Bilingual Case Relation Transformer に基づく 複数言語による物体操作指示文生成

○兼田寛大，神原元就，杉浦孔明（慶應義塾大学）

1. はじめに

ユーザと自然にコミュニケーションを行い介助や家事をサポートする生活支援ロボットは，少子高齢化や人手不足などの社会的課題の解決策として期待されている．生活支援ロボットの実用化のため様々な研究が行われており，その 1 つとして物体操作などの指示をロボットが正確に理解する技術が挙げられる．このためには実世界の画像と指示文が紐づいたマルチモーダルコーパスによる訓練が重要であるが，現時点でコーパスへの指示文付与は人手で行われているためコストが高く規模が限られるという問題がある．

そこで既存コーパスの拡充を行う際に，指示文の付与を自動化することでこの問題を解決することを考える．そのために本研究では画像から英語または日本語の指示文を生成することを目指す．本論文では，本タスクを Fetching Instruction Generation (FIG) タスクと呼ぶこととする．図 1 に本タスクの入力画像例を示す．この場合，英語では “move the teddy bear to the top right box,” 日本語では “左下の箱にある茶色の人形を右上の箱に移してください” などの指示文を付与することが望ましい．

しかし，FIG タスクにおいて正確な指示文を生成することは容易ではない．特に対象物体を正確に記述することは難しい．実際に既存研究では生成した 100 文のうち 34 の生成文において対象物体に関する誤り（動物の人形に対して “緑色の箱” と生成するなど）がみられた [1]．また，複数言語による FIG タスクを扱った研究は非常に少ない．

このような背景から，本論文では Bilingual CRT を提案する．Bilingual CRT は対象物体と目標領域の参照表現を含む指示文を，英語および日本語で生成する．また，提案手法は複数言語を 1 つのモデルで扱うことが可能である．Bilingual CRT は既存手法の CRT [1] を拡張し，Transformer Embedder および対象物体の領域画像に関する再構成損失を新たに導入する．これにより，対象物体に関する記述をより正確にする．

提案手法の主要な新規性は以下に示す通りである．

- エンコーダへの入力前の処理として，Transformer Embedder を導入する．
- 英語と日本語の 2 言語での生成を可能とする．
- 損失関数として対象物体の領域画像に対する再構成損失を新たに導入する．

2. 関連研究

マルチモーダル言語処理分野のサーベイ論文として，[2] が挙げられる．Vision and Language 分野で用いられる有名データセットとして，MSCOCO データセット [3]，Flickr30k [4]，STAIR Captions [5] などが挙げられる．これらのデータセットでは画像とそのキャプションがセットとなっており，主に画像キャプショニングタスクにて用いられる．一方で，本論文で扱う FIG タスクでは画像中の対象物体についての指示文を生成することが目的である．従って，本論文では対象物体を明示的に含む PFN-PIC データセット [6] を用いる．

図 1 FIG タスクで与えられる入力画像例．茶色の人形が対象物体 (ピンク)，右上の領域が目標領域 (水色)．

FIG タスクを扱う既存研究として，Multi-ABN [7] や ABEN [8] が挙げられる．Multi-ABN は生活支援ロボットのためのデータセット生成の自動化を問題提起し，ABEN は FIG タスクにおいて高い性能を示した．また，MTCM-AB [9] は指示文から対象物体を特定する言語理解モデルである．指示文に必要な情報が不足している場合や，同一の物体が画像内に存在する場合などにおいて，正確に対象物体を特定することができる．

提案手法の類似手法として，ORT [10] と CRT [1] が挙げられる．CRT は Transformer を用いた言語生成モデルである．CRT はエンコーダへの入力前の処理として線形変換を用いるが，提案手法では Transformer エンコーダを用いた処理に変更したという違いがある．

3. 問題設定

本論文では英語と日本語の 2 言語における FIG タスクを扱う．このタスクでは，与えられた画像に含まれる対象物体を目標領域へ移動させるための指示文を生成することを目指す．これらの指示文は対象物体および目標領域を特定するための位置参照表現を含むものとする．本タスクでは，与えられた画像に対して対象物体と目標領域を正確に特定することができる明確な指示文を生成することが望ましい．

FIG タスクにおいて，入力は対象物体，目標領域，コンテキスト情報を含む画像およびそれぞれの座標情報である．また，出力は対象物体と目標領域の参照表現を含む指示文である． 本タスクの評価として，[1] と同様の尺度を用いる．日本語の指示文の評価には SPICE および MOS を除いた尺度を用いる．ここで，本論文で使用する用語を以下のように定義する．

- **対象物体**：FIG タスクにおいて移動させる日常用品． （ペットボトル，缶，コップなど）

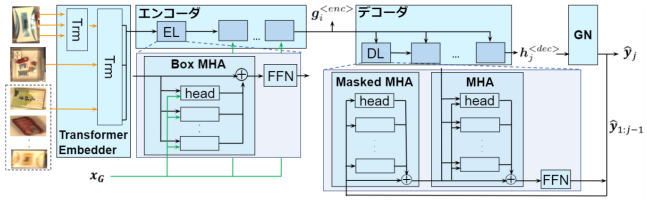

図 2 Bilingual CRT の構造．Trm，EL，FFN，DL，MHA，head，GN はそれぞれ Transformer レイヤ，エンコーダレイヤ，順伝搬型ネットワーク層，デコーダレイヤ，multi-head attention 層，attention head，ジェネレータを示す．

- **目標領域**：FIG タスクにおいて移動先となる 4 つの領域（右上，右下，左下，左上）のうちの 1 つ．
- **コンテキスト情報**：Up-Down Attention [11] により得られた画像中の領域群

また，本論文では対象物体の画像中における座標は与えられることを前提とする．本タスクを扱うために，画像および指示文に対して事前処理を行った．画像には MSCOCO データセット [3] により事前訓練された Up-Down Attention モデルを用い，物体検出を行った．また日本語の指示文は形態素解析エンジンである MeCab を用いて分かち書きに変換した．これは辞書の作成や評価において英語と同様の手法で実行するためである．さらに文末の句点，感嘆符および疑問符の削除を行った．これは日本語の指示文において句点，感嘆符および疑問符が全角および半角で混同されていたことによりそのままでは扱うことが難しかったためである．例として “動かして!” と “動かして！” はどちらも “動かして！” という文節を指すが，半角の記号が用いられているために前者はこのまま処理することができなかった．

4. 手法

4.1 モデル概要および多言語化

モデル入力は $\boldsymbol{x} = (\boldsymbol{X}^{<targ>}, \boldsymbol{x}^{<dest>}, \boldsymbol{X}^{<cont>})$ である．ここで，$\boldsymbol{X}^{<cont>} = (\boldsymbol{x}^{<1>}, \boldsymbol{x}^{<2>}, ..., \boldsymbol{x}^{<N>})$，$\boldsymbol{x}^{<i>} = (\boldsymbol{x}_V^{<i>}, \boldsymbol{x}_G^{<i>})$ である．さらに，$\boldsymbol{x}_V^{<i>}$ は対象物体について ResNet-50 [12] に入力して conv2_x 層，conv3_x 層，conv4_x 層から得た特徴量，目標領域について ResNet-50 に入力して conv5_x 層から得た特徴量，コンテキスト情報の各画像領域について ResNet-101 に入力して conv5_x 層から得た特徴量を表す．また，N はコンテキスト情報に含まれる物体の数を表す．$\boldsymbol{x}_G^{<i>}$ は各領域についての幾何的特徴量である [1]．

図 2 に Bilingual CRT の構造を示す．$\boldsymbol{x}_G$ および緑の矢印は幾何的特徴量を示し，オレンジの矢印は画像特徴量を示す．出力はトークン $\boldsymbol{y}_j$ である．ここで，j は予測するトークンのインデックスを示す．また，$\boldsymbol{g}_i^{<enc>}$ は対象物体の領域画像を再構成することを目的としたベクトルを表す．本モデルにおける注意機構の効果として，空間参照表現を獲得することができ，品質の高い指示文を生成することができる．提案手法は主に 3 つのモジュールからなり，それぞれを Transformer Embedder，エンコーダ，デコーダと呼ぶ．

Bilingual CRT は複数の言語で書かれた指示文を混合して学習することで，1 つのモデルで英語および日本語の指示文を生成する．そのために，学習時には指示文の文頭に言語に応じて<en>または<ja>トークンを挿入する．生成時にはいずれかのトークンを文頭に与えることで指示文が得られる．すなわち，モデルの切り替えは不要である．

4.2 Transformer Embedder

Transformer Embedder では $\boldsymbol{X}^{<targ>}$，$\boldsymbol{x}^{<dest>}$，$\boldsymbol{X}^{<cont>}$ を入力とする．ここで，$\boldsymbol{X}^{<targ>}$ は対象物体の領域画像を ResNet-50 に入力し，conv2_x 層，conv3_x 層，conv4_x 層から得られた出力を結合したベクトルである．

$\boldsymbol{X}^{<targ>}$ は Transformer レイヤ [13] に入力される．まず Query $\boldsymbol{Q}_T$，Key $\boldsymbol{K}_T$，Value $\boldsymbol{V}_T$ がパラメータ行列 $\boldsymbol{W}_q$，$\boldsymbol{W}_k$，$\boldsymbol{W}_v$ を用いて以下の式 (1) で計算される．

$$\boldsymbol{Q}_T = \boldsymbol{W}_q\boldsymbol{X},\ \boldsymbol{K}_T = \boldsymbol{W}_k\boldsymbol{X},\ \boldsymbol{V}_T = \boldsymbol{W}_v\boldsymbol{X} \quad (1)$$

ただし $\boldsymbol{X} = \boldsymbol{X}^{<targ>}$ である．次に，multi-head attention 層においては注意が計算される．multi-head attention 層は内部に独立した注意機構を N_T 個並列に持ち，$\boldsymbol{Q}_T$，$\boldsymbol{K}_T$，$\boldsymbol{V}_T$ をそれぞれ N_T 個に分割したものがそれぞれへの入力となる．この注意機構のことを attention head と呼ぶ．それぞれの attention head において，$\boldsymbol{\omega}_A = \frac{\boldsymbol{Q}_T\boldsymbol{K}_T^\top}{\sqrt{d_k}}$ が計算される．ここで，d_k は Key の隠れ層の次元数である．

その後，各ヘッドの出力 $\boldsymbol{h}_{sa} = \boldsymbol{V}_T\mathrm{softmax}(\boldsymbol{\omega}_A)$ が計算される．N_T 個の attention head の出力はパラメータ行列 $\boldsymbol{W}_M$ を用いて以下の式 2 のように計算される．

$$\boldsymbol{h}_{mh} = \{\boldsymbol{h}_{sa}^{<1>}, \boldsymbol{h}_{sa}^{<2>}, ..., \boldsymbol{h}_{sa}^{<N_T>}\}\boldsymbol{W}_M \quad (2)$$

最後に，$\boldsymbol{h}_{mh}$ を FFN 層と畳み込み層に入力し，$\boldsymbol{h}_{out}^{<trm>}$ を得る．その後，$\boldsymbol{h}_{out}^{<trm>}$，目標領域，コンテキスト情報を同様に Transformer レイヤに入力し，Transformer Embedder の出力 $\boldsymbol{h}_v \in \mathbb{R}^{N_A \times 512}$ を得る．ここで，N_A は入力された領域の数を示す．

4.3 エンコーダ

エンコーダは 3 層のエンコーダレイヤによって構成される．エンコーダレイヤは box multi-head attention 層と FFN 層で構成される．入力は Transformer Embedder からの出力 $\boldsymbol{h}_v = \boldsymbol{h}_{in}^{<enc>}$ および $\boldsymbol{h}_G$ をとる．ここで，$\boldsymbol{h}_G$ は $\boldsymbol{x}_G^{<i>}$ を結合したベクトルである．

Box multi-head attention 層では入力として $\boldsymbol{h}_{in}^{<el>}$ をとる．ここで，$\boldsymbol{h}_{in}^{<el>}$ は最初の層では $\boldsymbol{h}_{in}^{<enc>}$，それ以外の層では前のエンコーダレイヤの出力である．領域 m, n に対して相対位置ベクトル $\boldsymbol{\Lambda}(m,n) = \{\lambda(\delta w, w_m), \lambda(\delta h, h_m), \lambda(w_n, w_m), \lambda(h_n, h_m)\}$ が計算される．ただし，$\lambda(x,y)$，w_i，h_i，δw，δh はそれぞれ $\log(x/y)$，領域 i の幅，領域 i の高さ，$|r_{xmin}^m - r_{xmin}^n|$，$|r_{ymin}^m - r_{ymin}^n|$ を示す．その後，$\boldsymbol{\omega}_G^{mn} = \mathrm{ReLU}(f_{em}(\boldsymbol{\Lambda}(m,n)\boldsymbol{W}_G))$ が Transformer で用いられる positional encoding f_{em} およびパラメータ行列 $\boldsymbol{W}_G$ を用いて以下の式で計算される．次に，式 (1) と同様に $\boldsymbol{Q}_E$，$\boldsymbol{K}_E$，$\boldsymbol{V}_E$ が計算される．その後，式 (2) と同様の処理を行い，出力 $\boldsymbol{h}_{out}^{<el>}$ を得る．

4.4 デコーダ

デコーダは 3 層のデコーダレイヤによって構成される．デコーダレイヤは masked multi-head attention 層，multi-head attention 層，および FFN 層で構成される．

デコーダの入力は $\boldsymbol{h}_{in}^{<dec>}$ である．ここで，$\boldsymbol{h}_{in}^{<dec>} = \boldsymbol{h}_{out}^{<enc>}$ である．デコーダでは，$\hat{\boldsymbol{y}}_{1:j-1}$ を用いて自己回帰的にトークン予測を行う．まず，masked multi-head attention 層では式 (1) および式 (2) と同様に自己注意

$\boldsymbol{h}_{emb}$ を計算する．ただし，訓練時は j トークン以降をマスクすることで，過去の情報のみを参照するようにする．次に multi-head attention 層では，式 (1) と同様に $\boldsymbol{Q}_D$，$\boldsymbol{K}_D$，$\boldsymbol{V}_D$ が計算される．その後，式 (2) と同様の処理を行い，出力 $\boldsymbol{h}_j^{<dec>}$ を得る．

4.5 ジェネレータおよび損失関数

ジェネレータの入力は $\boldsymbol{h}_g = \boldsymbol{h}_j^{<dec>}$ であり，全結合層およびソフトマックス関数を用いて j 番目のトークンについての確率の予測値 $p(\hat{\boldsymbol{y}}_j)$ を計算する．

損失関数は交差エントロピー誤差関数と再構成損失の加重和を用いており，以下の式で示される．

$$L = -\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J}\log(p(w_{ij})) + \alpha\sum_{i=1}^{I}\|\boldsymbol{g}_i^{<inp>} - \boldsymbol{g}_i^{<enc>}\|$$

第 1 項は交差エントロピー誤差，第 2 項は再構成損失を表す．ここで，I，J，$p(w_{ij})$，α，$\boldsymbol{g}_i^{<inp>} \in \mathbb{R}^{1\times1200}$，$\boldsymbol{g}_i^{<enc>} \in \mathbb{R}^{1\times1200}$ はそれぞれサンプル数，各文の長さ，サンプル中のトークン w_{ij} の予測確率，重み，入力の対象物体の領域画像を 1 次元に変換整形したベクトル，対象物体の領域画像に関するエンコーダの出力を線形変換したベクトルを表す．ここで，再構成損失は入力画像を潜在空間にマッピングした後に再構成を行い，元の入力画像との ℓ_1 ノルムを計算した値である．そのため，再構成損失を小さくするように訓練させることで，入力画像の特徴を保持するように学習が行われる．対象物体の領域画像に関する再構成損失を導入することで，対象物体の色や形状の記述に関する誤りを減らすことが期待できる．

5. 実験設定

データセットとして PFN-PIC [6] を用いた．データセットの事前処理として，画像ごとに対象物体の座標，目標領域の情報 (右上，右下，左上，左下のいずれか)，および指示文がデータセットにおいてまとめられていたため，サンプルごとにセットとなるよう整形を行った．本研究では，訓練集合，検証集合，テスト集合をそれぞれ 81087，8774，898 となるように分割した．訓練集合と検証集合には PFN-PIC データセットの train を分割して用い，訓練集合はパラメータの学習に，検証集合はハイパーパラメータの検証に選択した．テスト集合には validation を用い，性能の評価に使用した．学習可能パラメータは 3400 万である．

提案手法のハイパーパラメータ設定は以下に示すとおりである。バッチサイズは 15，エポック数は 15，学習率は 5.0×10^{-4}，エンコーダを構成するエンコーダレイヤおよびデコーダを構成するデコーダレイヤの数は 3，Transformer Embedder を構成するエンコーダレイヤの数は 2，multi-head attention 層中の attention head の数は 4，α は 5.0×10^{-3} である．

本提案手法の学習はメモリ 11GB 搭載 GeForce RTX 2080 Ti ×1， 64GBRAM，そして Intel Corei9 9900K により行われる．Up-Down Attention [11] による特徴抽出のみメモリ 24GB 搭載 TITAN RTX ×1，256GBRAM，そして Intel Corei9 9820X という構成により行っている．学習は 60 分程度で完了し，1 サンプルあたりの推論にかかる時間は 44ms 程度であった．未知データに対して汎化性能の高いハイパーパラメータを選択するため，200 バッチ分訓練を行うごとに検証集合を用いてその時点での性能を評価した．最終的に 15 エポックまでで英語および日本語の CIDEr-D スコアの平均値が最も高いハイパーパラメータを選択し，テスト集合に対して評価を行った．

表 1　各手法による英語の生成文の品質評価

手法	BLEU4	METEOR	CIDEr-D	SPICE
ORT [10]	7.3 ± 1.4	17.4 ± 0.9	29.3 ± 2.3	26.7 ± 1.3
CRT [1]	14.9 ± 1.1	23.1 ± 0.7	96.6 ± 12.0	44.0 ± 2.3
提案手法	**16.4 ± 0.6**	**24.6 ± 0.3**	**115.8 ± 3.4**	**48.2 ± 0.1**

表 2　各手法による日本語の生成文の品質評価

手法	BLEU4	METEOR	CIDEr-D
CRT [1]	25.4 ± 1.1	29.4 ± 0.3	94.2 ± 4.7
提案手法	**25.8 ± 0.6**	**29.5 ± 0.6**	**116.5 ± 3.9**

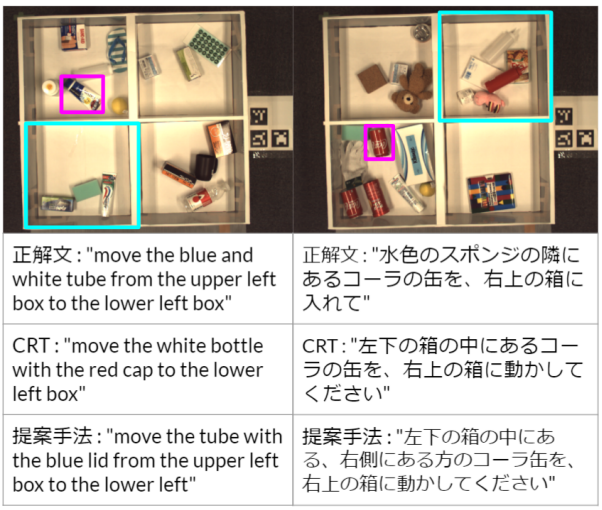

図 3　適切な指示文を付与できた例．

6. 実験結果

6.1 定量的結果

表 1 および表 2 に提案手法およびベースラインの結果を示す．表 1 は英語の指示文の結果であり，表 2 は日本語の指示文の結果である．なお実験は各手法につき 5 回学習およびテストを行い，表にはその平均値および標準偏差を示す．また実験の際，各論文の設定値を参考にして ORT [10] については 20 エポック，また CRT [1] については 10 エポック訓練した．

ベースラインは ORT と CRT とした．今回用いた評価尺度は BLEU4，METEOR，CIDEr-D，SPICE, MOS である．また 5 つの自動評価尺度における主要尺度は英語の指示文は SPICE，日本語の指示文は CIDEr-D とする．上記の評価尺度のうち，CIDEr-D および SPICE は画像キャプショニングタスク用の尺度であり，FIG タスクと性質が似ていることから使用した．BLEU4，METEOR に関しては今回のタスクに特化した尺度ではないものの，自然言語の分野において一般的な評価尺度であるため使用した．

表 1 より，主要尺度である SPICE に関しては ORT が 26.7，CRT が 44.0 であったのに対し，提案手法は 48.2 であった．CIDEr-D に関しては ORT が 29.3，CRT が 96.6 であったのに対し，提案手法は 115.8 であった．従って，既存手法と比較して提案手法が優れるという結果を得た．また表 2 より，日本語においても既存手法と比較して提案手法が優れるという結果を得た．

6.2 定性的結果

図 3 に適切な指示文を生成出来た例を示す．それぞれの図において，ピンクで囲まれた領域が対象物体を

表 3 被験者実験の結果．各手法における MOS の平均値および 95%信頼区間の上限および下限を示す．

手法	MOS
正解文 (Upper bound)	4.42 ± 0.11
ORT [10]	1.35 ± 0.08
CRT [1]	3.29 ± 0.19
提案手法	4.01 ± 0.16

表 4　各条件における英語の生成文の品質評価

	BLEU4	METEOR	CIDEr-D	SPICE
TE なし	15.9 ± 1.5	24.2 ± 0.8	110.7 ± 10.2	45.3 ± 3.8
RL なし	$\mathbf{16.5 \pm 1.3}$	24.5 ± 0.5	111.8 ± 6.9	46.8 ± 1.9
提案手法	16.4 ± 0.6	$\mathbf{24.6 \pm 0.3}$	$\mathbf{115.8 \pm 3.4}$	$\mathbf{48.2 \pm 0.1}$

示し，水色で囲まれた領域が目標領域を示す．

図 3 の左に示す画像では対象物体が左上の領域にある青と白のチューブ，目標領域が左下の領域である．正解文ではそれぞれの参照表現として対象物体に “blue and white tube, ” 目標領域に “lower left box” が当てられている．既存手法は対象物体について “white bottle with the red cap” と不正確な表現になっているが，提案手法はこのサンプルに対して対象物体の参照表現を “tube with the blue lid, ” 目標領域の参照表現を “lower left” としており，適切な指示文を出力できた．

図 3 の右に示す画像では対象物体が左下の領域にあるコーラの缶，目標領域が左上の領域である．ここで，左下の領域には 2 つのコーラの缶があるため，指示文は正確に対象物体を特定できることが重要である．正解文ではそれぞれの参照表現として対象物体に “水色のスポンジの隣にあるコーラの缶, ” 目標領域に “右上の箱” が当てられている．既存手法では対象物体の参照表現を “左下の箱の中にあるコーラの缶” としており，どちらのコーラを指しているのか不明瞭な表現となってしまっている．一方で，提案手法はこのサンプルに対して対象物体の参照表現を “右側にある方のコーラ缶, ” 目標領域の参照表現を “右上の箱” としており，“右側にある方” ともう 1 つのコーラを参照して対象物体を特定することで，品質の高い指示文を生成できた．

6.3　被験者実験

生成文の品質について，被験者による評価を行った．被験者は 20 代の男性日本語話者 5 名とした．実験をするにあたり，テスト集合の正解文およびテスト集合についての生成文から無作為にそれぞれ 50 文抽出したうえで，被験者への提示順も無作為とした．被験者には指示文と対応する画像を提示し，指示文の明瞭さによって以下に示すように 5 段階で評価させた．被験者には各自の作業速度で評価作業を行わせた．

1 : とても悪い, 2 : 悪い, 3 : 普通, 4 : 良い, 5 : とても良い

表 3 に，被験者実験によって得られた各手法における MOS の平均値および 95%信頼区間を示す．表より，ORT の MOS は 1.35，CRT の MOS は 3.29 であったのに対し，提案手法の MOS は 4.01 であった．また正解文の MOS は 4.42 であるため，提案手法によって生成された指示文は正解文に近い品質を達成したといえる．

6.4　Ablation Study

新規性である Transformer Embedder と再構成損失が性能にどの程度影響を与えるか検証するため，英語の指示文について Ablation study を行った．検証は各手法につき 5 回学習およびテストを行い，表 4 にはその平均値および標準偏差を示した．TE は Transformer Embedder，RL は再構成損失を表す．まず Transformer Embedder の有無を比較すると，主要尺度である SPICE に関して Transformer Embedder なしのモデルは 45.3 であった一方で，Transformer Embedder ありのモデルは 48.2 であった．従って，Transformer Embedder は性能向上に寄与していることが分かった．また、再構成損失の有無を比較すると，SPICE に関して再構成損失なしのモデルは 46.8 であった一方で，再構成損失ありのモデルは 48.2 であった．従って，再構成損失も同様に性能向上に寄与していることが分かった．

7.　結論

本論文では，対象物体および目標領域を含む画像から物体操作指示文を生成するタスクを扱った．提案手法の主要な貢献は以下に示す通りである．

- 英語および日本語の物体操作指示文を 1 モデルで生成する Bilingual CRT を提案した．
- Bilingual CRT では Transformer Embedder および再構成損失を新たに導入し，各評価尺度においてベースライン手法を上回る品質を達成した．
- 被験者実験において，人間の付与した指示文に近い品質の生成文を得られることを確認した．

謝辞

本研究の一部は，JSPS 科研費 20H04269，JST CREST，JST ムーンショット型研究開発事業 JPMJMS2011，NEDO の助成を受けて実施されたものである．

参　考　文　献

[1] K. Motonari and S. Komei, “Case Relation Transformer: A Crossmodal Language Generation Model for Fetching Instructions,” IROS, 2021.
[2] A. Mogadala, et al., “Trends in integration of vision and language research: A survey of tasks, datasets, and methods,” arXiv preprint arXiv:1907.09358, 2019.
[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, “Microsoft coco: Common objects in context,” ECCV, pp.740–755, 2014.
[4] P. Young, et al., “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” ACL, vol.2, pp.67–78, 2014.
[5] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, “Stair captions: Constructing a large-scale japanese image caption dataset,” arXiv preprint arXiv:1705.00823, 2017.
[6] J. Hatori, Y. Kikuchi, S. Kobayashi, et al., “Interactively picking real-world objects with unconstrained spoken language instructions,” ICRA, pp.3774–3781, 2018.
[7] A. Magassouba, K. Sugiura, and H. Kawai, “Multimodal attention branch network for perspective-free sentence generation,” CoRL, pp.76–85, 2020.
[8] T. Ogura, et al., “Alleviating the burden of labeling: Sentence generation by attention branch encoder–decoder network,” IEEE RA-L, vol.5, no.4, pp.5945–5952, 2020.
[9] A. Magassouba, K. Sugiura, and H. Kawai, “A multimodal target-source classifier with attention branches to understand ambiguous instructions for fetching daily objects,” IEEE RA-L, vol.5, no.2, pp.532–539, 2020.
[10] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, “Image captioning: Transforming objects into words,” arXiv preprint arXiv:1906.05963, 2019.
[11] P. Anderson, X. He, C. Buehler, et al., “Bottom-up and top-down attention for image captioning and visual question answering,” CVPR, pp.6077–6086, 2018.
[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” CVPR, pp.770–778, 2016.
[13] A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention is all you need,” arXiv preprint arXiv:1706.03762, 2017.