CCG と Coq を用いた日本語マルチモーダル推論システムの構築

秋山 雛乃[1] 石嶋 美咲[1] 石田 真捺[1] 高野 紗輝[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
{g1720501,g1720503,g1720504,g1720525}@is.ocha.ac.jp
suzuki.riko@is.ocha.ac.jp, hitomi.yanaka@riken.jp
mineshima.koji@ocha.ac.jp, bekki@is.ocha.ac.jp

1 はじめに

非テキストデータとテキストデータなどのモーダルの異なる情報間で推論することによって、新たな情報を獲得するマルチモーダル推論に関する研究が近年進んでいる。高度なマルチモーダル推論を行うシステムとして、画像と文を論理式による意味表示に変換することで、画像と文間の推論（Visual-Texual Entailment, VTE）を実現する VTE システム [9, 16] がある。しかし、VTE システムは英語を対象としており、その他の言語には対応していないという課題が残されていた。

そこで本研究では、VTE システムを日本語文に対応できるよう改良することを検討する。VTE システムを日本語文に適用するためには、いくつかの課題がある。VTE タスクの具体例として、図 1 が描く状況で、(1) と (2) の文がそれぞれ真であるか偽であるかを判定するという問題を考える。

図 1: 日本語 Visual-Textual Entailment の例

(1) 二匹の羊がいる。(TRUE)
(2) 犬が羊にさわっている。(TRUE)

図 1 の画像が表す状況では、(1) と (2) の文は真であると判断できる。この判断のためには、理論的には、(1) の文の意味を構成的に計算し、画像に現れる物体とその属性・関係・数量を認識した上で、画像に対する文の真偽値を計算する必要があり、複雑な意味処理を伴う。また、(1) の「二匹」のような日本語の数量表現は、「二匹の羊」「羊が二匹」「羊二匹」など、いくつかの構文に出現し、統語的に複雑である。さらに、日本語は項の省略が頻繁に生じるなど、英語の意味解析システムを単純には適用できない側面がある。このように日本語に即した統語・意味解析を行うように VTE システムを改良する必要がある。

そこで本研究では、先行研究の英語 VTE システム [9, 16] を日本語に適合させ、意味解析と推論器を改良することを試みる。英語 VTE システムでは、組合せ範疇文法 (Combinatory Categorial Grammar, CCG) [8] に基づく構文解析を利用し、意味表現と推論は、1 階述語論理 (FOL) とその定理証明器に基づくものを使用した。これに対して、日本語 VTE システムでは、日本語の CCG [15] に基づいて、(1) 意味表現としては、イベント意味論 (event semantics) [7] 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 [11] 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 [5, 12]。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・統一的に扱える日本語 VTE システムの実現を目指す。

2 システムの概要

システムの全体像を図 2 に示す。まず、日本語入力文から論理式に基づく意味表示を導出するため、入力文に対して CCG 構文解析器を用いて、導出木と呼ばれる統語構造から意味表示（論理式）を合成的に導出する。CCG 構文解析器は Jigg [6] と depccg [13] を用いる。導出木から意味表示への構成的なマッピングには、ccg2lambda [4, 14] を用いる。例として、「犬が羊にさわっている」という文の CCG 導出木を図 3 に示す。

出力の意味表示（図 2 の論理式 A）に現れる述語記号は、自動翻訳[1]を用いて日本語から英語に変換し、

[1] googletrans 2.4.0 (https://pypi.org/project/googletrans/) を使用した。

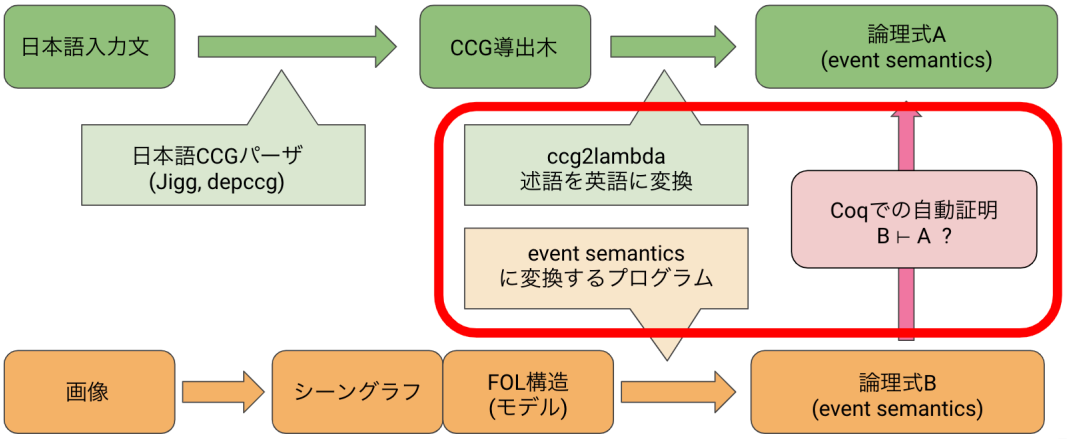


図 2: 日本語 VTE システムの全体像

必要に応じて翻訳ルールを人手で追加した。1 節で述べたように、英語の VTE システムの先行研究 [9] では、意味表現に FOL を採用したのに対して、本研究では意味表現として FOL を拡張した高階型理論を採用し、定理証明支援系 Coq [11] を用いてその推論を実現する。Coq は、高階の公理も扱うことができるため、数量表現を含む文をより簡潔に扱うことが可能となる（詳細は 3 節で述べる）。

次に、画像に対応するシーングラフ（グラフ構造）[1] を FOL 構造に変換し、さらに Coq の形式的表現（図 2 の論理式 B）に変換する。シーングラフは、1 階述語論理の構造（FOL 構造）に対応し [9]、画像に現れるエンティティとその属性・関係を記述する表現系である。Coq の型理論は FOL を拡張した体系であるため、シーングラフ及び FOL 構造を、Coq の体系に翻訳することが可能である。特に Coq における列挙型 (enumeration type) を利用することで、画像に現れるエンティティの簡潔な表現が可能となる（詳細は 4 節で述べる）。

最後に、入力文から変換された論理式 A と画像から変換された論理式 B の間に含意関係が成り立つかどうかを Coq によって自動証明する。B が A を含意する（$B \vdash A$）ならば、入力文はその画像に対して真であり、含意しないならば偽であると判定する。

3 文から論理式への変換

日本語 VTE システムの文から論理式への変換手法は、先行研究 [9] と比べて、(i) Neo-Davidsonian [7] に基づくイベント意味論を導入する点、及び、(ii) 数量表現を扱うためにリストを導入する点において異なる。

イベント意味論　「犬が羊にさわっている」という文は、非イベント意味論では (3) の論理式に、イベント意味論では (4) の論理式にマップされる。

(3) $\exists x_1 x_2.(\mathsf{dog}(x_1) \wedge \mathsf{sheep}(x_2) \wedge \mathsf{touch}(x_1, x_2))$

(4) $\exists e_3 x_1 x_2.(\mathsf{dog}(x_1) \wedge \mathsf{sheep}(x_2) \wedge \mathsf{touch}(e_3) \wedge \mathsf{subj}(e_3, x_1) \wedge \mathsf{obj}(e_3, x_2))$

論理式をイベント意味論で表現することの利点は主に 2 つある。まず、日本語の格表現をイベントを項にとる述語 (subj,obj) として表現することで、項の省略が起きた場合でも論理式が導出可能であり、例えば、「犬が羊にさわっている」から「犬がさわっている」への推論が導けるという点が挙げられる。

また、非イベント意味論では、自動詞は $\mathsf{run}(x_1)$ のように 1 項述語として、他動詞は (3) の $\mathsf{touch}(x_1, x_2)$ のように、2 項述語として表す必要があるため、自動詞・他動詞を統一して扱うことができないのに対して、イベント意味論では自動詞・他動詞をイベントを項にとる 1 項述語として統一的に扱うことができるという利点もある。このため、知識ベースとの接合や自動証明に適した意味表現となっていると言える。

数量表現　英語の VTE システムでは、FOL を使って数量表現を含む文を形式化した。例えば、FOL では (5) は次のように記号化される。

(5) （少なくとも）2 匹の猫がいる。
$\exists x \exists y.(\mathsf{cat}(x) \wedge \mathsf{cat}(y) \wedge (x \neq y))$

この論理式は数量 n が大きくなるにつれて複雑になり、自動証明が困難になるという問題がある。本研究では、Coq の型理論におけるリスト (list) 型を使うことでこの問題に対処する。例えば (6) と (7) は次のように記号化する。

(6) （少なくとも）2 匹の猫がいる。
$\exists x_1.(\mathsf{cat}(x_1) \wedge (\mathsf{length}(x_1) \geq 2))$

(7) ちょうど 2 匹の猫がいる。
$\exists x_1.(\mathsf{cat}(x_1) \wedge (\mathsf{length}(x_1) = 2))$

ここで、変数 x_1 の値となるのはエンティティのリストであり、「二匹の」という数量表現はそのリスト

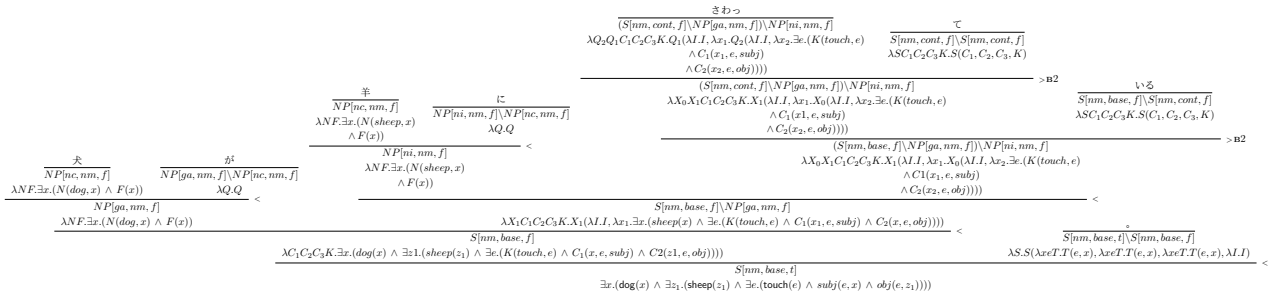


図 3: 「犬が羊にさわっている」の CCG 導出木

の長さによって表すことができる。述語 cat の項となるのは、エンティティのリストであるため、次の累積性の公理 [2] をリストを用いて定式化し、推論を行う。Coq での表現は以下のようになる。

(8) $\forall A : \mathsf{Type}.\forall x : A.\forall l : \mathsf{list}\ A.$
$\forall P : \mathsf{list}\ A \to \mathsf{Prop}.\ (Pl \to P[x] \to P(x :: l))$

4 画像から論理式への変換

シーングラフ [1] は画像中の物体とその属性、また物体間の関係をグラフ表現で表したものである。シーングラフは英語 VTE システム [9] と同じ方法で FOL 構造に変換することができる。本節では FOL 構造を Coq の論理式に変換する方法について述べる。まず、画像中のエンティティの情報は FOL 構造では (9) のように表され、論理式 (10) に変換することができる。

(9) $I(\mathsf{entity}) = \{d_1, d_2, d_3\}$

(10) $\forall x.(\mathsf{entity}(x) \leftrightarrow x = d_1 \wedge x = d_2 \wedge x = d_3)$

ここで、ドメイン（画像中の全エンティティの集合）には「d_1, d_2, d_3 というエンティティしかいない」という限定的（否定的）情報を加えるため、論理式は複雑なものとなる（詳細は [9] を参照）。そこで本研究では、Coq の列挙型を利用する。ドメインが d_1, d_2, d_3 という 3 つのエンティティからという情報を Coq では (9) 次のように表現する。

```
Inductive entity : Type := d1 | d2 | d3
```

この列挙型に基づく entity の定義により、否定的情報を含む命題の証明も可能になる。

5 評価実験

5.1 実験設定

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 構造がアノテートされている GRIM データセット [3] を用い、画像から FOL 構造（シーングラフ）への自動変換は今後の課題とする。

その一部を以下に示す。システムが複雑な言語現象が現れる推論を扱えるかについてテストするため、連言 (CONJ)、選言 (DISJ)、数量表現 (NUM)、否定 (NEG)、全称量化 (UNIV)、存在量化 (EXIST)、関係 (REL) という 7 つの言語現象ラベルを用いて各文に現れる言語現象ラベルを付与した。

(11) [CONJ 白い猫] が [EXIST いる]。
(12) [CONJ 白い猫] が [EXIST い][NEG ない]。
(13) [DISJ 犬か猫] が [EXIST いる]。
(14) [NUM 2 匹の] 猫が [EXIST いる]。
(15) [NUM ちょうど 2 匹の] 猫が黒い。
(16) [UNIV すべて] の猫は白い。
(17) 猫が犬に [REL 触れ] ている。
(18) [UNIV すべての] 人が自転車に [REL 触れ] ている。
(19) 人が自転車を [REL 支え] ている。
(20) 女の子が [EXIST 何か] を [REL 持っ] ている。
(21) 猫を [REL 見] ている犬が [EXIST いる]。

5.2 実験結果

日本語 CCG パーザとして Jigg [6] と depccg [13] の 2 つを用い、意味テンプレートとしてイベント意味論に基づくテンプレート (event) と非イベント意味論に基づくテンプレート (plain) の 2 つを用いて精度比較を行った。実験結果を表 1 に示す。表 1 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。

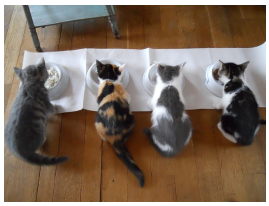
システムが証明に成功し正しく真偽を判定した例を図 4 に示す。

5.3 エラー分析

エラー分析の結果、以下の 3 つのエラーが挙げられた。1 つ目は構文解析エラーである。(13) は構文解析に失敗していた例である。

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 1: CCG パーザと意味論ごとの正答率



(a) 正解：TRUE

(b) 正解：FALSE

図 4: 少なくとも 2 匹の猫がいる。

(13)　[DISJ 犬か猫] が [EXIST いる]。

一方、(13) と同様の意味である (22) は正しく意味解析ができており、本システムも証明に成功していた。

(22)　[DISJ 猫が [EXIST いる] か、犬が [EXIST いる]]。

(13) のような名詞と名詞を選言でつないだ形の構文解析の精度向上は今後の課題とする。

2 つ目は GRIM データセットのアノテーション不備によるエラーである。GRIM データセットには「笑う」や「歩く」といった自動詞の情報が付与されていないため、以下に示す文は証明できなかった。

(23)　誰かが笑っている。

(24)　猫が歩いている。

画像とテキストの大規模データセットである Visual Genome [10] には自動詞の情報が付与されているため、Visual Genome を用いることで自動詞を含む文についても本システムを評価することが可能であるが、今後の課題とする。

3 つ目は Coq の自動証明のエラーである。例として以下の文を用いた場合、画像によって証明できるものとできないものに分かれた。

(25)　猫が犬に触れている。

(26)　人が自転車を支えている。

その理由として、GRIM の画像情報 (FOL 構造) に応じて、論理式中に出現する述語や変数などの順番が影響していると考えられる。画像推論のためのより頑健な証明システムを構築することが今後の課題である。

6　おわりに

本稿では CCG と Coq を用いた日本語マルチモーダル推論システムを提案した。イベント意味論に基づく文の意味解析を行うことで、数量表現を含む文についてシステムの正答率が上がった。また 1 階述語論理では数量表現の意味表示が複雑になるという問題があったが、Coq の型推論におけるリスト型を使うことで簡潔に表現できるようになった。

今後は日本語特有の複雑な表現を含む文でのテストを試みる。照応（「数匹の猫がいて、そのうち一匹は白い」）や比較（「タクシーより手前に人がいる」）などの現象を含む文をどのように扱うか検討したい。

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] 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.

[2] Manfred Krifka. Nominal reference, temporal constitution and thematic relations. In *Lexical Matters*, pp. 29–53. CSLI Publications, 1992.

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

[4] Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. ccg2lambda: a compositional semantics system. In *Proc. of ACL System Demonstrations*, pp. 85–90, 2016.

[5] Koji Mineshima, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. Higher-order logical inference with compositional semantics. In *Proc. of EMNLP*, pp. 2055–2061, 2015.

[6] Hiroshi Noji and Yusuke Miyao. Jigg: A framework for an easy natural language processing pipeline. In *Proc. of ACL System Demonstrations*, pp. 103–108, 2016.

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

[8] Mark Steedman. *The Syntactic Process*. MIT Press, 2000.

[9] Riko Suzuki, Hitomi Yanaka, Masashi Yoshikawa, Koji Mineshima, and Daisuke Bekki. Multimodal logical inference system for visual-textual entailment. In *Proc. of ACL-SRW*, pp. 386–392, 2019.

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

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

[12] Masashi Yoshikawa, Koji Mineshima, Hiroshi Noji, and Daisuke Bekki. Combining axiom injection and knowledge base completion for efficient natural language inference. 2019.

[13] Masashi Yoshikawa, Hiroshi Noji, and Yuji Matsumoto. A* CCG parsing with a supertag and dependency factored model. In *Proc. of ACL*, pp. 277–287, 2017.

[14] 田中リベカ, 峯島宏次, Pascual Martínez-Gómez, 宮尾祐介, 戸次大介. 日本語 CCG パーザに基づく意味解析・推論システムの提案. 言語処理学会第 22 回年次大会発表論文集, pp. 757–760, 2016.

[15] 戸次大介. 日本語文法の形式理論. くろしお出版, 2010.

[16] 鈴木莉子, 吉川将司, 谷中瞳, 峯島宏次, 戸次大介. テキスト情報と画像情報を組み合わせた論理推論システムの構築. 人工知能学会全国大会論文集, pp. 2L1J903–2L1J903, 2019.