多言語統語・意味情報コーパス Parallel Meaning Bank 日本語版の構築

谷中 瞳[1,2] 峯島 宏次[2] 山田 彬尭[3] 山口 悠[4]
窪田 悠介[5] Lasha Abzianidze[6] Johan Bos[6]
[1]理化学研究所 [2]お茶の水女子大学 [3]駿河台大学
[4]東京大学 [5]国立国語研究所 [6]University of Groningen

hitomi.yanaka@riken.jp, mineshima.koji@ocha.ac.jp, akitaka001@gmail.com,
yamaguchi.b93@gmail.com, kubota@ninjal.ac.jp, {l.abzianidze, johan.bos}@rug.nl

1 はじめに

Parallel Meaning Bank (PMB) [2] は、多言語・多ジャンルのテキストに対して、組合せ範疇文法 (Combinatory Categorial Grammar, CCG) [20, 25] に基づく統語解析情報と、談話表示理論 (Discourse Representation Theory, DRT) [11] に基づく意味解析情報を付与したコーパスである。元のコーパスは、Tatoeba[1]や新聞記事をはじめとした、12 種類の多言語コーパスを採用している。PMB は Version 2.2.0 までは英語、2 種類のゲルマン語（オランダ語・ドイツ語）、1 種類のロマンス語（イタリア語）の計 4ヶ国語を対象言語としており、西洋圏外の言語は含まれていなかった。そこで現在では、アジア言語として日本語と中国語の２ヶ国語を PMB コーパスに追加する試みが行われている。本論文では、その最初の報告として、日本語版の構築方針と進捗状況を報告する。

2 日本語 PMB のアノテーション

PMB コーパスの構築方法は、深い意味情報のアノテーションを効率的でかつ一貫した仕方で実現するため、CCG 構文解析器を軸とした自動アノテーションと人手による修正を組み合わせた設計となっている。具体的には、(1) まず英語コーパスに対してすでに付与されている様々な語彙情報を単語間のアラインメントをとることによって他の言語にマップし、(2) その上で CCG 構文解析器を利用して単語レベルから句・文レベルの統語・意味情報を自動付与する。(3) さらに、PMB アノテーションのために設計された Web インターフェイスを用いてアノテーションの各層において人手による修正を行い、自動解析モデルの再学習を行う、というブートストラップ的手法を採用している。日本語 PMB の構築においてもこの方針を採用する。

[1]https://tatoeba.org

図 1 に、PMB コーパス自動アノテーションのパイプラインを、図 2 に日本語のアノテーション例を示す。以下、アノテーションの各層の概要を説明する。

トークン化 英語のトークン化では、CCG による意味解析を考慮して、複単語表現 (MWE) が 1 つのトークンとして扱われる点に特徴がある。図 2 の例に対応する英語文では、固有表現の *The Statue of Liberty* と *New York* が一語として扱われており、日本語でもこの分割に従うことで語彙情報のマッピングが容易となる。日本語を含めて全 6 言語のトークン化には、Elephant [9] を用いる。Elephant の日本語解析用のモデルは、日本語 Universal Dependencies コーパス[2]で学習した UDpipe tokenizer[3]による PMB コーパスの自動解析結果をベースとし、それに人手による修正（後述する Bits of Wisdom, BoW）を加え、再学習したものを使用している。

単語アラインメントに基づく語彙情報付与 英語 PMB コーパスでは、各トークンに対して、品詞・固有表現情報を拡張した意味現象タグ [3]（2020 年 1 月現在で 76 種）、VerbNet [19] に基づく意味役割、WordNet [10] に基づく語義タグ (synset)、DRS に現れる述語 (symbol) の情報がアノテートされている。これらの語彙情報は単語間アラインメントに基づいて、英語から日本語を含む他の言語へ自動付与される（図 1 の Projection）[4]。Aligner は GIZA++ [16] を用いた。例えば、図 2 の例では、対応する英語文との単語間アラインメントに基づいて、「ニューヨーク」というトークンに対して意味現象タグ (sem) として GPE (geo-political entity)、DRS に現れる述語 (sym) として new~york、語義タグ (syn) として city.n.01 が付与されている。このような語彙情報と後述の CCG 統語範疇 (cat) に基づいて、意

[2]UD Japanese-GSD Version 2.3.
[3]http://ufal.mff.cuni.cz/udpipe
[4]ただし、オランダ語・ドイツ語・イタリア語の場合、意味現象タグは各言語の Semantic tagger を用いて自動付与されている。

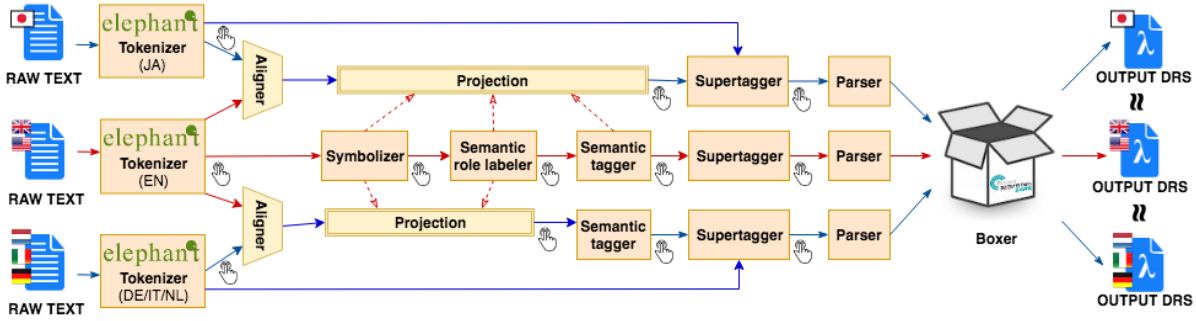

図 1: PMB コーパスのアノテーションパイプライン

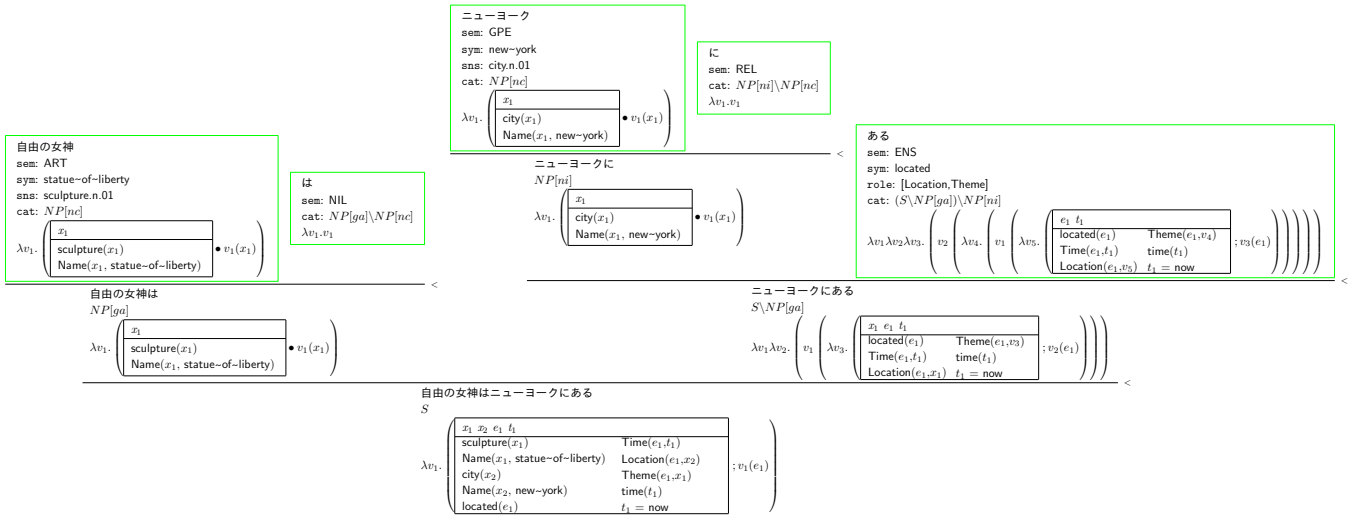

図 2: アノテーション例（id-40/2851）。終端ノードに付与された情報は以下の通り。意味現象タグ (sem): ART(artifact), NIL(empty semantics), GPE(geo-political entity), REL(relation), ENS(present simple); シンボル (sym); 語義タグ (syn); 意味役割 (role); 統語範疇 (cat). この情報に基づいて、CCG 構文解析により、文の談話表示構造 (DRS) が自動的に導出される。

味解析器 Boxer により、ラムダ項として表現される各トークンの意味表示が自動的に決定される。

CCG 構文解析　CCG は、単語の統語範疇と少数の組合せ規則に基づいて、文の統語構造（導出木）から意味表現へのマッピングが明確かつ簡潔に与えられる点に特徴がある。また、いわゆる Supertagging により、確率モデルに基づいて各トークンに統語範疇を割り当てることで、高速な構文解析が可能となる。PMB では、Supertagging と CCG 構文解析に EasyCCG [13] を用いている。日本語の場合、日本語独自のタイプ変換規則 [21] を EasyCCG に追加し、文の導出木を自動付与する。Supertagging のモデルは、depccg [23] による PMB コーパスの自動解析結果をベースとし、それに人手で修正を加え、再学習したものに基づく。現在のところ、CCG の統語情報については、各トークンの統語範疇（図 2 の終端ノードの cat タグの情報）のみが人手で修正可能であり、その変更に基づいてシステムが自動的に再解析を行い、導出木が修正される。

意味合成による談話表示構造の導出　Boxer [7] は、談話表示理論とラムダ計算に基づく構成的意味論を実装した意味解析器であり、PMB 構築のため、多言語に対応した意味割り当てに拡張されている。PMB では、上述のように単語の意味表示は自動的に割り当てられ、そこからラムダ計算に基づく意味合成によって、文の意味表示として談話表示構造 (DRS) [11] が導出される。図 2 の文全体の DRS は、この文が導入する談話指示物 (discourse referent)、すなわち、「自由の女神」と「ニューヨーク」にそれぞれ対応するエンティティ x_1 と x_2、動詞「ある」が導入するイベント e_1 と時間 t_1、及び、これら談話指示物間に成り立つ条件 (condition) の集合からなり、これにより文の意味情報・時間情報の形式的表現が与えられる。

Bits of Wisdom (BoW)　PMB コーパスの一部はすでに公開されており、PMB explorer [5]から検索・閲覧が可能である。同時に PMB explorer は、オンライ

[5]https://pmb.let.rug.nl/

	文数	トークン数	BoW
英語	334,529	3,324,139	322,996
ドイツ語	221,224	2,336,565	14,252
イタリア語	124,479	955,094	10,520
オランダ語	48,497	462,551	8,878
日本語	91,687	955,686	5,247
中国語	45,111	338,250	2,047

	言語	Gold	Silver	Bronze
トークン化	英語	22,990	5,365	274,854
	日本語	1,418	18	88,586
意味現象タグ	英語	13,039	105,818	184,352
	日本語	296	187	89,554
CCG 統語範疇	英語	10,104	4,441	288,655
	日本語	175	199	89,664

表 1: 統計情報：'BoW' は BoW によるデータ修正回数

ンで PMB のデータを編集できる Wiki 形式の Web インターフェイスとしても機能しており、自動付与されたアノテーション情報の大部分は、PMB Explorer 上で編集可能である。この人手による追加・修正情報は Bits of Wisdom (BoW) [5] と呼ばれ、その編集履歴をインターフェイスからたどることができる。BoW は、図 1 中の☝が描かれている各層において追加することが可能であり、その情報はシステムに直ちに反映され、自動解析ツールによる再解析が行われる。さらに、BoW を学習データとして各プロセスの言語解析モデルの再学習を定期的に行うことによって、自動アノテーションの精度向上を図っている。アノテーションの整合性・一貫性を保持するため、自動付与したアノテーション情報と、人手で編集したアノテーション情報との間に不整合が生じた場合は、PMB Exploer 上で conflict として表示され、アノテータが再チェックしやすい工夫が施されている。

PMB コーパスはアノテーションの質に応じて、Gold（アノテータによるチェックを経たもの）、Silver（少なくとも一つの BoW を含むもの）、Bronze（BoW を含まないもの）の 3 段階に区分されている。2019 年 10 月から日本語 PMB コーパスの自動アノテーション、及び、言語学の専門家を含むチームによるチェック作業を開始した。表 1 に 2020 年 1 月現在までの日本語 PMB の統計情報を示す[6]。

3 日本語固有の問題点

PMB プロジェクトは、当初、西洋諸語を基に設計されたため、既存の意味カテゴリーのタグセットが、日本語にきれいに対応しないケースがある。PMB を日本語に拡張する際に浮上した問題点のうち、言語学的、また類型論的に重要性が高いもの二点を報告する。

[6]この統計情報は PMB explorer から確認可能である。

課題 1 (表出的意味) 敬意表現や終助詞などの話者の態度／談話管理に関わる表現は、理論言語学においては、命題／真理条件的な意味とは別のレベルに属するものとして分析され、各単語／形態素の語彙的意味として慣習化された「表出的意味」を表すと考えられている [18, 14]。PMB においては、まだ表出的意味の構成的意味論の扱いが整備されておらず、意味タグを用意し、適切に意味表示に組み込むことができるように DRS を修正することは今後の課題である。類型論的な広がりを視野に入れた本プロジェクトでは、言語普遍性を捉えるためになるべく共通した意味タグのセットを各言語に用いたいという要請と、多様な言語のそれぞれにおいて明示的にコード化されている意味的区別をなるべくきめ細かに捉えたいという要請との間で現実的な妥協を探る必要があり、タグセットの設計には慎重な検討が必要となる。

課題 2 (呼応表現) 一般に、迂言的な表現に対して意味現象タグを付与する場合、いずれかの形態素にその意味を代表させて、残りの形態素の意味を空にするという方針が可能である。しかし、呼応表現を伴う日本語のモダリティ表現 [24] [26] には、このような方針を取りづらいものがある。例えば、「もしかすると、罠かもしれない」という事例では、「もしかすると」と「かもしれない」のどちらが認識的用法の意味の主要な要素であるのか判断が難しい (一方のみが表れる文「もしかすると罠？」「罠かもしれない」が副詞と文末表現両方が表れる文とほぼ同義である点に注意)。

4 関連する言語資源との比較

競合する言語資源は少ないが、最有力のものとして国語研 NPCMJ コーパス[7]とその関連ツール Treebank Semantics[8]がある。NPCMJ/Treebank Semantics の PMB との主な違いは、統語論と意味論のマッピングに関して Scope Control Theory (SCT) [8] と呼ばれる理論を採用している点と、PMB の BoW の付加 (2 節参照) に対応する人手での修正用のインターフェースを持たない点である。SCT という独自の理論に基づく実装を伴う Treebank Semantic に比べて、PMB は構成的 DRT の亜種の一つである λDRT [15] に依拠するため、構成的意味論の計算過程が完全に透明である。また、コーパス構築の方法論に関しても、適切な構成的意味論を単語列に付与するというタスクは非常に複雑なものであることを考えると、PMB が採用する、アノテーションの各層での人手での修正を直接自動処理の結果とシームレスに統合できる枠組みには有効性があると考えられる。

[7]http://npcmj.ninjal.ac.jp/ 2020 年 1 月時点で人手で修正した句構造のパーズ・ツリー 3 万文を公開中。

[8]http://www.compling.jp/ajb129/ts.html パイプライン処理で文の論理的意味表示を自動生成するシステム。

5 おわりに一今後の展望

自然言語処理 自然言語処理分野では、PMB の豊富なアノテーション情報を用いた応用研究が既に進められている。例えば、ニューラル含意関係認識モデルの事前学習として意味現象タグの予測を行うことで認識精度を向上させる研究 [1]、表層レベルよりも深い機械翻訳モデルの評価手法として、機械翻訳モデルが学習したベクトル表現の品質を意味現象タグの予測精度で評価する研究 [6]、PMB のアノテーション情報を用いることで推論データを効率的に自動構築し、含意関係認識モデルのデータ拡張を行う研究 [22] などがある。

既存の構文構造が付与された多言語パラレルコーパスには依存構造が付与された Universal Dependencies (UD) コーパスがあるが、PMB は構成的意味論に基づくより高度な統語情報・意味情報が付与された唯一の大規模多言語パラレルコーパスであり、UD コーパスと同様にニューラルネットに基づくモデルの学習や評価に容易に活用できるという特徴がある。近年の多言語モデルの研究の発展に伴い、今後さらに PMB コーパスの統語情報・意味情報を用いた多言語モデルの学習改善や品質評価といった応用が期待される。

言語学 UD や句構造文法のツリーバンクなどの資源は理論言語学や類型論の知見に基づくものの、多くの場合言語学的知見を一部取り込むのみであるため、言語学研究自体への成果の還元はまだ少ない。この点で、構成的意味論を明示した多言語資源である PMB には大きな将来性がある。まず第一に、ツリーバンク構築が統語論研究の実証的検証にほかならないというのと同じ理由で、構成的意味論を付与したツリーバンクの構築は形式意味論・語彙意味論研究の実証的検証にほかならないという点で、類型論的に多様な言語の言語データに明示的な構成的意味論情報を付与する作業は重要である。

また、構築したコーパスは理論中立的に、形式意味論・語彙意味論研究における理論構築の際に参照するデータベースとしての利用が期待できる。特に、複文の構成的意味論に関しては通言語的観点からの理論的研究が膨大に存在するが、[4] など最近の研究を除くとコーパスデータを積極的に活用したものは稀である。DRT は [11] などに代表されるように、時間表現の詳細な意味的分析に強みを持つ理論であるため、構成的意味論を DRT で付与した PMB は、埋め込み節のテンス解釈 [17, 12] を明示的に可視化したデータベースとしての理論研究への活用などが期待できる。

謝辞 本研究は国立国語研究所共同研究プロジェクト「構成的意味論情報を付与した多言語資源の構築(フィージビリティスタディ)」「対照言語学的観点から見た日本語の音声と文法」の助成を受けたものである。

参考文献

[1] Mostafa Abdou, Artur Kulmizev, Vinit Ravishankar, Lasha Abzianidze, and Johan Bos. What can we learn from semantic tagging? In *Proc. of EMNLP*, pp. 4881–4889, 2018.

[2] Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proc. of EACL*, pp. 242–247, 2017.

[3] Lasha Abzianidze and Johan Bos. Towards universal semantic tagging. In *Proc. of IWCS*, pp. 1–6, 2017.

[4] Daniel Altshuler. *Events, States and Times: An essay on narrative discourse in English*. de Gruyter, 2016.

[5] Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. A platform for collaborative semantic annotation. In *Proc. of EACL*, pp. 92–96, 2012.

[6] Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proc. of IJCNLP*, pp. 1–10, 2017.

[7] Johan Bos. Open-domain semantic parsing with boxer. In *Proc. of NODALIDA*, pp. 301–304, 2015.

[8] Alastair Butler. *Linguistic Expressions and Semantic Processing: A Practical Approach*. Springer, 2015.

[9] Kilian Evang, Valerio Basile, Grzegorz Chrupała, and Johan Bos. Elephant: Sequence labeling for word and sentence segmentation. In *Proc. of EMNLP*, pp. 1422–1426, 2013.

[10] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[11] Hans Kamp and Uwe Reyle. *From Discourse to Logic*. Dordrecht: Kluwer Academic Publishers, 1993.

[12] Yusuke Kubota, Jungmee Lee, Anastasia Smirnova, and Judith Tonhauser. Cross-linguistic variation in temporal adjunct clauses. In *Cahier Chronos*, Vol. 25, pp. 141–161. Rodopi, Amsterdam/Atlanta, 2012.

[13] Mike Lewis and Mark Steedman. A* CCG parsing with a supertag-factored model. In *Proc. EMNLP*, pp. 990–1000, 2014.

[14] Eric McCready. *The Dynamics of Particles*. PhD thesis, 2005.

[15] Reinhard Muskens. Combining Montague semantics and discourse representation. *Linguistics and Philosophy*, Vol. 19, No. 2, pp. 143–186, 1996.

[16] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51, 2003.

[17] Toshiyuki Ogihara. *Tense, Attitudes, and Scope*. Kluwer Academic Publishers, Dordrecht, 1996.

[18] Christopher Potts. *The Logic of Conventional Implicature*. PhD thesis, 2003.

[19] Karin Kipper Schuler. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. PhD thesis, 2005.

[20] Mark Steedman. *The Syntactic Process*. MIT Press, Cambridge, Mass., 2000.

[21] Sumire Uematsu, Takuya Matsuzaki, Hiroki Hanaoka, Yusuke Miyao, and Hideki Mima. Integrating multiple dependency corpora for inducing wide-coverage Japanese CCG resources. *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 14, No. 1, pp. 1–24, 2015.

[22] Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proc. of *SEM*, 2019.

[23] 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.

[24] 工藤浩. 副詞と文の陳述的なタイプ. 日本語の文法 3 モダリティ, pp. 161–234. 岩波出版, 東京, 2000.

[25] 戸次大介. 日本語文法の形式理論. くろしお出版, 東京, 2010.

[26] 渡辺実. 国語構文論. 塙書房, 東京, 1971.