 29th International Conference on Computational Linguistics (COLING 2022) (Shirai et al. 2022) で発表したものです．

参考文献

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). “Flamingo: A Visual Language Model for Few-shot Learning.” *arXiv preprint arXiv:2204.14198*.

Alikhani, M., Nag Chowdhury, S., de Melo, G., and Stone, M. (2019). “CITE: A Corpus of Image-Text Discourse Relations.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 570–575.

Balntas, V., Riba, E., Ponsa, D., and Mikolajczyk, K. (2016). “Learning Local Feature Descriptors with Triplets and Shallow Convolutional Neural Networks.” In *Proceedings of the British Machine Vision Conference*, pp. 119.1–119.11.

Bosselut, A., Levy, O., Holtzman, A., Ennis, C., Fox, D., and Choi, Y. (2018). “Simulating Action Dynamics with Neural Process Networks.” In *Proceedings of the 6th International Conference on Learning Representations*.

Dalvi, B., Huang, L., Tandon, N., Yih, W.-t., and Clark, P. (2018). “Tracking State Changes in Procedural Text: a Challenge Dataset and Models for Process Paragraph Comprehension.”

In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1595–1604.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep Residual Learning for Image Recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Isola, P., Lim, J. J., and Adelson, E. H. (2015). "Discovering States and Transformations in Image Collections." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). "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. Association for Computational Linguistics.

Loshchilov, I. and Hutter, F. (2019). "Decoupled Weight Decay Regularization." In *Proceedings of the 7th International Conference on Learning Representations*.

Maeta, H., Sasada, T., and Mori, S. (2015). "A Framework for Procedural Text Understanding." In *Proceedings of the 14th International Conference on Parsing Technologies*, pp. 50–60.

Miech, A., Laptev, I., and Sivic, J. (2018). "Learning a Text-Video Embedding from Incomplete and Heterogeneous Data." *arXiv preprint arXiv:1804.02516*.

Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. (2019). "Howto100m: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2630–2640.

Mori, S., Maeta, H., Yamakata, Y., and Sasada, T. (2014). "Flow Graph Corpus from Recipe Texts." In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pp. 2370–2377.

Neubig, G., Nakata, Y., and Mori, S. (2011). "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.

Nishimura, T., Hashimoto, A., and Mori, S. (2019). "Procedural Text Generation from a Photo Sequence." In *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 409–414.

Nishimura, T., Hashimoto, A., Ushiku, Y., Kameko, H., and Mori, S. (2021). *State-Aware Video Procedural Captioning*, pp. 1766–1774. Association for Computing Machinery, New York, NY, USA.

Nishimura, T., Tomori, S., Hashimoto, H., Hashimoto, A., Yamakata, Y., Harashima, J., Ushiku, Y., and Mori, S. (2020). "Visual Grounding Annotation of Recipe Flow Graph." In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4275–4284.

Pan, L.-M., Chen, J., Wu, J., Liu, S., Ngo, C.-W., Kan, M.-Y., Jiang, Y., and Chua, T.-S. (2020). *Multi-Modal Cooking Workflow Construction for Food Recipes*, pp. 1132–1141. Association for Computing Machinery.

Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., and Torralba, A. (2017). "Learning Cross-modal Embeddings for Cooking Recipes and Food Images." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Shirai, K., Hashimoto, A., Nishimura, T., Kameko, H., Kurita, S., Ushiku, Y., and Mori, S. (2022). "Visual Recipe Flow: A Dataset for Learning Visual State Changes of Objects with Recipe Flows." In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3570–3577. International Committee on Computational Linguistics.

白井圭佑，橋本敦史，牛久祥孝，栗田修平，亀甲博貴，森信介 (2022). 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. 言語処理学会第 28 回年次大会, pp. 129–134. [K. Shirai et al. (2022). Construction of a Dataset Considering State Changes of Objects in the Cooking Domain and Proposal of a Retrieval Model for the Dataset. Proceedings of the 28th Annual Meeting of the Association for Natural Language Processing, pp. 129–134.].

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. (2014). "Grounded Compositional Semantics for Finding and Describing Images with Sentences." *Transactions of the Association for Computational Linguistics*, pp. 207–218.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2022). "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models." *arXiv preprint arXiv:2206.04615*.

Tandon, N., Sakaguchi, K., Dalvi, B., Rajagopal, D., Clark, P., Guerquin, M., Richardson, K., and Hovy, E. (2020). "A Dataset for Tracking Entities in Open Domain Procedural Text." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6408–6417.

Ushiku, A., Hashimoto, H., Hashimoto, A., and Mori, S. (2017). "Procedural Text Generation from an Execution Video." In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 326–335.

Yagcioglu, S., Erdem, A., Erdem, E., and Ikizler-Cinbis, N. (2018). "RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes." In *Proceedings of the 2018*

Conference on Empirical Methods in Natural Language Processing, pp. 1358–1368.

Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. (2019). "From Recognition to Cognition: Visual Commonsense Reasoning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, Y., Yamakata, Y., and Tajima, K. (2021). "MIRecipe: A Recipe Dataset for Stage-Aware Recognition of Changes in Appearance of Ingredients." In *Proceedings of the 3rd ACM International Conference on Multimedia in Asia*, pp. 1–7. Association for Computing Machinery.

略歴

白井　圭佑：2017 年愛媛大学工学部卒業．2020 年京都大学大学院情報学研究科修士課程修了．現在同大学院博士課程．自然言語処理，マルチメディアに関する研究に従事．言語処理学会会員．

橋本　敦史：2005 年京都大学工学部卒業．2013 年京都大学大学院情報学研究科にて博士（情報学）取得．2012 年より同大助手，2015 年より同大助教として勤務．2018 年よりオムロンサイニックエックス株式会社シニアリサーチャー．2021 年より慶應義塾大学特任講師を兼任して現在に至る．主に機械学習，画像処理，および，料理や組立作業を対象とした人と機械のインタラクション理解に関する研究に従事．IEEE，IEICE，IPSJ 各会員．

西村　太一：2019 年九州大学芸術工学部卒業．2020 年京都大学大学院情報学研究科修士課程修了．現在同大学博士課程．修士（情報学）．マルチメディア，自然言語処理，コンピュータビジョンの研究に従事．学術振興会特別研究員 (DC1)．2023 年言語処理学会論文賞受賞．

亀甲　博貴：2018 年東京大学大学院工学系研究科博士課程修了．博士（工学）．同年より京都大学学術情報メディアセンター助教．自然言語処理，ゲーム AI 等に関する研究に従事．言語処理学会，情報処理学会各会員．

栗田　修平：2013 年京都大学理学部卒業．2015 年同大学院理学研究科物理学教室修士課程修了．2019 年京都大学大学院情報学研究科にて博士（情報学）取得．現在国立研究開発法人理化学研究所革新知能統合研究センター研究員．深層学習を用いた自然言語処理ならびにコンピュータビジョンの研究に従事．言語処理学会，情報処理学会，人工知能学会，ACL 各会員．

森　　信介：1998 年京都大学大学院工学研究科電子通信工学専攻博士後期課程修了．同年日本アイ・ビー・エム株式会社入社．2007 年より京都大学学術情報メディアセンター准教授．2016 年同教授．現在に至る．計算言語学ならび

に自然言語処理の研究に従事．博士（工学）．1997 年情報処理学会山下記念研究賞受賞．2010 年，2013 年情報処理学会論文賞受賞．2010 年第 58 回電気科学技術奨励賞．2023 年言語処理学会論文賞受賞．言語処理学会，情報処理学会，日本データベース学会各会員．

（2023 年 2 月 27 日　受付）
（2023 年 5 月 24 日　再受付）
（2023 年 6 月 16 日　採録）