Multimodal Attention Branch Networkに基づく把持命令文の生成

Sentence Generation for Fetching Instruction based on Multimodal Attention Branch Network

小椋 忠志 *1 Tadashi Ogura　Aly Magassouba*1 Aly Magassouba　杉浦 孔明 *1 Komei Sugiura　平川 翼 *2 Tsubasa hirakawa　山下 隆義 *2 Takayoshi Yamashita

藤吉 弘亘 *2 Hironobu Fujiyoshi　河井 恒 *1 Hisashi Kawai

*1情報通信研究機構 National Institute of Information and Communications Technology

*2中部大学 Chubu University

Domestic service robots (DSRs) are a promising solution to the shortage of home care workers. Nonetheless, one of the main limitations of DSRs is their inability to naturally interact through language. Recently, data-driven approaches have been shown to be effective for tackling this limitation, however, they often require large-scale datasets, which is costly. Based on this background, we aim to perform automatic sentence generation for fetching instructions, e.g., "Bring me a green tea bottle on the table." In this paper, we propose a method that generates sentences from visual inputs. Unlike other approaches, the proposed method has multimodal attention branches that utilize subword-level attention and generate sentences based on subword embeddings. In the experiment, we compared the proposed method with a baseline method using four standard metrics in image captioning. Experimental results show that the proposed method outperformed the baseline in terms of these metrics.

1. はじめに

高齢者人口の増加により，日々のケアとサポートの必要性が着実に増加している．高齢者を物理的に支援できる生活支援ロボットは，在宅介護労働者の不足に対する有望な解決策である [Iocchi 15]．これにより，必要なサポート機能を提供できる標準化された生活支援ロボットの必要性が高まっている．

一方で，生活支援ロボットの主な制限の 1 つに，言語を介して自然に相互作用できない点がある．実際，ほとんどのロボットに対してユーザが様々な表現で指示することは困難である．近年の研究では，data-driven のアプローチがあいまいな指示の処理に有効であることが示されている [Anderson 18, Magassouba 18].

しかしながら，これらのアプローチは大規模なデータセットを必要とすることが多く，その構築は時間と費用を要する．その要因は，専門家による画像に対応する文章のアノテーション作業にある．したがって，生活支援ロボットにおける命令文の自動生成手法は，このコストを大幅に削減し，アノテーション作業の負担を軽減することが期待できる．

本研究は，生活支援ロボットにおける把持命令文の自動文生成に着目する．このタスクは，画像内の目標物を指定して自然な把持命令文を生成することである．自然な把持命令文の例は，" Bring me a green tea bottle on the table." である．この命令文のように，"a green tea bottle on the table" という参照表現がしばしば用いられる．参照表現は，"table" などのランドマークに関する相対的な関係性が記述されている表現である．この課題は，言語と環境との間の多対多マッピングを解決する必要があるため，特に困難である．

本稿では，入力画像から把持命令文を生成する手法を提案する．提案手法は，視覚的入力と言語的出力の両方に対応するために，Visual Attention Branch（VAB）と Linguistic Attention Branch（LAB）で構成される．さらに，文を生成するために Generation Branch（GB）が導入されている．

提案手法は，Attention Branch Network（ABN）[Fukui 19] を拡張し，複数の Attention branch を持つ．Attention branch によって出力される Attention map は，予測するラベルが与えられた画像の最も有益な部分を強調表示する．提案手法では，通常ブラックボックスとされる深層ネットワークとは異なり，VAB が提供する Attention map はモデルの視覚的な説明として機能する．同様に，LAB が提供する Attention map は，subword の関係性の説明として機能する．

提案手法は，我々がこれまでに開発した Multimodal-ABN（Multi-ABN）[Magassouba 19] と基本構造を共有している．一方，Multi-ABN との主な違いは，LAB で使用される subword-level の注意機構と，BERT [Devlin 18] に基づく subword embedding が文生成に用いられる点である．

本研究の独自性は以下である．

- LAB と GB を導入することにより，ABN を拡張する手法を提案する．
- 提案手法は，文生成に BERT に基づく subword embedding を導入する．

2. 問題設定

本研究は，生活支援ロボットにおける把持命令文のための自然言語の生成を対象とする．以降，このタスクを把持命令文生成（Fetching Instruction Generation; FIG）タスクと呼ぶ．FIG タスクにおける典型的な場面の例を図 1に示す．右図の緑枠の物体を把持対象とした命令文は，例えば "Bring me the blue glass on the same level as the teddy bear on the metal wagon." という文で表現される． この FIG タスクの例では，命令文が把持対象物体を一意に記述する必要がある．この例のように，同じタイプの物体が多数存在する可能性がある場合，曖昧さを避けるために参照表現を含む文を生成する．参照表現を使用すると，対象となる物体を周囲の環境に対して一意に特徴付けることが可能となる．図 1では，対象物体を他

連絡先: 小椋忠志，情報通信研究機構，京都府相楽郡精華町光台 3-5，tadashi.ogura@nict.go.jp

図 1: 左 : 生活支援ロボットが把持対象物体を観察する例．右 : ロボットから得られるカメラ画像の例．緑，青の線はそれぞれ Target（blue glass），Source（metal wagon）のバウンディングボックスである．把持命令文の例 : “Bring me the blue glass on the same level as the teddy bear on the metal wagon.”

のオブジェクトと区別するために，“on the same level as the teddy bear on the metal wagon” という参照表現が用いられている．

この問題を解くことは，適切な表現が対象物体自体とその周囲に依存するため，特に困難である．例えば，図 1における命令文の参照表現は，他に “glass near the bear doll” や “blue tumbler glass on the second level of the wagon” と言ったように表現することができる．したがって，言語と環境との間の多対多マッピングを扱うこととなる．

FIG タスクにおいて，本稿では以下の入出力を想定する．

- **入力:** （環境中を巡回して得た）カメラ画像
- **出力:** 与えられた Target と Source に対して最も可能性の高い生成文

ここで，本稿では Target と Source を以下のように定義する．

- **Target:** ペットボトルやコップなどのロボットが把持するべき日常的な物体
- **Source:** テーブルや棚などの Target が置かれている家具等

本研究では，入力画像の収集にシミュレーション環境 (図 1) を用いる．FIG タスクでは，多様な環境構成において文を生成することを目的としている．そのため，低コストに多様な構成を作成できるシミュレーション環境を利用する．また，再現性の面に置いてシミュレーション環境は有利である．

シミュレーション環境における標準プラットフォームロボットとして，HSR [Yamamoto 19] を使用する．シミュレーション環境には，Unity エンジンに基づいた 3 次元環境である SIGVerse [Mizuchi 17] を用いる．

データ収集フェーズでは，HSR は日常的な家具や候補物体が配置された家庭環境内を巡回する． その後，Target と Source の候補が含まれる RGB 画像を，HSR に搭載されているカメラを用いて収集する．

3. 提案手法

提案手法は，BERT に基づく subword embedding および subword-level の注意機構を導入することによって，Multi-ABN を拡張する．図 2に提案手法の概要図を示す．提案手法は，エンコーダ，デコーダ，Linguistic Attention Branch (LAB)，Visual Attention Branch (VAB)，Generation Branch (GB) によって構成される．提案手法は以下のような特徴を持つ．

- Attention branch と Perception branch によって構成される ABN とは異なり，提案手法は LSTM に基づく encoder-decoder ネットワークによって構成され，Perception branch を保有しない．

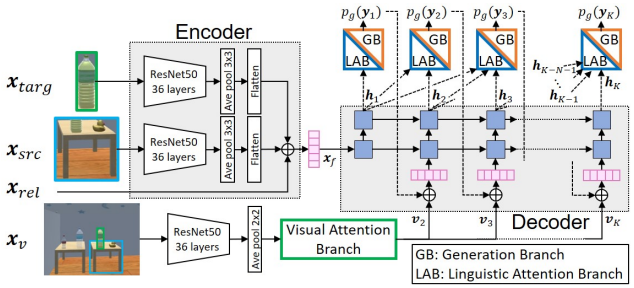


図 2: 提案手法の構成の概要

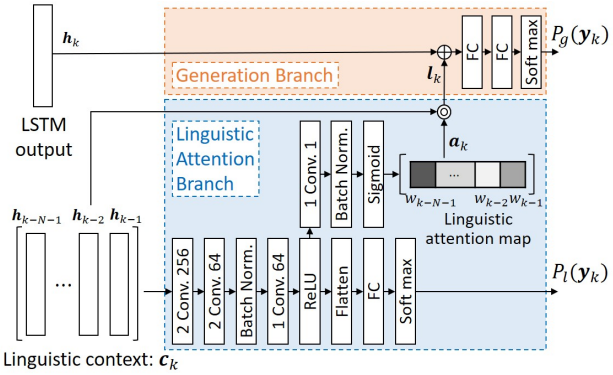


図 3: 提案手法における LAB と GB の構造

- Multi-ABN と異なり，文生成において BERT に基づくエンコーディングを用いて subword 単位の生成を行う．
- subword-level の注意機構を構成するために，新規構造の LAB および GB を保有する．潜在空間の注意機構を持つ Multi-ABN とは異なり，解釈可能な注意機構を取り扱う．

3.1 入力

図 2は提案手法のネットワーク構造を示している．シーン i について，入力セット $\mathbf{x}_i$ を次のように定義する．

$$\mathbf{x}_i = \{\mathbf{x}_v(i), \mathbf{x}_{src}(i), \mathbf{x}_{targ}(i), \mathbf{x}_{rel}(i)\} \quad (1)$$

可読性の観点から，インデックス i を省略し，$\mathbf{x}_i$ を $\mathbf{x}$ と表記する．これを用いて，入力を次のように定義する．

- $\mathbf{x}_v$: 全体の RGB 画像
- $\mathbf{x}_{targ}$: $\mathbf{x}_v$ のうちクロップされた Target 画像
- $\mathbf{x}_{src}$: $\mathbf{x}_v$ のうちクロップされた Source 画像
- $\mathbf{x}_{rel}$: $\mathbf{x}_v$，$\mathbf{x}_{targ}$，$\mathbf{x}_{src}$ の相対位置特徴

相対位置特徴 $\mathbf{x}_{rel}$ は，[Magassouba 19] に示された方法によって算出する．

3.2 ネットワーク構造

3.2.1 エンコーダ

エンコーダは，視覚情報を潜在空間特徴に変換する．潜在特徴は，後にデコーダによって文としてデコードされる． エンコーダの入力は，図 2に示すように Target $\mathbf{x}_{targ}$，Source $\mathbf{x}_{src}$ および相対位置特徴 $\mathbf{x}_{rel}$ である．Target 画像と Source 画像の両方が畳み込みニューラルネットワーク（convolutional neural network; CNN）によってエンコードされる．本稿では，ResNet-50 [He 16] を特徴抽出として利用する． エンコード

処理には，ResNet-50 のうち 36 番目のレイヤーを出力層として抽出を行う．その後，Global Average Pooling（GAP）と次元削減のためのフラット処理が行われる．$\mathbf{x}_f$ は，2 つのエンコードされた視覚的特徴と相対位置特徴 $\mathbf{x}_{rel}$ を結合することで得られる．

3.2.2 デコーダ

デコーダは，多層 LSTM を用いて各ステップ k に対して，エンコードされた特徴 $\mathbf{x}_f$ から $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_K\}$ の一連の潜在空間特徴を生成する． これらの潜在空間特徴に基づいて，LAB のおよび GB では把持命令文に対応する subword 列を予測する．LSTM のステップ k でのセルは，直前の出力 subword $\mathbf{y}_{k-1}$ の埋め込みベクトルと VAB から得られる画像特徴 $\mathbf{v}_k$ によって初期化される．本手法では Multi-ABN [Magassouba 19] における VAB と同様の VAB を用いる．その後，LSTM の隠れ状態は図 2に示すように伝播し，各ステップ k に対して出力 $\mathbf{h}_k$ が生成される．

3.2.3 Linguistic Attention Branch

図 3に LAB のネットワーク構造を示す．LAB は，LSTM の出力のうち後ろから N 個を扱う言語コンテキスト $\mathbf{c}_k = \{\mathbf{h}_{k-N-1}, \mathbf{h}_{k-N}, \cdots, \mathbf{h}_{k-1}\}$ が入力される．$\mathbf{h}_k$ はステップ k の LSTM の出力で，パラメータ N は定数である．すなわち，言語コンテキスト $\mathbf{c}_k$ の次元は $N \times d$ となる．ここで，d は LSTM の隠れ状態の次元である．したがって，LAB は $\mathbf{c}_k$ の各項に重みを付ける次元 $1 \times N$ の Attention map を作成することを目的としている．$\mathbf{c}_k$ は 3 つの 1 次元畳み込み層，バッチ正規化（BN）および ReLU によって処理される．その後，subword $\mathbf{y}_k$ は，全結合層（FC）および softmax 層を経て予測される．並行して，Attention map $\mathbf{a}_k$ は，2 番目の畳み込み層から分岐され，1×1 の畳み込み層，BN，および Sigmoid 関数を経て得られる．Attention map $\mathbf{a}_k$ の次元は $1 \times$ N であり，$\mathbf{a}_k = \{w_{k-N-1}, w_{k-N}, \cdots, w_{k-1}\}$ とする．ここで，w_k は隠れ状態 $\mathbf{h}_k$ に対応する重みである．また，$\mathbf{l}_k$ は重み付けされた言語コンテキストで，次のように定義する．

$$\mathbf{l}_k = \{\mathbf{o}_{k-N-1}, \mathbf{o}_{k-N}, \cdots, \mathbf{o}_{k-1}\} \tag{2}$$

ここで，$\mathbf{o}_k$ はそれぞれ次のように求められる．

$$\mathbf{o}_k = (1 + w_k) \odot \mathbf{h}_k \tag{3}$$

ここで，$\odot$ は Hadamard 積である．LAB における loss L_l は VAB と同様に cross-entropy によって計算される．

3.2.4 Generation Branch

図 3上部は，GB の構造を示している．GB は，把持命令文を構成する一連の subword を生成する．入力 $\mathbf{h}_k$ および $\mathbf{l}_k$ は，FC 層によって連結および処理され，そこから次の subword $p_g(\mathbf{y}_k)$ の尤度が予測される．cross-entropy loss L_g は，BG で最小化される．

3.2.5 損失関数

ネットワークのグローバル損失関数 L_{ABEN} は次のように得られる．

$$L_{\text{ABEN}} = L_v + L_l + L_g \tag{4}$$

ここで L_v，L_l，および L_g はそれぞれ VAB，LAB，GB に基づく cross-entropy loss である．

4. 実験

4.1 データセット

データセットはシミュレーション上の家庭環境で収集された．ロボットは自動的に環境内を巡回し，指定された地点におけるの画像を収集した．環境は，日常的な物体と家具によって構成されている．収集された各画像における Source と Target のバウンディングボックスはシミュレーション環境から自動的に得られる．これらの画像には，各 Target に対し把持命令文を与えるように指示された 3 人の異なるラベラーによってアノテーションされた．各画像には複数の候補 Target と Source が含まれている可能性がある．データセットとして，308 のユニーク画像と 1,099 の一意な Target 候補からなる 2,865 の画像と文のペアを収集した．

データセットは，training セット，validation セット，および test セットとして，それぞれおよそ 8 割，1 割，および 1 割に分割される．無効なサンプルを削除した後，2,295 個の training サンプル，264 個の validation サンプル，306 個の test サンプルがそれぞれ得られた．training セット，validation セット，test セットの間に重複はないため，test セットは unseen で，この実験は open-test である．

4.2 パラメータ設定

提案手法のパラメーター設定を表 2に示す．学習率が 1×10^{-4} の Adam を最適化手法に用いた． BERT に基づく埋め込みベクトルの次元は 1024 で，デコーダには 2 層の Multi-layer LSTM を用いた（図 3）．LSTM における各セルの次元数は 1024 で，言語コンテキスト $\mathbf{c}_k$ のパラメーター N を 10 に設定した．$\mathbf{c}_k$ の各要素は，デコーダの出力 $\mathbf{x}_f$ で初期化した．GB には 2 つの全結合（FC）層があり，いずれも 1024 ノードを持つ．$\mathbf{x}_{rel}$ の各次元は，その平均と標準偏差がそれぞれ 0 と 1 になるように標準化された．入力 $\mathbf{x}_{targ}$，$\mathbf{x}_{src}$，および $\mathbf{x}_v$ は，ResNet に入力される直前に $224 \times 224 \times 3$ の画像としてサイズ変更処理された．

学習は，32GB の GPU メモリ，768GB の RAM，Intel Xeon 2.10 GHz プロセッサ，Tesla V100 を搭載したマシンで実施された． 事前実験での損失収束に十分であった 100 epoch での学習を行った．

4.3 定量評価

表 1に定量的な評価結果として，画像キャプションで利用されている標準的な尺度による結果を示す．[Magassouba 19] で提案された Multi-ABN をベースライン手法として提案手法との比較を行った．Multi-ABN は，これまで Visual Semantic Embedding [Vinyals 15] および Speaker モデル [Yu 17] と比較して優れていると報告されている．自然言語の処理に適したパラフレーズ辞書を持つ METEOR および画像キャプショニングに適した CIDEr の尺度において，validation セットに対する 2 つの尺度の値が最大となる際のモデルを最適なモデルとして選択した．

定量評価実験の結果は，提案手法が 4 種類のすべての尺度において Multi-ABN を上回ることが示された．特に，CIDEr の値は，Multi-ABN と比較して 0.198 ポイントの大幅改善が見られた．これらの結果は，subword-level の文生成と注意機構の有効性を示唆している．

4.4 定性評価

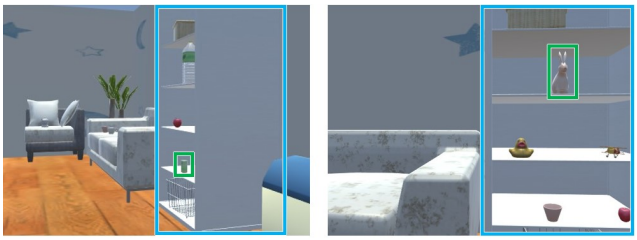
図 4に実験における定性的な結果を示す．図の上部は入力画像で，中央と下部の表は参照文（Ref1，Ref2，Ref3）および Multi-ABN と提案手法によって生成された文の結果である．

表 1: 標準尺度に基づく定量評価実験結果

Method	Evaluation metric						
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
Multi-ABN [Magassouba 19](baseline)	0.550	0.389	0.274	0.191	0.412	0.209	0.453
Proposed method	**0.586**	**0.418**	**0.302**	**0.212**	**0.471**	**0.220**	**0.651**

表 2: 提案手法のパラメータ設定

Opt. method	Adam (Learning rate $= 1.0 \times 10^{-4}$, $\beta_1 = 0.7, \beta_2 = 0.99999$)
Backbone CNN	ResNet-50
LSTM	2 layers, 1024-dimensional cell
N	10
Generation Branch	FC: 1024, 1024
Batch size	32



Ref1: “pick up the can juice”	**Ref1:** “take the rabbit please”
Ref2: “pick up a can on the lower shelf of cabinet”	**Ref2:** “let me have an rabbit item on the upper shelf of cabinet”
Ref3: “i want to bring a small can of juice”	**Ref3:** “take a white stuffed toy from the top of the shelf”
Multi-ABN: “go get a can bring on the shelf”	**Multi-ABN:** “please lift the rabbit stuffed on the upper shelf of the shelf”
Proposed method: “go get me a can from the kitchen shelf”	**Proposed method:** “give me the pink rabbit on the upper part of the shelf”

図 4: 実験における入力画像，参照文，および生成文の例

図 4の左側の結果では，Multi-ABN とは異なり，提案手法が意味的かつ構文的に成功した文生成の結果を示している．実際，Target は “go get me a can from the kitchen shelf” という文から一意に識別できる．一方，Multi-ABN によって生成された文は，Target である “can” と Source の “shelf” を出力出来ているものの，構文的に正しい文章ではない．正しい例としては，“go get a can on the shelf” というように，余分な単語 “bring” を除く文章が好ましい．

同様に図 4の右側においても，提案手法は Multi-ABN よりも正しく文章を生成できている．Target に対して “give me the pink rabbit on the upper part of the shelf” という文章を生成している．これは，参照表現を含む意味的かつ構文的に成功した文生成の結果である．一方で，Multi-ABN による生成文は，“please lift the rabbit stuffed on the upper shelf of the shelf” となっており，over-generation が発生している．

これらの結果から，LAB における subword-level の注意機構と subword 生成の貢献により，提案手法は Multi-ABN よりも自然な文生成が可能であることが示唆される．

5. おわりに

マルチモーダル言語理解のためのほとんどの data-driven アプローチには，大規模なデータセットが要求される．しかしながら，このようなデータセットの構築には時間と費用を要する．そこで本稿では，画像を指定して把持命令文を生成する手法を提案した．提案手法は，LAB と GB を導入し，ABN を拡張した．また，subword-level の注意機構および文生成を可能にした．提案手法は，生活支援ロボットのための画像と文の対応を含む大規模データセットを構築するための自動文生成への利用が期待できる．

謝辞

本研究は，JST CREST，SCOPE，NEDO の支援を受けたものである．

参考文献

[Anderson 18] Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., and Hengel, van den A.: Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments, in *IEEE CVPR*, pp. 3674–3683 (2018)

[Devlin 18] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv preprint arXiv:1810.04805* (2018)

[Fukui 19] Fukui, H., Hirakawa, T., Yamashita, T., and Fujiyoshi, H.: Attention Branch Network: Learning of Attention Mechanism for Visual Explanation, in *IEEE CVPR*, pp. 10705–10714 (2019)

[He 16] He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in *IEEE CVPR*, pp. 770–778 (2016)

[Iocchi 15] Iocchi, L., Holz, D., Solar, Ruiz-del J., Sugiura, K., and Van Der Zant, T.: RoboCup@ Home: Analysis and Results of Evolving Competitions for Domestic and Service Robots, *Artificial Intelligence*, Vol. 229, pp. 258–281 (2015)

[Magassouba 18] Magassouba, A., Sugiura, K., and Kawai, H.: A Multimodal Classifier Generative Adversarial Network for Carry and Place Tasks from Ambiguous Language Instructions, *IEEE RAL*, Vol. 3, No. 4, pp. 3113–3120 (2018)

[Magassouba 19] Magassouba, A., Sugiura, K., and Kawai, H.: Multimodal Attention Branch Network for Perspective-Free Sentence Generation, *CoRL* (2019)

[Mizuchi 17] Mizuchi, Y. and Inamura, T.: Cloud-based Multimodal Human-robot Interaction Simulator Utilizing ROS and Unity Frameworks, in *IEEE/SICE SII*, pp. 948–955 (2017)

[Vinyals 15] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D.: Show and Tell: A Neural Image Caption Generator, in *IEEE CVPR*, pp. 3156–3164 (2015)

[Yamamoto 19] Yamamoto, T., Terada, K., Ochiai, A., Saito, F., Asahara, Y., and Murase, K.: Development of Human Support Robot as the research platform of a domestic mobile manipulator, *ROBOMECH journal*, Vol. 6, No. 1, p. 4 (2019)

[Yu 17] Yu, L., Tan, H., Bansal, M., and Berg, T. L.: A Joint Speaker-Listener-Reinforcer Model for Referring Expressions, in *IEEE CVPR*, pp. 7282–7290 (2017)