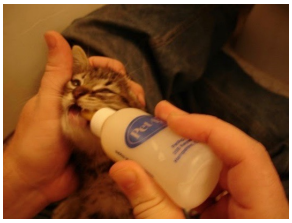
マルチモーダル推論評価のための日本語データセットの試案

飯野 早貴[1]　石田 真捺[1]　小谷野 華那[1]　松本 留奈[1]
鈴木 莉子[1]　谷中 瞳[2,1]　峯島 宏次[1]　戸次 大介[1]
[1] お茶の水女子大学　[2] 理化学研究所
{g1720502,g1720504,g1720544,g1720540,suzuki.riko,bekki}@is.ocha.ac.jp,
hitomi.yanaka@riken.jp, mineshima.koji@ocha.ac.jp

1　はじめに

自然言語のテキストデータと画像などの非テキストデータを接合し、知識獲得を行うマルチモーダル推論の研究が、近年盛んに行われている。画像とテキストという2種類のデータを統合的に処理する研究として、画像から新たなキャプションを生成する研究 [5] や、画像に対して文の真偽判定を行う Visual Reasoning [7] の研究がある。例えば、次の図1の例を考えよう。



(1)　人が猫の顔を手で押さえて、猫にミルクを飲ませている。(TRUE/FALSE)

図 1: 画像推論の例：画像は Visual Genome [8] による。

この画像が表す状況で (1) の文が真であるのか偽であるのか判定するというプロセスは、文と画像の複雑な処理を要する。まず、(1) の文の構造と単語の意味に基づいて文の意味を合成し、その上で画像に現れるエンティティとその性質・関係を認識し、この文と画像が表す情報を組み合わせて文の真偽値を決定する必要がある。さらに、(1) の文の意味を理解するためには、使役（「飲ませて」）や照応（「猫の顔」と「猫に」の「猫」が同じ対象を指す）など、複雑な言語現象を適切に扱う必要がある。また、(1) の例では、猫が飲んでいるものがミルクかどうかを画像だけから判断することは人にとっても困難であるが、仮に「猫がミルクを飲んでいる」という文がこの画像のキャプションとして付与されていれば、真偽の判定は容易となる。しかし、画像から獲得できる情報とキャプションから獲得できる情報をどのように組み合わせて (1) の真偽値を判定できるのかは、まったく自明ではない。

本研究は、このような複雑なマルチモーダル推論を評価するためのデータセットを導入することを目的とする。既存の画像キャプションデータセット [4, 10] では、画像に付与されたキャプションは比較的短く単純なものにとどまっており、また画像に現れる物体やその性質・関係の認識も、ごく限られた空間的パターンに限定されるという傾向がある。さらに、画像キャプションに含まれる言語現象には一定の傾向があることが指摘されており [1]、データセットの様々なバイアスの要因となりうる。こうした理由から、英語では、複雑な Visual Reasoning タスクの評価・学習のために設計されたデータセットとして、NLVR [6], CLEVR [2], NLVR2 [7] などが提案されている（2節を参照）。一方、日本語では、大規模な画像キャプションデータセット [9] はあるものの、複雑な言語現象を含む文と画像のペアからなる画像推論のデータセットは存在しない。

以上の背景をふまえて、本研究では、図1にあるような画像に対して文の真偽評価を行うタスクを Visual-Textual Reasoning (VTE) タスクと呼び、VTE タスクを評価するための日本語データセットを構築する。本稿では、データセットの設計と現在までに構築した部分について報告する。構築したデータセットは、研究利用可能な形式で公開する予定である。

2　関連研究

英語では、複雑な画像推論タスクを含むデータセットとして、NLVR [6] と CLEVR [2] がある。これらは、自動合成された文や画像 (synthetic data) を用いており、その点で自然さ、多様性に欠けるという問題がある。NLVR2 [7] は、ウェブから収集した写真を用いてクラウドソーシングにより構築された大規模な画像推論データセットである。2枚の画像ペアに対してそれぞれに写っている対象物の数量・比較や、その位置関係を記述するようワーカーに指示を与えることで、従

来の画像キャプションと比べて、より多様で複雑な文を収集するように設計されている。しかし、NLVR2では、画像内の物体の性質・関係を記述する文と、「どちらの画像も〜を含んでいる (Both images contain ...)」のように画像の外からメタ的な記述を行う文とが混在しており、そのため、量化や否定、数量表現などの言語現象がどちらのレベルに生じているのか（画像内の物体・出来事の記述か、メタ的な記述か）、区別されていないという問題がある。

STAIR Captions [9] は、1 枚の画像に対して 5 つのキャプションが付与されている日本語の大規模データセットである。キャプションを作成する際のガイドラインとして、「だ・である調で書く」「画像に映っていないものを想像して書かない」「感情・意見を書かない」といった条件を与えることで客観的なキャプションを記述する工夫が施されている。しかし、画像と文間の言語的に複雑な推論をターゲットとして設計された英語データセットとは異なり、あくまでキャプション（説明文）の付与を目的としているため、日本語の構文の多様性・複雑性は十分に考慮されていない。

3 日本語 VTE データセットの構築

本節では、以上の先行研究をふまえて、多様な言語現象を含む複雑な文と画像からなる日本語 VTE タスクのデータセットを構築する方法について説明する。

3.1 画像-文の推論ペアの構築

画像の選択 まず、Visual Genome [8] の 108,077 件のデータから画像が 100 件以上ある WordNet synset (以下、synset）を含むデータ 1381 件を取り出した。このうち、synset の上位語が car.n.01, furniture.n.01, natural_object.n.01, animal.n.01 のいずれかである 120 件の画像をアノテーションの対象とした。

次に、各 synset の画像から、ランダムに 16 枚の画像を選択した。この中から NLVR2 の画像選択の方法を参考に、複雑な文を作成しやすい画像を 8 枚選択した。具体的には、(1) synset の物体が写っていない、(2) synset の物体のみが写っている（背景が無地など）、(3) イラスト画像である、のいずれかの条件を満たす画像は複雑な文を作成することが困難であると予想されるため、除外した。また、以下の条件を満たす画像を優先的に選択した。

- 2 つ以上の synset の物体が写っている
- synset の物体と他の物体との関係が写っている（「ソファの上に猫がいる」など）
- synset の物体が活動している（「猫が走っている」など）
- 様々な物体が写っていたり synset の特徴を表したりしている

日本語文の作成 本研究では 2 種類の方法で、画像-文ペアの作成を試みる。まず基本的な方法として、アノテーション対象となる 1 枚の画像に対して、真もしくは偽となる文を作成した。その際、日本語の多様な構文を取り込むため、日本語意味論データセット [3] を参考に、推論で重要な役割を果たす意味現象 12 種類を分類し（詳細は 3.2 で述べる）、それぞれに特徴的な語彙・構文をまとめ、文作成のためのガイドラインを作成した。

もうひとつの方法として、NLVR2 [7] を参考に、複数の物体やそれらの間の多様な属性・関係を捉えた複雑な文を作成しやすいように、2 枚の画像をペアにして真ないし偽となる文を付与する方法を試みた。より具体的には、上記の方法で選択した 8 枚の画像を 2 枚ずつ 4 組に分け、その 4 組のうち 2 組に関して真であり、残りの 2 組に対して偽となるような日本語文を以下のルールを含むガイドラインに従って作成した。

1. 2 枚の画像を 1 つと見なしその全体に当てはまる文を作成するか、2 枚の画像をそれぞれ比較する。
2. 15 文字以上の文にする。
3. 単体のみについての言及は避ける (「犬がいる」など)。

2 枚の画像に対して文を作成する際には、label_one_image（いずれか 1 枚に当てはまる文を作成する）, label_left_image（左の画像に当てはまる文を作成する）, label_right_image（右の画像に当てはまる文を作成する）, label_each_image（2 枚の画像を比較しそれぞれに当てはまる文を作成する）, label_all_image（2 枚の画像を 1 つとみなし全体に当てはまる文を作成する)、という 5 つのラベルを使用した。このうち、label_left_image と label_right_image は label_one_image の特殊ケースであるため、それぞれ重複して付与することを認める。

画像ペアの具体例を図 2 に示す。「2 頭のシマウマがいて、互いに違う方向を向いている」という文は、画像ペア (i) と (ii) に対しては偽、(iii) の右側の画像と (iv) の左側の画像に対して真となるため、label_one_image もしくは、左右どちらの画像に対して真であるのかを明示して、label_left_image と label_right_image を同時に付与することができる。「立ったまま地面に顔をつけているシマウマはいない」という文は、(ii) と (iv) のペアのそれぞれの画像に対して真であり、(i) と (iii) のそ


画像ペア (i)


画像ペア (ii)


画像ペア (iii)


画像ペア (iv)

図 2: 画像ペアの例

れぞれの画像に対して偽となるため、label_each_image が付与される。「顔を地面に近づけて立っているシマウマが 2 頭いる」という文は、(ii) と (iv) を合わせて 1 つの画像とみなしたとき真となり、(i) と (iii) を合わせて 1 つとみなしたときは偽となるため、この場合 label_all_image を付与することができる。この方法により画像ペアに対して体系的に真偽ラベルを付与することができる。

3.2 意味現象タグの付与

構築した文がどの程度言語現象を網羅しているかを分析するため、上記のガイドラインに基づき、各文に現れる意味現象のタグ付けを行う。表 1 に示す基本的な 12 種類の意味現象タグを用意した。このうち、「猫がいる」における「いる」は存在を表し、存在量化 (EXIST) タグを付与されるのに対して、「犬が走っている」の場合は、動詞がイベントの進行相を表すため、アスペクト (ASP) タグが付与される。タグ付けには、brat[1]を使用した。以下にタグ付けの例を示す。意味現象タグ付与のガイドラインはデータセットとともに公開する予定である。

(2) カモ [CONJ と] [CARD 8 羽の] 白鳥が [UNIV 全て] 水の [REL 上] に [EXIST いる]。

(3) [CARD 15 匹以上の] 羊が [EXIST い] [CONJ て]、[ANA その] 内 [CARD 2 匹以上の] 子羊が [EXIST いる]。

(4) 顔を地面に [REL 近づけ] [CONJ て] [ASP 立っている] シマウマが [CARD 2 頭] [EXIST いる]。

[1]http://brat.nlplab.org/

現象	タグ	例
連言	CONJ	犬と猫
		安くて美味しい
選言	DISJ	犬か猫
		立っているか座っている
全称量化	UNIV	どの猫も白い
存在量化	EXIST	猫がいる
色属性	COLOR	白い猫
		赤く光る
関係	REL	靴を履いた猫が
		机の上にいる
否定	NEG	寝そべっていない猫
照応	ANA	数匹の猫がいて、
		そのうち一匹は白い
数量	CARD	3 匹の猫が寝ている
比較	COMP	犬より小さい猫
アスペクト	ASP	犬が走っている
タ	TA	メガネをかけた男性

表 1: 意味現象タグの一覧

3.3 データセットの概要

2020 年 1 月時点で、画像 592 枚に対して真もしくは偽となる文を合計 680 件付与した。構築した文の語彙数は 7,448 語、各文の平均単語数は 17 語であった。単語分割には、Mecab[2]を使用した。データセットに含まれる 680 文のうち、277 文を抽出し、意味現象タグのアノテーションを行った。そのうち、真の文は 177 文、偽の文は 100 文である。表 2 に意味現象タグの分布（出現数及び、アノテーション対象の全 277 文のうち、当該の意味現象タグを含む文の割合）を示す。否定 (NEG) や比較（COMP）など、従来の画像キャプションには現れにくい言語現象が含まれていることがわかる。

構築した画像-文ペアと真偽ラベルの例を図 3 に示す。

3.4 Human performance の分析

構築したデータセットの難易度を推定するため、人間のパフォーマンスの正答率を調査した。構築したデータのうち、正解ラベルが真である文と偽である文をランダムに 198 文（真の文 97 件、偽の文 101 件）抽出し、4 名の作業者がそれぞれの画像-文ペアについて真偽判定を行った。文を作成した作業者は、自分が作成した文については回答しないようにした。実験の結果、正答率の平均は 0.958、標準偏差は 0.008 であった。

[2]http://taku910.github.io/mecab/

タグ	出現数	割合 (%)
REL	217	22.60
ASP	199	20.73
CARD	113	11.77
EXIST	105	10.94
COLOR	80	8.33
CONJ	69	7.19
NEG	41	4.27
COMP	37	3.85
TA	28	2.92
UNIV	25	2.61
DISJ	24	2.50
ANA	22	2.29
合計	960	100

表 2: 意味現象タグの分布（277 文が対象）

- 白い机の上にまな板があって、その上に切られた野菜が置いてある。(正解：TRUE)
- 切られていない人参がある。(正解：FALSE)

図 3: 画像-文ペアと真偽ラベルの具体例

正答率が低かった問題として、図 4 の例がある。1 人の作業者は FALSE と回答した。これは、カゴの後ろにあるぬいぐるみのうち、大きい方はテーブルの上にあるのか、テーブルよりも奥に置かれているのかの判定が難しいという要因が考えられる。このように対象に様々な属性・関係を帰属させる基準にゆれが生じるケースについては、人による正答率に基づいて選別するなど、データの質を改善する余地が残されている。

4 おわりに

本研究では、日本語 VTE データセットの設計と現在までの構築部分について報告した。構築したデータセットに対する意味現象タグのアノテーション結果は、本データセットが多様かつ複雑な言語現象を含んでいることを示している。また、人手によるパフォーマンス評価を行った結果、作成した文には複雑な言語現象が多く含まれているにもかかわらず、人にとっては自然で比較的に容易な推論であることが確認できた。今後

図 4: テーブルの上にカゴが一つ、ミシンが一台、ぬいぐるみが二つある。(正解：TRUE)

の展望として、日本語 VTE データセットのさらなる拡張を進めるとともに、このデータセットを用いた現行のマルチモーダルシステムの評価を進めていく。

謝辞. 本研究の一部は、JST AIP-PRISM JPMJCR18Y1、および JSPS 科研費 JP18H03284 の助成を受けたものである。

参考文献

[1] Malihe Alikhani and Matthew Stone. “caption” as a coherence relation: Evidence and implications. In *Proc. of the Second Workshop on Shortcomings in Vision and Language*, pp. 58–67, 2019.

[2] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

[3] Ai Kawazoe, Ribeka Tanaka, Koji Mineshima, and Daisuke Bekki. An inference problem set for evaluating semantic theories and semantic processing systems for japanese. In *New Frontiers in Artificial Intelligence*, pp. 58–65, 2017.

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Proc. of ECCV*, pp. 740–755, 2014.

[5] Takashi Miyazaki and Nobuyuki Shimizu. Cross-lingual image caption generation. In *Proc. of ACL*, pp. 1780–1790, 2016.

[6] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proc. of ACL*, pp. 217–223, 2017.

[7] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proc. of ACL*, pp. 6418–6428, 2019.

[8] Ranjay Krishna *et al.* Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. 2016.

[9] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proc. of ACL*, pp. 417–421, 2017.

[10] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, Vol. 2, pp. 67–78, 2014.