Case Relation Transformer に基づく 対象物体及び目標領域の参照表現を含む物体操作指示文生成

Generating Object Manipulation Instructions Including Referring Expressions of Target Objects and Destinations Based on Case Relation Transformer

神原 元就 *1　　杉浦 孔明 *1
Motonari Kambara　　Komei Sugiura

*1慶應義塾大学
Keio University

The purpose of this paper is to extend the dataset based on a cross-modal generative language generation model. We propose a Case Relation Transformer (CRT) that generates a fetching instruction sentence from an image, such as “Move the blue flip-flop to the lower left box.” Unlike existing methods, CRT uses Transformer to capture the visual and geometric features of objects in an image. The Case Relation Block allows the CRT to process the object. We conducted comparative experiments and human evaluations. Experimental results showed that CRT outperformed the baseline methods.

1. はじめに

家事を支援するためにユーザーとの自然なコミュニケーションが可能な家庭用サービスロボット（DSR）は，高齢者や障害者の在宅介護労働者の不足に対する有望な解決策である．このようなコミュニケーションスキルを向上させるためのロボット工学の研究は数多くあるが，コーパスが十分に大きくないため，それらのほとんどは近年発展を遂げたディープニューラルネットワークを活かし切れていない．これは主に，テキストデータが画像等の実世界のデータに接地されているクロスモーダルコーパスを構築するのに非常にコストがかかるためである．特に，画像へのテキスト付与はアノテータによって行われることが多く，時間と費用がかかる．

このような背景を踏まえ，本研究では画像データを用いてテキストデータを拡張することに焦点を当てる．この際，一部の画像のみアノテータによってテキストデータが付与されていると仮定する．本研究の目的は “Move the blue and white tissue box to the top right bin.” 等，与えられた画像に対し指示文を生成することによりデータを拡張することである．以降、この目標タスクを把持命令文生成（FIG）タスクと呼ぶ．このようなデータの拡張は，クロスモーダル言語理解モデルの精度の向上に貢献すると考えられる．実際，Zhao らはクロスモーダル言語生成モデルによる拡張データにより，クロスモーダル言語理解モデルの性能が向上しうると報告している [Zhao 21].

FIG タスクの難しさは，文のあいまいさが目標物体だけでなく周囲の物体に関する表現にも依存する点にある．実際，FIG タスクにおいて生成された文の品質は，単純な画像キャプションタスクの場合よりもはるかに劣っている [Ogura 20]．また，Section 4. に示すように，参照文と既存手法による生成文の間には，品質に大きな差がある．

本論文では，対象物体と目標領域の空間参照表現を含む物体移動指示文生成モデル Case Relation Transformer（CRT）を提案する． CRT は，Case Relation Block（CRB），Transformer エンコーダ，及び Transformer デコーダの 3 つの主要モジュールで構成される．既存手法 [Ogura 20] と異なり，CRT は Transformer を利用する．これにより，FIG タスクにおける物体の画像特徴量と幾何的特徴量を統合する．さらに，Object Relation Transformer（ORT [Herdade 19]）とは異なり，CRT は対象物体と目標領域の両方を処理できる．

図 1:　FIG タスクにおける典型的なシーン画像例

本研究の独自性は以下である．

- Transformer モデルに基づく FIG タスクのためのクロスモーダル言語生成モデル，CRT を提案する．
- CRB を導入し ORT を拡張することにより，目標物体，対象領域，及びコンテクスト情報を扱う．

2. 問題設定

2.1 タスク説明

本研究は，DSR のためのクロスモーダル言語指示文生成，すなわち FIG タスクを対象とする．指示文には空間参照表現が含まれているとする．空間参照表現は，注目している物体とその周囲の物体との位置関係によって記述された参照表現であり，これによって物体を特定する．具体例としては，“books on the desk” や “cans next to the plastic bottles” 等がある．このタスクでは，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できる明確な指示文を生成することが望ましい動作である．例えば，図 1 のような画像が与えられたとき，“Move the blue and white tissue box to the top right bin.” のような指示文が生成されることが望ましい．

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:**
  - 対象物体，目標領域，及びコンテクスト情報の各領域を含む画像
  - 対象物体及び目標領域の領域の座標値

連絡先: 神原元就，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，motonari.k714@keio.jp

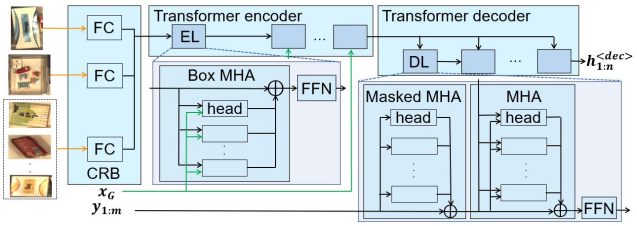

図 2: 提案手法の構成図

- **出力:**
  - 対象物体を目標領域へ移動する指示文

ここで，本稿では対象物体，目標領域，及びコンテクスト情報を以下のように定義する．

- **対象物体:** ペットボトルや缶などのロボットが把持するべき日常的な物体
- **目標領域:** 対象物体を移動する方向であり，右上，右下，左上，左下の 4 つの方向のうちの 1 つ
- **コンテクスト情報:** Up-Down Attention [Anderson 18] 等の物体検出器で検出された領域群

3. 提案手法

提案手法は，物体の相対位置から空間参照表現をモデル化できる Object Relation Transformer（ORT [Herdade 19]）に基づいている．特徴抽出は [Herdade 19] に従って Up-Down Attention [Anderson 18] を使用する．

3.1 入力

ネットワーク入力 $\boldsymbol{x}$ は以下で定義される．

$$\boldsymbol{x} = (\boldsymbol{x}^{<targ>},\ \boldsymbol{x}^{<dest>},\ \boldsymbol{X}^{<cont>})$$

$$\boldsymbol{X}^{<cont>} = (\boldsymbol{x}^{<1>}\ \boldsymbol{x}^{<2>}, ..., \boldsymbol{x}^{<N>})$$

$$\boldsymbol{x}^{<i>} = (\boldsymbol{x}_V^{<i>}, \boldsymbol{x}_G^{<i>}) \tag{1}$$

ここで，$\boldsymbol{x}^{<targ>}$，$\boldsymbol{x}^{<dest>}$，$\boldsymbol{X}^{<cont>}$, 及び N はそれぞれ対象物体，目標領域，コンテクスト情報についての特徴量，及びコンテクスト情報に含まれる領域数を表す．$\boldsymbol{x}_V^{<i>}$ 及び $\boldsymbol{x}_G^{<i>}$ は領域 i についての画像特徴量及び幾何的特徴量を表す．

対象物体及び目標領域の事前処理には ResNet-50 を使用する．具体的には，conv5_x 層からの出力を，$\boldsymbol{x}_V^{<targ>}$ 及び $\boldsymbol{x}_V^{<dest>}$ として使用する．コンテクスト情報については [Anderson 18] と同様の手順で物体検出及び特徴抽出を行う．物体検出には Faster R-CNN [Ren 16] を使用し，ResNet-101 の conv5_x 層からの出力を $\boldsymbol{X}_V^{<cont>}$ として使用する．対象物体，目標領域，及びコンテクスト情報に含まれる領域ごとに，1×2048 次元の画像特徴量を抽出した．

領域 i の幾何的特徴量，$\boldsymbol{x}_G^{<i>}$ は $\boldsymbol{x}_G^{<i>} = [r_{xmin}^{<i>}/W, r_{ymin}^{<i>}/H, r_{xmax}^{<i>}/W, r_{ymax}^{<i>}/H]$ で定義される．ここで $(r_{xmin}^{<i>}, r_{ymin}^{<i>}, r_{xmax}^{<i>}, r_{ymax}^{<i>})$ はそれぞれ，領域 i の x 座標と y 座標の最小値及び最大値を表す．また，W 及び H はシーン画像の幅及び高さを表す．

3.2 ネットワーク構造

図 2 に CRT のネットワーク構造を示す．図中，$\boldsymbol{x}_G$ 及び $\boldsymbol{y}_{1:m}$ は，それぞれ幾何的特徴量と入力文を表す．オレンジ及び緑の矢印はそれぞれ画像特徴量及び幾何的特徴量を表し，$\boldsymbol{h}_{1:n}^{<dec>}$ はトークン列を表す．また，FC，EL，FFN，DL，MHA，及び “head” は，それぞれ全結合層，エンコーダレイヤ，フィードフォワードネットワーク層，デコーダレイヤ，Multi-head attention 層，及びアテンションヘッド [Vaswani 17] を表す．

CRT は CRB，Transformer エンコーダ，及び Transformer デコーダの三つの主要モジュールで構成される． 各モジュールの詳細を以下に示す．

3.2.1 Case Relation Block

CRB ではまず入力 $\boldsymbol{x}$ を全結合 (FC) 層 $f_{FC}(\cdot)$ により線形変換する．出力 $\boldsymbol{h}_V \in \mathbb{R}^{N \times d_{model}}$ は以下のようにして得られる．

$$\boldsymbol{h}_V = \left\{ f_{FC}(\boldsymbol{x}_V^{<targ>}), f_{FC}(\boldsymbol{x}_V^{<dest>}), f_{FC}(\boldsymbol{x}_V^{<cont>}) \right\}$$

ここで d_{model} は次元数を表す．これらをこの順序で結合し，出力を条件付けする．$\boldsymbol{x}_G^{<i>}$ $(i = 1, 2, \cdots, N)$ についてもこの順序で結合することにより，$\boldsymbol{h}_G \in \mathbb{R}^{N \times 4}$ を得る．$\boldsymbol{h}_V$ 及び $\boldsymbol{h}_G$ は Transformer エンコーダの入力である．

3.2.2 Transformer エンコーダ

Transformer エンコーダは，6 層のエンコーダレイヤで構成され，各エンコーダレイヤは，Box multi-head attention 層及び FFN 層で構成される．各エンコーダレイヤへの入力は $\boldsymbol{h}_{in}^{<el>}$ 及び $\boldsymbol{h}_G$ である．ここで，$\boldsymbol{h}_{in}^{<el>}$ は一つ前のエンコーダレイヤの出力であり，特に最初のレイヤへの入力は $\boldsymbol{h}_V$ である．

最初に Box multi-head attention 層において，領域 m 及び n の変位ベクトル $\boldsymbol{\Lambda}(m, n)$ がそれらの情報から以下のように計算される．

$$\boldsymbol{\Lambda}(m, n) = \{\lambda(\delta w, w_m), \lambda(\delta h, h_m), \lambda(w_n, w_m), \lambda(h_n, h_m)\}$$

$$\lambda(x, y) = \log(x/y)$$

ここで，w_i 及び h_i は領域 i の幅及び高さを表す．また，δw 及び δh は $|r_{xmin}^m - r_{xmin}^n|$ 及び $|r_{ymin}^m - r_{ymin}^n|$ を表す．幾何的注意の重み $\boldsymbol{\omega}_G^{mn}$ は $\boldsymbol{\omega}_G^{mn} = \mathrm{ReLU}(f_{em}(\boldsymbol{\Lambda}(m, n)\boldsymbol{W}_G)$ のように計算される．ここで，$f_{em}(\cdot)$ は Transformer で用いられる位置エンコーディングを表す．クエリ $\boldsymbol{Q}_E$，キー $\boldsymbol{K}_E$，及びバリュー $\boldsymbol{V}_E$ はそれぞれ $\boldsymbol{Q}_E = \boldsymbol{W}_{qe}\boldsymbol{h}_{in}^{<el>}$，$\boldsymbol{K}_E = \boldsymbol{W}_{ke}\boldsymbol{h}_{in}^{<el>}$，及び $\boldsymbol{V}_E = \boldsymbol{W}_{ve}\boldsymbol{h}_{in}^{<el>}$ のように計算される．ここで，$\boldsymbol{W}_{qe}$，$\boldsymbol{W}_{ke}$，及び $\boldsymbol{W}_{ve}$ はそれぞれ，$\boldsymbol{Q}_E$，$\boldsymbol{K}_E$，及び $\boldsymbol{V}_E$ の重みを表す．

次に，$\boldsymbol{\omega}_G^{mn}$，$\boldsymbol{Q}_E$，$\boldsymbol{K}_E$，及び $\boldsymbol{V}_E$ は N_E 個のアテンションヘッドに入力される．この時，$\boldsymbol{Q}_E$，$\boldsymbol{K}_E$，及び $\boldsymbol{V}_E$ は N_E 等分される．このアテンションヘッドは Box multi-head attention 層内に並列実装されている．それぞれのアテンションヘッド内において，画像特徴量に基づく注意の重み $\boldsymbol{\omega}_A^{mn}$ は以下のように計算される．

$$\boldsymbol{\omega}_A^{mn} = \frac{\boldsymbol{Q}_E \boldsymbol{K}_E^\top}{\sqrt{d_k}} \tag{2}$$

ここで，$d_k = d_{model}/N_E$ は $\boldsymbol{K}_E$ の次元数を表す．注意の重み $\boldsymbol{\omega}^{mn} \in \mathbb{R}^{N \times N}$ は $\boldsymbol{\omega}_A^{mn}$ 及び $\boldsymbol{\omega}_G^{mn}$ からソフトマックス関数によって以下のように計算される．

$$\boldsymbol{\omega}^{mn} = \frac{\boldsymbol{\omega}_G^{mn} \exp \boldsymbol{\omega}_A^{mn}}{\sum_{l=1}^{N} \boldsymbol{\omega}_G^{ml} \exp \boldsymbol{\omega}_A^{ml}} \tag{3}$$

ここで，$\boldsymbol{\omega}_A^{mn}$ は $\boldsymbol{h}_{in}^{<el>}$ から式 (2) のように計算される．その後，各ヘッドの出力である自己注意 $\boldsymbol{h}_{sa}$ は $\boldsymbol{h}_{sa} = \boldsymbol{\omega}^{mn}\boldsymbol{V_E}$ のように計算される．N_E 個のアテンションヘッドからの出力は以下のように結合される．

$$\boldsymbol{h}_{mh} = \{\boldsymbol{h}_{sa}^{<1>}, \boldsymbol{h}_{sa}^{<2>}, ..., \boldsymbol{h}_{sa}^{<N_E>}\}\boldsymbol{W}^M \tag{4}$$

$$\boldsymbol{h}_{sa}^{<i>} = \boldsymbol{\omega}^{mn}\boldsymbol{V}_E\boldsymbol{W}_i^V$$

ここで，$\boldsymbol{W}^M$ は $\boldsymbol{h}_{mh}$ の重みを表す．$\boldsymbol{h}_{mh}$ は $\boldsymbol{h}_{in}^{<el>}$ と結合される．最終的に各エンコーダレイヤの出力 $\boldsymbol{h}_{out}^{<el>}$ は $\boldsymbol{h}_{out}^{<el>} = f_{FFN}(\boldsymbol{h}_{mh})$ のようにして得られる．ここで，$f_{FFN}(\cdot)$ は FFN 層を表す．

3.2.3 Transformer デコーダ

Transformer デコーダは，Masked multi-head attention 層，Multi-head attention 層，及び FFN 層で構成されるデコーダレイヤを 6 層重ねたものからなる．Transformer デコーダの入力は $\boldsymbol{h}_{out}^{<enc>}$ 及び $\boldsymbol{y}_{1:m}$ であり，$\boldsymbol{y}_{1:m}$ は入力文を表す．Transformer デコーダからの出力はトークン列 $\boldsymbol{h}_{1:n}^{<dec>}$ で表される．$\boldsymbol{y}_{1:k}$ 及び $\boldsymbol{h}_{1:k}^{<dec>}$ は以下のようなトークン列として定義される．

$$\boldsymbol{y}_{1:k} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_k)$$

$$\boldsymbol{h}_{1:k}^{<dec>} = (\boldsymbol{h}_1^{<dec>}, \boldsymbol{h}_2^{<dec>}, ..., \boldsymbol{h}_k^{<dec>}) \tag{5}$$

訓練時には $m = N$ とする．ここで N は文長を表す．また，テスト時には j 番目の単語を予測するとして $m = 1 : j - 1$，$n = j$ とする．

使用した損失関数を以下に示す．

$$L = -\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J}\log(p(\hat{y}_{ij})) \tag{6}$$

ここで，I 及び J はサンプル数およびそれぞれの指示文長を表す．また，i 及び j はこれらのインデックスを表す．さらに，$p(\hat{y}_{ij})$ は予測単語 $\hat{y}_{ij}$ についての予測確率を表す．サンプルは対象物体，目標領域，コンテクスト情報，及び指示文で構成される集合として定義される．

4. 実験

4.1 データセット

再現性の観点から，公開されている標準データセット PFN-PIC を使用した [Hatori 18]．データセットには対象物体の領域と目標領域に関する情報が含まれており，本研究における評価のためには十分であった．実験では PFN-PIC データセットを使用してモデルを検証した．

PFN-PIC データセットは英語の物体操作指示文を含む [Hatori 18]．指示文は Amazon Mechanical Turk を利用したクラウドソーシングを介してアノテータによって付与された．アノテータは対象物体及び目標領域について直感的で明確な表現を使用して文章を付与するように指示された．[Hatori 18] で報告されているように，データセットにはあいまいな，又は誤った表現が含まれている．本研究ではこれらの表現は除外しなかった．

PFN-PIC データセットには，対象物体ごとに少なくとも三人のアノテータによって指示文が付与された．データセットは訓練集合と検証集合の二つに分かれている．訓練集合は，1,180 の画像，25,900 の対象物体，及び 91,590 の指示文で構成される．検証集合は，20 の画像，352 の対象物体，及び 898 の指示文で構成される．

本研究では PFN-PIC データセットの訓練集合を訓練集合及び検証集合に分割した．パラメータの訓練には訓練集合を使用し，適切なハイパーパラメータを選択するために検証集合を使用した．PFN-PIC データセットの検証集合をテスト集合として使用し，性能を評価した．訓練，検証，及びテスト集合には，それぞれ 81,087，8,774，及び 898 のサンプルが含まれていた．

4.2 パラメータ設定

CRT のハイパーパラメータの値は ORT に基づいた．具体的には，Transformer エンコーダのエンコーダレイヤ及び Transformer デコーダのデコーダレイヤはそれぞれ 6 層であり，アテンションヘッドは 8 個である．最適化関数には Adam を使用し，学習率は 5.0×10^{-4} とした．訓練の際エポック数は 10 とし，バッチサイズは 15 とした．

表 1: 各手法による生成文の，BLEU4，ROUGE-L，METEOR，CIDEr-D，SPICE による評価結果

Method	BLEU4	ROUGE-L	METEOR	CIDEr-D	SPICE
ABEN [Ogura 20]	**15.2**$_{\pm 0.8}$	46.8$_{\pm 1.0}$	21.2$_{\pm 0.8}$	18.2$_{\pm 1.8}$	23.4$_{\pm 2.1}$
ORT [Herdade 19]	8.0$_{\pm 1.2}$	39.4$_{\pm 0.7}$	17.3$_{\pm 0.7}$	27.9$_{\pm 2.8}$	26.4$_{\pm 1.3}$
Ours	14.9$_{\pm 1.1}$	**49.7**$_{\pm 1.0}$	**23.1**$_{\pm 0.7}$	**96.6**$_{\pm 12.0}$	**44.0**$_{\pm 2.3}$

訓練可能パラメータ数は 5,900 万であった．提案手法の学習はメモリ 11GB 搭載 GeForce RTX 2080 Ti，メモリサイズ 64GB の RAM，及び Intel Corei9-9900K により行われた．なお，Up-Down Attention [Anderson 18] による特徴抽出のみメモリ 24GB 搭載 TITAN RTX，メモリサイズ 256GB の RAM，及び Intel Corei9-9820X という構成により行われた．

事前学習とファインチューニングは合わせて 1 時間程度で完了した．未知集合に対して最高の性能を持つモデルを選択するため，学習時，3000 サンプル分行うごとに検証集合に対してその時点での性能を評価した．最終的に主要尺度である SPICE [Anderson 16] が検証集合に対して最高を記録した際のモデルを用いてテスト集合に対して評価を行った．

4.3 定量的結果

実験では，BLEU4，ROUGE-L，METEOR，CIDEr-D，そして SPICE の 5 つの自然言語生成タスク用標準自動評価尺度により生成文を評価する．評価における主要尺度は画像キャプションタスクの標準的尺度である SPICE とした．

CRT と ORT [Herdade 19]，及び ABEN [Ogura 20] を比較した． 表 1 は定量的結果を示している．各手法に対し 5 回実験を実行した．表は，各尺度の平均と標準偏差を示している．

CRT と ORT を比較することにより，CRB が性能にどの程度貢献したかを調べた．また，CRT と ABEN を比較することにより，Transformer エンコーダ-デコーダの有効性を調べた．ORT は 20 エポック訓練を行った． ORT に関して，PFN-PIC データセット [Hatori 18] に最適なパラメータ設定で結果を示した．ABEN は 30 エポック訓練を行った．ABEN の訓練集合では，全てのサンプルを使用したとき訓練が終了条件を満たさなかったため，4,000 サンプルのみ使用した．

まず CRT と ABEN を比較する．表より，SPICE が 20.6 ポイントと大幅に改善されたことが分かる．これより，Transformer エンコーダ-デコーダが FIG タスクへの効果的なアプローチであったことが示された．

次に，CRT と ORT を比較する．表より，SPICE が 17.3 ポイントと大幅に改善されたことが分かる．これより CRT が CRB によって対象物体及び目標領域の参照表現を正しく扱えることが示された．

4.4 Ablation studies

入力情報の種類について Ablation studies を行った．表 2 に 5 回の実験の平均と標準偏差を示す．どの入力特徴量が性能向上に最も貢献しているかを調べた．入力特徴量 $\boldsymbol{x}^{<targ>}$，$\boldsymbol{x}^{<dest>}$，及び $\boldsymbol{X}^{<cont>}$ の組み合わせについて，表 2 に示す 4 つの条件（a）から（d）を検討した．

条件（a）を条件（b）及び（c）と比較すると，SPICE はそれぞれ 1.5，9.5 ポイント減少した．この結果より，$\boldsymbol{x}^{<targ>}$ が性能向上に最も貢献したことが分かった．条件（b）を条件（c）と比較すると，SPICE は 8.0 ポイント減少した． このことから，$\boldsymbol{x}^{<dest>}$ は $\boldsymbol{X}^{<cont>}$ よりも性能向上に貢献した．一方，条件（c）を条件（d）と比較すると，SPICE は 1.5 ポイント減少した．このことから $\boldsymbol{X}^{<cont>}$ も性能向上に貢献することが分かった．

表 2: Ablation studies の結果

Ablation condition				BLEU4	ROUGE-L	METEOR	CIDEr-D	SPICE
Condition	$\boldsymbol{X}^{<cont>}$	$\boldsymbol{x}^{<dest>}$	$\boldsymbol{x}^{<targ>}$					
(a)	✓	✓		$13.3_{\pm 1.0}$	$44.0_{\pm 0.5}$	$20.4_{\pm 0.4}$	$37.0_{\pm 2.8}$	$33.0_{\pm 0.8}$
(b)	✓		✓	$10.3_{\pm 0.6}$	$44.6_{\pm 1.1}$	$19.7_{\pm 0.7}$	$81.8_{\pm 7.0}$	$34.5_{\pm 2.5}$
(c)		✓	✓	$\mathbf{14.9_{\pm 1.0}}$	$49.3_{\pm 1.1}$	$23.0_{\pm 0.5}$	$92.1_{\pm 6.1}$	$42.5_{\pm 2.5}$
(d)	✓	✓	✓	$\mathbf{14.9_{\pm 1.1}}$	$\mathbf{49.7_{\pm 1.0}}$	$\mathbf{23.1_{\pm 0.7}}$	$\mathbf{96.6_{\pm 12.0}}$	$\mathbf{44.0_{\pm 2.3}}$

Ref: “grab the cola can near to the white gloves and put it in the upper right box”
ABEN: “move the white bottle with black labels from the upper left box to the upper left box”
Ours: “move the red can from the bottom right box to the top right box”

図 3: 実験における入力画像，参照文，及び生成文の例

4.5 定性的結果

図 3 に定性的結果を示す．この図では，上部に入力画像，下部に参照文と ABEN 及び CRT による生成文が示されている．水色及び赤の領域は，それぞれ対象物体及び目標領域を表す．

図のサンプルは正しく文が生成された例を示す． サンプルでは，対象物体及び目標領域はそれぞれ「右下のボックスにあるコーラ缶」及び「右上」である．この入力画像中には二つのコーラ缶があるため，指示文で対象物体を明確に指定する必要がある．参照文では，缶は “the cola can near to the white gloves” と表現されていた．CRT による生成文では，対象物体は “the red can from the bottom right box.” と表現された．これより CRT が，参照文とは異なるものの妥当な空間参照表現を使用して特定に成功したことが示された．一方，ABEN による生成文では，対象物体及び目標領域はそれぞれ “the white bottle with black labels from the upper left box”，及び “the upper left box” と表現された． これらの表現はいずれも不正確であった．

4.6 被験者実験

参照文，ベースライン [Ogura 20] による生成文，及び CRT による生成文を比較する被験者実験を行った．評価尺度として Mean Opinion Score（MOS）を使用した．被験者は 20 代の 5 人とした．被験者にはそれぞれの作業速度で評価を依頼した．実験では 50 枚の画像がテスト集合から無作為に抽出された．それらの画像と付随する参照文及び生成文は，被験者に無作為に提示された．被験者は次のように 5 段階で指示の明瞭さの観点から文を評価した．

1： とても悪い　2： 悪い　3： 普通　4： 良い　5： とても良い

表 3 は MOS の平均及び標準偏差を示す．表から，参照文の MOS は 3.8 であり，これがこの実験の上限値と見なされる．CRT 及び ABEN の MOS はそれぞれ 2.6，1.2 だった．CRT と ABEN の間の統計的有意性は $p = 8.8 \times 10^{-35}$ (< 0.001) である．これより，被験者実験においても CRT が ABEN を上回ることが示された．

表 3: 正解文及び生成文についての，MOS 値による評価実験の結果

Method	MOS
Reference sentences (upper bound)	3.8 ± 1.2
ABEN [Ogura 20] (baseline)	1.2 ± 0.4
Ours	2.6 ± 1.5

5. おわりに

クロスモーダル言語理解のためのほとんどのデータ駆動型アプローチには，大規模なコーパスが必要である．ただし，このようなコーパスの構築には時間と費用がかかる．そこで，対象物体及び目標領域の参照表現を含む物体移動指示文を生成できるクロスモーダル言語生成モデルである Case Relation Transformer（CRT）を提案した．CRT では，対象物体，目標領域，及びコンテクスト情報間の関係を扱うために，Case Relation Block が導入された．CRT は，FIG タスクにおいて主要尺度でベースラインを上回った．

参考文献

[Anderson 16] Anderson, P., Fernando, B., Johnson, M., and Gould, S.: Spice: Semantic propositional image caption evaluation, in *ECCV*, pp. 382–398Springer (2016)

[Anderson 18] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering, in *CVPR*, pp. 6077–6086 (2018)

[Hatori 18] Hatori, J., Kikuchi, Y., Kobayashi, S., Takahashi, K., Tsuboi, Y., Unno, Y., Ko, W., and Tan, J.: Interactively picking real-world objects with unconstrained spoken language instructions, in *ICRA*, pp. 3774–3781IEEE (2018)

[Herdade 19] Herdade, S., Kappeler, A., Boakye, K., and Soares, J.: Image captioning: Transforming objects into words, in *NeurIPS*, pp. 11137–11147 (2019)

[Ogura 20] Ogura, T., Magassouba, A., Sugiura, K., Hirakawa, T., Yamashita, T., Fujiyoshi, H., and Kawai, H.: Alleviating the Burden of Labeling: Sentence Generation by Attention Branch Encoder–Decoder Network, *RA-L*, Vol. 5, No. 4, pp. 5945–5952 (2020)

[Ren 16] Ren, S., He, K., Girshick, R., and Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks, *TPAMI*, Vol. 39, No. 6, pp. 1137–1149 (2016)

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.: Attention is all you need, in *NeurIPS*, pp. 5998–6008 (2017)

[Zhao 21] Zhao, M., Anderson, P., Jain, V., Wang, S., Ku, A., Baldridge, J., and Ie, E.: On the Evaluation of Vision-and-Language Navigation Instructions, *arXiv preprint arXiv:2101.10504* (2021)