形式意味論に基づく自然言語と画像間のマルチモーダル推論システムの構築

理学専攻・情報科学コース　鈴木 莉子 （指導教員：戸次 大介）

1　はじめに

近年では，画像，音声，テキストといった複数のモダリティからの情報を組み合わせて新しい知識を獲得することを目的としたマルチモーダル研究が注目されている．本研究では，文と画像の情報を一階述語論理 (FOL) のモデルと論理式を用いて表現し，数量や否定を含む複雑な言語現象を伴う文を画像から推論するシステムを構築した．さらに，システムの対応言語を日本語に拡張すると共に，画像に対して日本語文と真偽値を付与した評価用データセットの作成を試みた．

2　画像の意味表現

画像の意味表現として，scene graph と First-order logic (FOL) structure の 2 種類の構造を紹介する．どちらの表現手法も画像に写る物体情報やその属性情報，そして物体間の関係について記述できる．

2.1　Scene graph

Scene graph は画像に写る物体情報やその属性情報，そして物体間の関係をグラフ構造で記述したものである [3]．Scene graph の各ノードは物体情報やその属性情報に対応し，各エッジは物体間の関係に対応する．

2.2　FOL structure

論理推論に使うことができる画像情報の意味表現としては，一階述語論理のモデル (FOL structure) を用いる手法がある [2]．FOL structure $\mathcal{M}$ は，ドメイン D と解釈関数 I から定義される [1]．ドメインは空でない集合であり，画像中に存在するエンティティの情報を表す．解釈関数は各 n 項述語を D^n の部分集合に対応づける関数であり，ドメイン中のエンティティがもつ属性と関係を表現する．

3　画像と文間の推論システムの構築

本章では画像と文間の推論システムの構築手順について説明する．まず画像と英語文を入力とするマルチモーダル推論システムについて説明し，次に本システムを日本語文に対応させる方法について述べる．

3.1　画像と英語文間の推論システム

本節では画像と英文間の論理推論システムについて説明する．提案手法では，画像情報の意味表現として FOL の論理式と構造 (モデル) を採用する．本システムの全体像を図 1 に示す．本システムは大きく分けて 4 つのモジュールに分けられる．

(a) **Graph Translator**: Scene Graph や FOL structure で記述された画像の意味情報を 2 通りの方法で論理式に変換する．

(b) **Semantic Parser**: 意味解析・推論システム ccg2lambda[5] を用いて文を論理式に変換する．

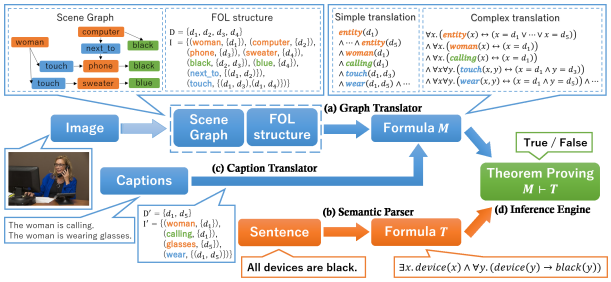


図 1: システムの全体像

(c) **Caption Translator**: 画像に付与されたキャプション情報を FOL structure に変換する．キャプション情報を FOL structure に統合することにより，キャプションでのみ記述できる関係を反映した意味表現を取得できる．

(d) **Inference Engine**: 各画像の論理式を前提，文の論理式を帰結として，その画像が文を含意するか否かを定理証明器を用いて推論する．含意関係が成り立つ場合，その画像は結論の文が成り立つ状況を表すものとみなせる．

3.2　画像と日本語文間の推論システム

マルチモーダル推論システムの対応言語を日本語に拡張するために，意味解析と推論器を改良することを試みる．日本語文に対応したシステムでは，(1) 意味表現としては，イベント意味論 (event semantics)[6] に基づくものを採用し，(2) 推論器としては，FOL をその部分系として含む高階型理論に基づく Coq[7] を用いる．Coq は自然演繹に基づく半自動の定理証明支援系として知られているが，その自動証明部分を利用することで，定理証明に基づく自然言語推論の研究が行われている．これらの成果をふまえて，数量表現などを含む複雑な推論を自然言語の構造に即した形で効率的・統一的に扱える画像と日本語文間の論理推論システムを構築した．

4　評価実験

4.1　実験に用いたデータセット

本実験では，画像に対して scene graph が付与された大規模データセット Visual Genome[4] と，画像に対して FOL structure と真または偽の文が付与されたデータセットである GRIM[2] の 2 種類のデータセットを用いる．否定や量化表現を含む評価文を作成し，各画像に対してひょか分が真または偽であるか 2 名のアノテータが判断し，正解ラベルを付与した．

4.2　サーカムスクリプションによる解析性能の評価

画像の意味表現を，原始論理式の連言への翻訳 (Simple) とサーカムスクリプションに基づく翻訳

(Hybrid) を用いて画像の意味情報を論理式に変換した場合における，推論精度 (F 値) と，定理証明にかかった平均時間 (秒) を比較する．データセットは Visual Genome と GRIM を用い，それぞれの実験結果を報告する．定理証明器は Prover9 を用いる．表 1 に GRIM データセットを用いた実験結果を示す．

手法	F 値 / 速度 (秒)				
	論理結合子	数詞	量化	関係	否定
Simple	68.21 / 8.9	80.89 / 8.8	0.0 / 10.4	73.12 / 9.3	64.45 / 9.0
Hybrid	84.04 / 12.1	95.36 / 9.2	**75.77** / **35.0**	88.19 / 13.3	**88.19** / **19.9**

表 1: GRIM を用いたモデルの翻訳方法の比較．精度 (F 値) と証明に要した速度を示す．

サーカムスクリプションにより，特に *every* や *all* など量化を含む文による検索の精度の改善が見られ，さらに *not* など否定を含む文の精度も大きく改善した．サーカムスクリプションにより，全ての分類において検索の精度の改善が見られるが，同時に検索にかかる時間も増大していることが分かる．

4.3 キャプションからモデルへの変換の評価

キャプションから FOL モデルへの変換の正しさを，GRIM のキャプションを用いた含意関係認識を行うことで評価した．キャプションより更新されたモデルを用いて，各画像ごとに真と偽の 2 種類のキャプションとの含意関係を判定した結果は，適合率，再現率，F 値がそれぞれ 0.91, 0.73, 0.81 であった．再現率が低くなった原因として，キャプションに含まれる物体が画像中のどの物体を指すかが一意に決められず，モデルの更新に失敗したケースが見られた．

4.4 画像と日本語文間の推論システムの評価

画像と日本語文間の推論システムの精度を評価するため，否定や量化などを含む複雑な日本語文 20 文を入力として，各文につき真となる画像と偽となる画像をそれぞれ 3 件程度用い，計 115 問の真偽判定を行なった．画像は FOL structure がアノテートされている GRIM データセット [2] を用いる．

実験結果を表 2 に示す．表 2 から，イベント意味論に基づくテンプレートを用いることで，非イベント意味論に基づくものと比べて正答率が向上したことがわかる．また，入力文の言語現象ごとに見てみると，特に数量表現についての精度が向上していた．

CCG パーザ	意味論	正答率 (%)
Jigg	plain	70.4
	event	74.8
depccg	plain	71.3
	event	77.4

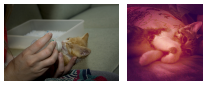
表 2: CCG パーザと意味論ごとの正答率

5 日本語マルチモーダルデータセット

2021 年 1 月時点で，画像 592 枚に対して真もしくは偽となる日本語文を合計 680 件付与した．構築した画像-文ペアと真偽ラベルの例を図 2 に，画像ペアに対して文と真偽ラベルを付与した例を図 3 に示す．

✓ ベリーを含めて多くの食べ物がある．
✓ 座っている人はいない．
✗ 全ての人間が座って何かを食べている．
✗ 白い服を着た男性が椅子に座っている．

図 2: 画像-文ペアと真偽ラベル (✓:TRUE，✗:FALSE) の具体例．



(i) TRUE

(ii) FALSE

(iii) FALSE

(iv) TRUE

- 1 匹の子猫が哺乳瓶をくわえている．

図 3: 4 組の画像ペアに対して，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付与した例．

6 おわりに

本論文では，画像情報とテキスト情報を論理ベースの意味表現を用いて統一的に扱い，論理推論を行うシステムを提案した．CCG の意味解析と定理証明を組み合わせることで，文と画像に含まれる属性・関係だけでなく，否定や量化・数量表現を含む意味情報を扱うことが可能であることを示した．さらに，システムの対応言語を日本語に拡張した日本語マルチモーダル推論システムを提案した．加えて，日本語マルチモーダルデータセットの設計と概要について報告した．

今後の展望として，日本語マルチモーダル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．

参考文献

[1] Patrick Blackburn and Johan Bos. *Representation and Inference for Natural Language: A First Course in Computational Semantics*. Center for the Study of Language and Information, Stanford, CA, 2005.

[2] Manuela Hürlimann and Johan Bos. Combining Lexical and Spatial Knowledge to Predict Spatial Relations between Objects in Images. In *Proc. of the Workshop on Vision and Language*, 2016.

[3] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *IEEE/ CVF International Conference on Computer Vision and Pattern Recognition*, pp. 3668–3678. IEEE Computer Society, 2015.

[4] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, Vol. 123, No. 1, pp. 32–73, 2017.

[5] Koji Mineshima, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. Higher-order logical inference with compositional semantics. In *EMNLP*, 2015.

[6] Terence Parsons. *Events in the Semantics of English: A study in subatomic semantics*. MIT Press, 1990.

[7] The Coq Development Team. *The Coq Proof Assistant: Reference Manual: Version 8.9.0*. INRIA, 2019.