レシピフローグラフへのVisual Groundingアノテーション

西村 太一[1]，友利 涼[1]，橋本 隼人[1]，橋本 敦史[2]，山肩 洋子[3]，原島 純[4]，牛久 祥孝[2]，森 信介[5]
[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] 東京大学大学院 情報理工学系研究科，[4] クックパッド株式会社，
[5] 京都大学 学術情報メディアセンター

{nishimura.taichi.43x,tomori.suzushi.72e,hashimoto.hayato.73e}@st.kyoto-u.ac.jp,
{atsushi.hashimoto, yoshitaka.ushiku}@sinicx.com, yamakata@mi.u-tokyo.ac.jp,
jun-harashima@cookpad.com, forest@i.kyoto-u.ac.jp

1 はじめに

手順書は手順の流れを示すフローグラフ [6] で表現することで，その意味表現を定義することができる．この意味表現を料理ロボット [2] などの実世界上で動作するシステムに応用する上で，手順書の記述内容と画像や動画中の物体との間でグラウンディング (Visual grounding) を行うことは重要な課題である．手順書においては，手順に応じて物体が視覚的に変化するため，手順の流れを考慮しながら Visual grounding が行えることが望ましい．この文脈を加味した Visual grounding を本研究では特に *Contextual visual grounding* と呼ぶ．

本研究ではドメインを料理ドメインに限定し，Contextual visual grounding に対応するデータセットを提案する．図 1 にアノテーションの概要を示す．手順書と各手順に対応する写真，フローグラフが与えられ，アノテータはフローグラフの頂点に対応する物体領域をバウンディングボックス (BB) を用いてアノテーションを行う．加えて，各 BB に “動作中”，“動作済” の 2 つの状態を示すラベルも付与する．これらのアノテーションにより，モデルが画像の物体領域と手順書の記述内容の間で文脈を加味しながらグラウンディングができるようになる．アノテーションの結果，本研究では，272 のフローグラフコーパスとそれに付与した 2,300 の BB を含んだレシピフローグラフ BB(r-FG-BB) データセットを構築した．

2 アノテーション方法

2.1 レシピフローグラフ (r-FG) コーパス

r-FG コーパスは，レシピテキストに対して，手順内容を無閉路有向グラフ (DAG) で表現したフローグラフからなるコーパスである．図 1 のフローグラフの項目にその例を示す．フローグラフの各頂点はレシピ中の重要な表現であるレシピ固有表現 (r-NE) を表し，それらの関係性をラベル付きの辺で繋ぐことで手順の流れを表現している．表 1 に r-NE の種類と辺のラベルの一覧を示す．

表 1: r-NE のタグ一覧 (左) と，辺のラベル一覧 (右).

タグ名	意味
F	食材
T	道具
D	継続時間
Q	分量
Ac	調理者の動作
Af	食材の動作
Sf	食材の状態
St	道具の状態

ラベル名	意味
Agent	主語
Targ	対象
Dest	方向
F-comp	食材デ
T-comp	道具デ
F-eq	同一の食材
F-part-of	食材の一部
F-set	食材の集合
T-eq	同一の道具
T-part-of	道具の一部
A-eq	同一の動作
V-tm	動作を行う
other-mod	その他

2.2 BB アノテーション

本論文で提案する r-FG-BB データセットは，r-FG コーパスの各レシピの画像に BB を付与し，各 BB とフローグラフの頂点である r-NE との間でアノテーションを行なうことで拡張したものである．今回は r-NE のタグの 8 種類のうち，写真から判別がしやすい F，T，Ac の 3 種類に限定してアノテーションを実施した．

図 2 に F，T，Ac のそれぞれのアノテーションルールを示す．未加工の材料が写っていた場合にその領域を BB で囲い，F とアノテーションする．また，道具が画像内に一部分でも写っていた場合はその領域を BB で囲い，T とアノテーションする．Ac に付与する場合は，Ac の示す調理者の動作の結果得られた中間生成物を BB で囲い，アノテーションを行う．

2.3 動作属性

Ac に対して BB をアノテーションする場合は，BB がその動作中の状態を表すか (Ac_{ing})，動作後の結果の状態を表すか (Ac_{ed}) という動作の属性もアノテーションする．本論文では，これを動作属性と呼ぶ．例えば，図 2 の Ac の項目の “切る” という動作と BB との間では，上では動作中 (Ac_{ing}) というラベルが割り当てられ，下では，動作済 (Ac_{ed}) が割り当てられる．いずれの属性か判別するのが難しい場合，アノテータは動作不明 (Ac_{unc}) というラベルを付与する．

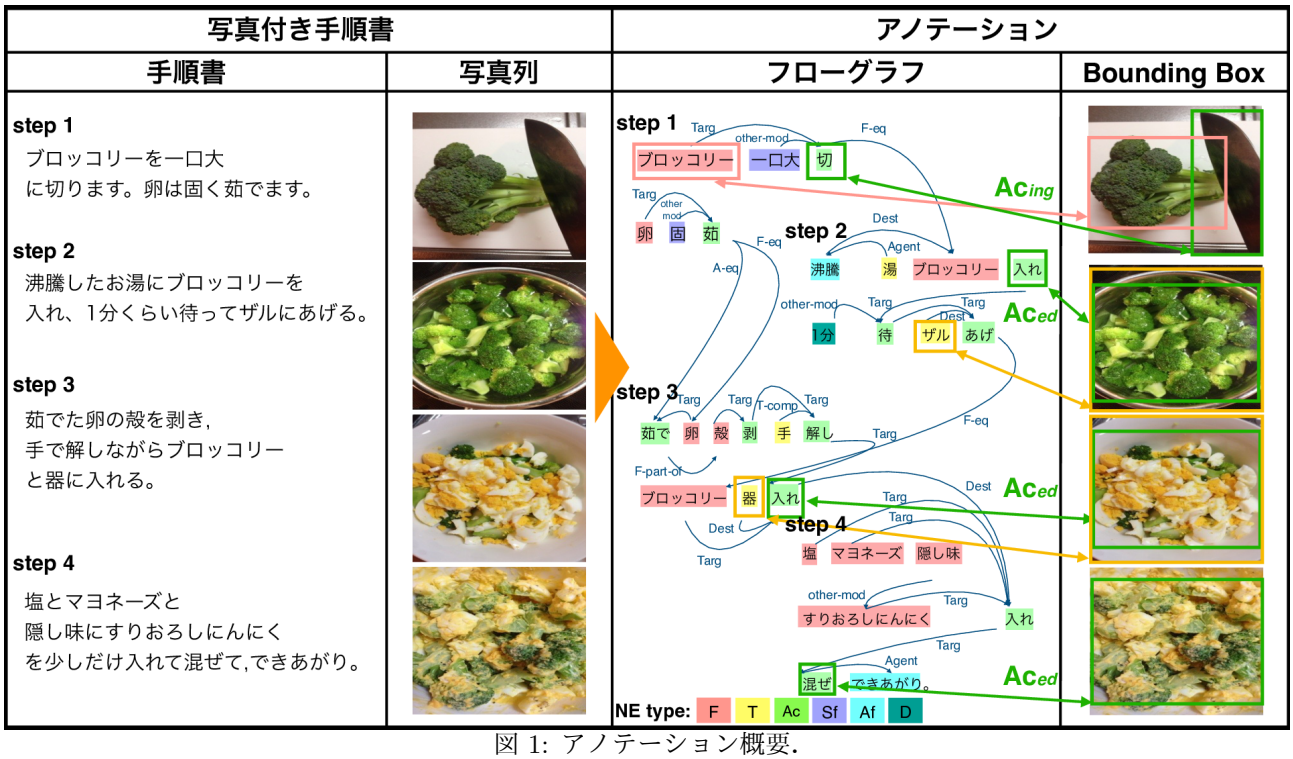

図 1: アノテーション概要.

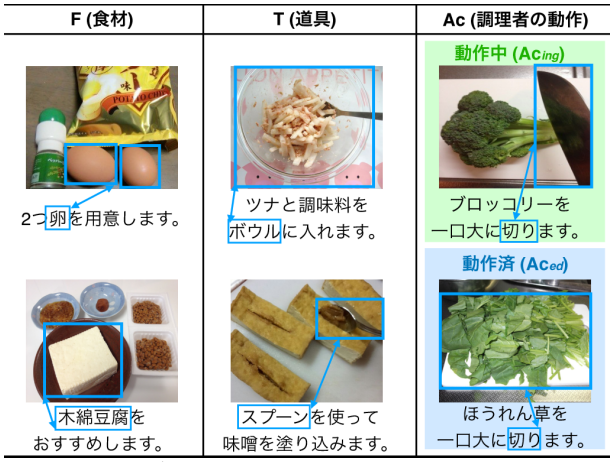

図 2: アノテーションルール.

3 アノテーション結果

272 レシピのフローグラフに対し，前節の手順で 2 名のアノテータにアノテーションを行ってもらった．本節では，2 名のアノテーション結果の一致率とそこから r-FG-BB データセットを構築する手順について説明する．

3.1 一致率

2 名のアノテータ A, B のアノテーション結果 S_A, S_B を用いて，r-NE，BB のアノテーションの一致率を以下の IoU を用いて測定した．

$$\mathrm{IoU}(S_A, S_B) = \frac{|S_A \cap S_B|}{|S_A \cup S_B|}, \qquad (1)$$

3.1.1 r-NE

r-NE の場合，S には BB に紐づいた r-NE の集合が当たる．r-NE の一致率を計算した結果，F，T，Ac で一致率がそれぞれ 81%，88%，76%となり，マイクロ平均は 79%となった．3 種類の r-NE 対象の中で，T がもっとも高い一致率を示し，Ac がもっとも低い一致率となった．T がもっとも高くなった理由として，道具は形状が変化しないため，アノテータの間で一致しやすくなったためであると考えられる．一方で，Ac に対して BB をアノテーションする場合，物体は動作中あるいは動作後の中間状態であり，アノテータの間でどの動作に付与するべきかが分かれたと考えられる．

3.1.2 BB

BB の場合，S は BB の矩形領域が当たる．アノテータの付与した BB のそれぞれの組に対して，IoU を計算した結果，IoU は 0.77 という結果となり，十分に高い一致率となった[1]．

3.1.3 動作属性

Ac にアノテーションを行う場合，アノテータは同時に動作属性も付与しなければならない．2 名のアノテータの動作属性の結果から，混合行列を計算し，その一致率を計算した結果，動作属性は 87%が一致したことが分かった．

3.2 r-FG-BB データセットの構築手順

以下の手順で r-FG-BB データセットを構築した．

[1] コンピュータビジョンの領域では，IoU が 0.5 を越えると BB が十分に重なったとしてデータセットの構築手法を評価している [7].

表 2: レシピあたりの Bounding box に付与した r-NE 数と，r-NE あたりの Bounding box の数．

	F	T	Ac	(Ac_{ing}, Ac_{ed})	合計
#r-NEs	0.16	1.28	2.44	(0.62, 1.82)	3.88
#BBs/#r-NEs	2.00	2.00	2.28	(2.59, 2.18)	2.18

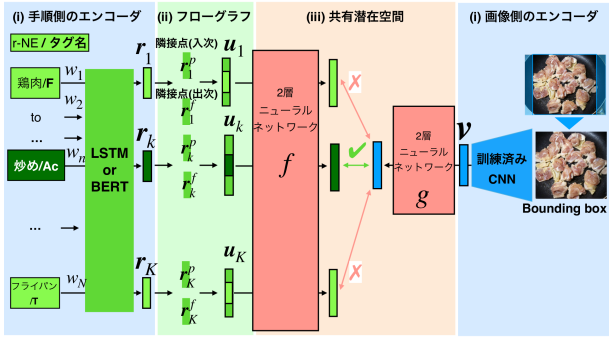

図 3: シンボルグラウンディングへの提案手法の概要．

1. 2 名のアノテータ間で r-NE の結果が一致したもののみを選択する．その結果，アノテーション結果のうち，79%の r-NE が抽出された．その後，Ac に関しては，動作属性が Ac_{unc} の場合のみを除外し，それ以外を残す．

2. その中から，2 名のアノテーションした BB 間で IoU が 0.7 を超えたもののみを抽出する．

最終的に，r-FG-BB データセットとして，272 レシピ，2,300 の r-NE と BB の組を得た．表 2 にレシピごとの BB に付与した r-NE の平均と r-NE ごとの BB 数の平均を示す．この表では，Ac_{ing} と Ac_{ed} は 2 名のアノテータ間で一致した場合にのみ数え上げている．

4 実験

本論文では，r-FG-BB データセットが Contexual visual grounding に有効か否かを調べるために，2 つの実験を行なった．本節ではその結果を報告する．

4.1 シンボルグラウンディング

4.1.1 問題設定

Contextual visual grounding に r-FG-BB データセットが有効であることを示すために，画像中の物体領域が手順書内のどの部分に対応するのかを正確に予測する問題 (シンボルグラウンディング) に取り組んだ．この問題は BB と対応する手順 $W = (w_1, w_2, \ldots, w_n, \ldots, w_N)$($w_n$ は n 番目の手順中の単語，N は W の単語数) が与えられた時，モデルが手順中の r-NE 列 $R = (r_1, r_2, \ldots, r_k, \ldots, r_K)$(ただし，$r_k$ は k 番目の F，T，Ac に属する r-NE を，K は手順中の r-NE 数) の中から BB に紐づいた r-NE である $\hat{r}_k$ を選択するという問題として定式化できる．

4.1.2 提案手法

この問題に取り組む上で考慮しなければならない点は，手順 W ごとに異なる r-NE 数を持つという点である．そのため，BB と手順を共有潜在空間に埋め込む手法を用いる．この手法により，BB と r-NE 列中の全ての r-NE との組み合わせで類似度を計算し，最も高い類似度のものを選択することで，手順 W が異なる r-NE 数を保持していても対応することができる．

図 3 に提案手法の概要を示す．(i) BB と r-NE 列を特徴ベクトルに変換する．BB に対しては，画像のエンコーダを用いて画像ベクトル $\boldsymbol{v}$ を得る．同じように，手順 W を手順のエンコーダに通し，手順中の単語ベクトル列 $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n, \ldots, \boldsymbol{w}_N)$ を得て，k 番目の r-NE の特徴ベクトルを $\boldsymbol{W}$ から取り出し，r-NE ベクトル列 $\boldsymbol{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_k, \ldots, \boldsymbol{r}_K)$ を得る．(ii) モデルがフローグラフを参照する時は，r-NE 列から 2 つの特徴ベクトルを $\boldsymbol{r}_k$ に結合する．それは，$\boldsymbol{r}_k$ がフローグラフ上で終点となる頂点集合の平均ベクトル $\boldsymbol{r}_k^p$ と，$\boldsymbol{r}_k$ がフローグラフ上で始点となる頂点集合の平均ベクトル $\boldsymbol{r}_k^f$ である．これらの特徴ベクトルが，フローグラフ上の前後関係の情報を含んでいると期待し，r-NE ベクトルを改めて $\boldsymbol{u}_k = \mathrm{concat}(\boldsymbol{r}_k^p, \boldsymbol{r}_f, \boldsymbol{r}_k^h)$ として得る．ここで，concat(·) はベクトルの結合を表す．なお，フローグラフ表現を用いない時は，r-NE ベクトル $\boldsymbol{u}_k = \boldsymbol{r}_k$ である．(iii) これらの特徴ベクトルを用いて，2 つのニューラルネットワーク f (テキスト側)， g (画像側) を r-NE ベクトルと BB のベクトルの間で埋め込み表現を得られるように Triplet margin loss [1] を用いて同時に学習を行う．

表 3: Symbol grounding の実験結果．

ベースライン (無作為に選択)	再現率	適合率	F1
F	0.023	**0.250**	0.042
T	0.091	0.039	0.054
Ac	0.429	0.221	0.291
Total	0.189	0.174	0.181
LSTM，フローグラフなし			
F	**0.250**	**0.250**	**0.250**
T	0.480	0.461	0.471
Ac	0.634	0.662	0.648
Total	0.580	0.592	0.586
LSTM，フローグラフあり			
F	0.125	**0.250**	0.167
T	0.772	0.654	0.708
Ac	0.612	0.603	0.607
Total	0.608	0.602	0.605
BERT，フローグラフなし			
F	0.000	0.000	0.000
T	0.720	**0.750**	0.734
Ac	0.762	0.716	0.739
Total	0.733	0.717	0.725
BERT，フローグラフあり			
F	0.000	0.000	0.000
T	**0.782**	**0.750**	**0.766**
Ac	**0.785**	**0.761**	**0.772**
Total	**0.734**	**0.750**	**0.742**

4.1.3 結果

r-FG-BB データセットの全ての r-NE と BB のペアを用いて実験を行なった．この実験では，r-FG-BB データセットの 80%を学習に，10%を検証に，10%を評価に用いた．画像のエンコーダとして，ImageNet [3] で訓練済みの ResNet-50 [5] を用いた．手順側のエ

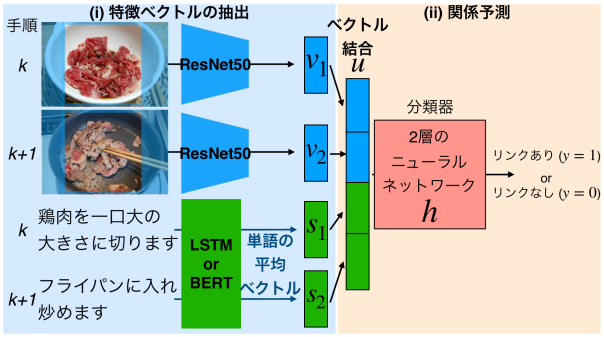

図 4: BB の関係予測への提案手法の概要.

ンコーダとして，LSTM と BERT [4] の 2 種類のエンコーダを，50 万のレシピで学習させたものを用いて実験を行なった．

表 3 にその結果を示す．テキスト側からのエンコーダの種類に関係なく，フローグラフを参照するモデルの方がフローグラフを参照しない場合と比べて高い性能を示した．このことから，フローグラフから得られるグラフ上の前後関係の情報はモデルがシンボルグラウンディングを行う上で役に立つことが分かる．また，テキスト側のエンコーダの間で比較した時，LSTM よりも BERT の方がより高い性能を示した．全体を通してどの手法でも F の結果が T や Ac に比べて低いことが分かる．これは，表 2 にあるように，F のアノテーション数は極めて少ないため，F を正しく予測するようにモデルを学習することができなかったためであると考えられる．

4.2 BB の関係予測

4.2.1 問題設定

r-FG-BB データセットでは，各 BB がレシピフローグラフの頂点と結びついている．そのため，BB 間に関係があるか否かを BB に紐づいたフローグラフ上のノード同士が繋がっているか否かで正解データ ($y = 1$ or 0) を作成することができる．この正解データを用いて，2 つの連続する手順に付与した BB と対応する手順を与え，BB 間に関係があるか否か ($\hat{y} = 1$ or 0) を予測する問題に取り組んだ．

4.2.2 提案手法

図 4 にモデルの概要を示す．入力は k 番目と，$(k+1)$ 番目の手順と，それらに対応する画像中の BB である．(i) BB を ResNet50 を通じてそれぞれベクトル $\boldsymbol{v}_1, \boldsymbol{v}_2$ に変換する．一方，手順は LSTM や BERT に入力し，出力の各単語の平均ベクトル $\boldsymbol{s}_1, \boldsymbol{s}_2$ を手順ベクトルとして得る．(ii) 次に，得られたベクトルを結合して特徴ベクトル $\boldsymbol{u} = \mathrm{concat}(\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{s}_1, \boldsymbol{s}_2)$ を得る．最終的に得られた特徴ベクトルを 2 層の全結合層と，最終層にシグモイド関数を持つニューラルネットワークに与えることで関係が存在する確率を出力する．

4.2.3 結果

比較手法として，Random と Cosine similarity の 2 つのベースラインを用意した．Random は，ランダムに BB 間で関係があるか否かを決定する．この手法は，作成した正解データに偏りがあるかどうかを調べるために用意した．Cosine similarity は，モデルが 2 つの画像ベクトル $\boldsymbol{v}_1, \boldsymbol{v}_2$ の間で余弦類似度を計算し，その類似度が α を超えている場合に関係があると予測するベースラインである．このベースラインは，視覚的に類似する場合は同じ物体であろうと考え用意したものである．提案手法のモデルを学習させるために，r-FG-BB データセットを 80%を訓練用，10%を検証用，10%をテストに分割し，実験を行なった．

表 4: 手法ごとの正解率.

ベースライン	正解率
Random	0.483
Cosine similarity ($\alpha = 0.8$)	0.533
提案手法	
画像	0.667
画像 + LSTM	0.583
画像 + BERT [4]	**0.817**

表 4 に，ベースラインと提案手法の結果を示す．ベースラインの結果より，提案手法のような 2 値分類モデルを用いなければ解くのは難しいことが分かる．提案手法では画像のみで学習した場合に比べて BERT を用いた場合は高い性能で関係を予測することができた．LSTM を用いた場合は，BERT に比べて低い性能となった．このことから，このタスクにおいても言語モデルの性能が重要であることが分かる．

5 おわりに

本論文では，Contextual visual grounding のためのデータセットとして，r-FG-BB データセットを作成し，その詳細について述べた．手順書のフローグラフ表現の各ノードに対して BB をアノテーションすることによって，手順書の記述内容と物体の中間状態に対してグラウンディングをすることができるようになった．2 つの実験の結果，作成した r-FG-BB データセットが Contextual visual grounding に有効であることが分かった．

参考文献

[1] Balntas et al. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Proc of BMVC*, pp. 1–11, 2016.

[2] Bollini et al. Interpreting and executing recipes with a cooking robot. In *Experimental Robotics*, pp. 481–495, 2013.

[3] Deng et al. Imagenet: A large-scale hierarchical image database. In *Proc of CVPR*, pp. 248–255, 2009.

[4] Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc of NAACL-HLT*, pp. 4171–4186, 2019.

[5] He et al. Deep residual learning for image recognition. In *Proc of CVPR*, pp. 770–778, 2016.

[6] Mori et al. Flow graph corpus from recipe texts. In *Proc of LREC*, pp. 2370–2377, 2014.

[7] Papadopoulos et al. Extreme clicking for efficient object annotation. In *Proc of ICCV*, pp. 4930–4939, 2017.