Vol.2022-IFAT-146 No.3
Vol.2022-DC-124 No.3
2022/3/25

意味役割と概念フレームを付与したNPCMJ-PTによるタグの推定

竹内 孔一[1,a)]　岩本 潤季[2]　バトラー アラステア[3]　長崎 郁[4]　パルデシ プラシャント[5]

概要：現在，述語項構造シソーラス (Predicate-Argument Structure Thesaurus (PT)) に基づく概念フレームと意味役割を NPCMJ（NINJAL Parsed Corpus of Modern Japanese) に付与した NPCMJ-PT を構築している．NPCMJ は Web 上で公開されている日本語の構文木データであり，構文木に即して，述語項構造シソーラスの概念フレームと意味役割を付与している．意味役割とは述語と係り関係にある要素との意味的な関係を示すもので，英語圏では PropBank や FrameNet が構築されている．PropBank 形式の意味役割は AMR (Abstract Meaning Representation) といった文書の抽象的な意味構造を記述する枠組に応用されていることから，文書情報アクセスにおける利用が期待されている．NPCMJ-PT は言語学習や分析のためのデータであるが，PropBank 形式の意味役割を付与していることから工学的な応用も期待できる．本稿では，現状の付与データの内容と量について記述するとともに，一部のアノテーションデータを機械学習モデルに適用することで意味役割タグがどの程度推定できるのか評価を行ったので報告する．

Annotation Results of Semantic Role Labels and Frames in NPCMJ-PT and Estimation of Semantic Role Labels

1. はじめに

述語の意味構造を表す方法の一つとして意味役割と概念フレームが挙げられる [1], [5], [7], [8], [12]．これは述語と係り関係にある要素 (項) との組合せにより抽象的に解釈できる意味の鋳型であり，鋳型ごとに鋳型に関連する要素が同様に項として現れるという考えである．鋳型のことを概念フレームと呼び，鋳型における要素と述語との関係を意味役割とする．

例えば「走る」という述語には複数の語義が考えられるが，例えば【移動】に関する概念と【表情の出現】に関する下記の表現があったとする．すると，それぞれの概念において，他の述語でも同様の要素が出現することがわかる．

(1) [Arg0:動作主 選手が] [Arg2:経路 コースを] [移動 走る]
(2) [Arg0:動作主 バスが] [Arg2:経路 市内を] [移動 運行する]
(3) [Arg1: 対象 (身体部分) 恐怖が] [Arg2: 着点 (身体部分) 顔に] [表情の出現 走る]
(4) [Arg1: 対象 (身体部分) 疲れが] [Arg2:着点 (身体部分) 顔に] [表情の出現 滲む]

ここで例文 (1) から (4) の述語は「走る」「運行する」「走る」「滲む」であり，述語の [ ] の左下に記述しているは概念フレームを表す．また，その係り関係にある要素 (項) の [ ] には意味役割が記述されている．Arg0, Arg1 が PropBank 形式の意味役割で，動作主などが名前の意味役割である．この例にあるように【移動】を概念フレームとすると，【移動】では述語が違っていても，移動の動作主[*1]と経路に関する表現が出現する．また，【表情の出現】でも同様に，述語が異なっても同じ概念を持つ場合，ある感情的な特徴と身体が

[1] 岡山大学学術研究院自然科学学域
Graduate School of Natural Science and Technology, Okayama University, 3-1-1, Tsushima-naka, Kita-ku, Okayama, 700–8530, Japan
[2] 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University
[3] 弘前大学
Hirosaki University
[4] 名古屋大学
Nagoya University
[5] 国立国語研究所
National Institute for Japanese Language and Linguistics
[a)] takeuc-k@okayama-u.ac.jp

[*1] この場合は，動作主であると同時に移動対象である．意味役割が複数重なっている場合は PropBank の付与方針と同様に，動作主を優先して付与する [2]．意味役割の優先度順位も指定されている．

要素としてあわられる．このようなことから概念フレームを中心とした研究が盛んに行われており，PropBank[8] や FrameNet[1] といった大規模な言語資源も開発されている．また工学的な利用が期待されることから [7]，CoNLL-2005 などの競争的な国際ワークショップのタスクに採用されている [4]．

意味役割付与データの構築には係り関係の意味的な分類と概念フレーム辞書が必要であるため，開発にコストがかかる．近年では概念フレームを作成せずに係り関係のみを構文に従って解く方法が採用されている．例えば下記に示す Universal Dependency (UD) では nsubj，obj として述語と項の関係を付与している[*2]．

(5) [$_{\text{nsubj}}$ He] opened [$_{\text{obj}}$ the door].
(6) [$_{\text{nsubj}}$ The door] opened.

しかしながらこの例にも示すように，意味的に主語と目的語が入れ替わることができる表現に対して，あくまで主語に対して nsubj を付与するため自動詞表現の場合による「開けられた対象」の door と「開ける動作主」の He の違いを同定することはできない．また概念フレームがないため上記の (1) から (4) の例を同様に関係付けすることは難しい．

日本語の述語に対して意味役割と概念フレームを付与したデータとして述語項構造シソーラスが公開されている[*3]．概念フレーム辞書であるため，例文が端的な能動態に限られているため表現の幅が限られている．そこで，NPCMJ-PT では，国立国語研が公開する統語・意味解析コーパス NPCMJ[*4]に意味役割と概念フレーム付与することで概念フレーム辞書を更新するとともに PropBank や FrameNet に近いデータを構築し，言語学習者や言語分析者に解析済み日本語事例を提供することを目的としている．

以下では NPCMJ-PT の現状を記述するとともに，付与データの一部を利用して深層学習モデルである BERT+CRF を利用して意味役割の推定の実験を行った．実験結果は高い精度を得ることができなかったが，先行研究の意味役割付与精度 [11] と比較しつつより高い精度のモデル作成するための改善方法について検討する．

2. NPCMJ-PT のデータ

NPCMJ は統語解析情報が付与されたツリーバンクである．文は形態素に分割されて非終端記号が詳細に付与されている (図 1 参照)[9]．現在，99,840 文に統語情報が付与されている．テキストデータは統語情報とともに公開されており Web 上で検索が可能である．

NPCMJ に対して述語と項を自動で抽出して [3] 意味役割と述語の概念フレームを人手で付与する．付与作業の表

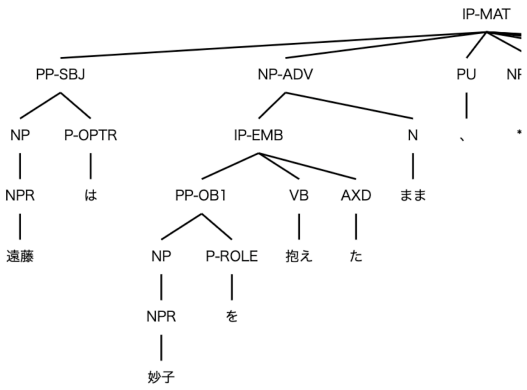

図 1 NPCMJ の統語解析済みデータの例

示部分を図 2 に示す．図 2 に示すように 1 文には複数の述

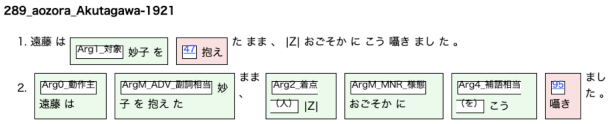

図 2 NPCMJ の文に対して意味役割と概念フレームを付与する例

語があるため，述語それぞれについて概念フレームを付与する．概念フレームは Frame ID という番号で示されており，「抱える」は 47 番，「囁く」は 95 番であることが示されている．これらの番号による概念フレームは述語項構造シソーラスで検索できる．薄緑のボックスが項を表しており，その中に付与した意味役割が提示されている．NPCMJ では省略項や痕跡が付与されており，そうした要素にも意味役割を付与している．

表 1 に NPCMJ-PT で現在付与した述語数と意味役割の数について記述する．付与した述語数 (述べ数) が付与した

表 1 付与した述語数と意味役割の量

内容	件数
付与した文数	39,523
付与した述語数 (述べ数)	75,523
付与した述語数 (種類)	12,696
付与した概念フレーム (種類)	1018
付与した項の数	151,281

概念フレームの頻度であるため，概念フレームは約 7.5 万件，意味役割を付与した項の数は約 15 万件と少なくない量を付与している．また付与した概念フレームは 1018 タイプである．述語項構造シソーラスでは現在 1097 種類の概念フレームを定義しており，そのほとんどが NPCMJ-PT に出現していることがわかる．

出現している述語の種類を頻度の多いものから順に表 2 に示す．概念フレームで最も頻度が多いのはコピュラである．8443 件あるうち，「です」が 6488 件であった．これは NPCMJ では，(7) に見るような「の」を「です」や「であ

[*2] Stanford Parser の出力結果を参考にした．
[*3] http://pth.cl.cs.okayama-u.ac.jp/ (2022/3/2 アクセス)
[*4] https://npcmj.ninjal.ac.jp/

表 2 付与した述語数と意味役割の量 (上位 10 件)

概念フレーム (Frame ID)	例	件数
コピュラ (895)	です	8443
着点への移動 (17)	行く	2515
生成 (124)	作る	2494
伝達 (95)	伝える	2218
変更 (407)	変更する	1915
絡まる (29)	絡む	1578
存在 (529)	居る	1254
認識 (101)	思う	857
行う (251)	実行する	823
決定 (333)	確定する	727

る」に相当するものと見なしており，コピュラの頻度にはこの「の」の頻度も含まれているからである．

(7) [Arg2:補語相当（は） 三百弗] [895:です の] [Arg1:対象 小切手] (aozora_Akutagawa-1921)

次に意味役割の内訳について示す．まず上位 10 件の PropBank 形式の意味役割を表 3 に示す．NPCMJ-PT の

表 3 PropBank 形式の意味役割の内訳 (上位 10 件)

意味役割のラベル	件数
Arg1	60,364
Arg0	39,565
Arg2	25,282
ArgM-ADV	6,487
ArgM-TMP	4,243
ArgM-LOC	2,242
ArgM-MNR	1,790
Arg3	1,632
ArgM-PRX	1,454
ArgM-NEG	1,385

中で現れた PropBank 形式の意味役割の種類は全部で 29 種類であった．その中で最も頻度が多く出現したのは表 3 に示すように Arg1 で，次に Arg0 である．Arg1 は対象を表しておりこれは述語が述べる変化対象などを指す．一方 Arg0 は主に動作主を表しており，述語のイベントを引き起こしている動作主体に付与されている．Arg1 から Arg5 までは必須項で，ArgM は付加詞である．ArgM の定義については文献 [10] に示している．

また名前の意味役割の上位 10 件を表 4 に示す．NPCMJ-PT では 72 種類の意味役割がに出現している．そのうち最も頻度が高い意味役割は表 4 に示すように対象であった．続いて，動作主，経験者である．名前の意味役割は PropBank 形式と異なり，役割の意味的なタイプを名前で表現しているため必須項でもどのような役割で出ているかがわかりやすい．4 番目に多い補語相当（は）は「は」で現れた項の説明にあたるもので，コピュラ文によく現れる．事例は例文 (7) に示すように「小切手」が「三百弗」(300 ドル) であったことを意味する．

また，対象については細分類の対象（生成物）などが現

表 4 名前の意味役割の内訳 (上位 10 件)

意味役割のラベル	件数
対象	50,502
動作主	29,113
経験者	11,048
補語相当（は）	8,328
副詞相当	6,487
時間	4,244
着点	3,803
場所	3,533
対象（動作）	2,834
対象（生成物）	2,606

れている．これは PropBank 形式の意味役割では必須項に分類されるもので Arg1 から Arg5 のどれかに割り振られている．

このように PropBank 形式の意味役割と名前の意味役割はそれぞれ独立に付与されているため頻度分布が異なっており，名前の意味役割の方が種類数も多く細分化された結果が示されている．

3. 推定モデル

文の中の意味役割を推定する枠組として文と文中のどの述語に対する意味役割を求めるかを入力とし，これに対して文中のトークンのどこが意味役割かを識別する．例えば入力文を S とし，そのトークン列を s_i とする．述語位置 p を教えることで，各トークンに対する識別結果のタグ列 $T = t_1, .., t_n$ を出力する．この識別モデルを表すと式 (1) のようになる．

$$\hat{T} = \arg\max_{T} P(T|S, p) \tag{1}$$

ここで，$P(T|S, p)$ は下記のように求める．

$$P(T|S, p) = P(t_1, .., t_n|s_1, .., s_n, p) \tag{2}$$

この $P(T|S, p)$ を CRF と BERT を利用して具現化する[*5]．スコア関数 c を仮定すると文献 [6] から下記のように定義する．

$$c(S, T) = \sum_{i=0}^{n} A_{t_i, t_{i+1}} + \sum_{i=1}^{n} P_{i, t_i} \tag{3}$$

式 (3) の A は状態遷移行列を表しており，付与すべきタグ列の遷移スコアを表している．ニューラルネットワークでは学習できる重みで表す．P_{i,t_i} は i 番目のトークンに対してタグ t_i を出力するスコアであり BERT の対応トークンの出力ベクトルを割り当てる．これらを利用して式 (2) は下記のように表される．

$$P(T|S, p) = \frac{e^{c(S,T)}}{\sum_{\tilde{T} \in T_{all}} e^{c(S,\tilde{T})}} \tag{4}$$

[*5] CRF-LSTM モデル [6] を参考に改変する．

ここで T_{all} は全ての可能なタグ列を表す．これを最大化するように学習データから各重みを調整する．

タグとして IOB2 形式を利用する．具体的には下記のようになる．ここでは「華やか」の述語に対する意味役割と

入力文	冬	の	星空	は
出力タグ	B-Arg1	I-Arg1	I-Arg1	I-Arg1
	明るい	星 ..	華やか	です
	O	O ..	O	O

して Arg1 を形態素毎に IOB2 形式で付与している[*6]．

入力文のトークン化とベクトル化は BERT を利用する．具体的には東北大学乾研究室の事前学習モデルを利用する[*7]．12 層の BERT でトークン化には MeCab を利用したモデルを使用する[*8]．

モデルの全体像を図 3 に示す．図 3 の例では「華やか」

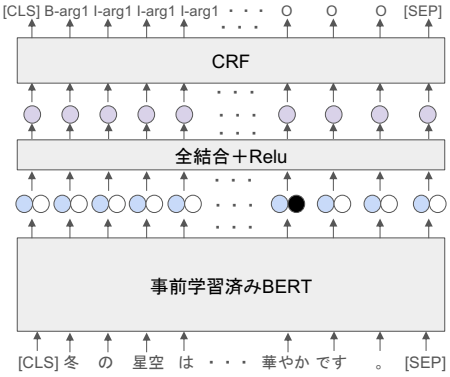

図 3 BERT と CRF を利用した意味役割付与モデル

の述語に対する意味役割を付与している．ここで図中の●がトークン「華やか」が対象の述語であることを学習する層に教えている．学習する部分として BERT の最終出力層のユニットに結合する重みと CRF 内の重みを学習する．

4. 意味役割ラベル推定実験

NPCMJ-PT の一部を利用して意味役割ラベルの付与実験を行う．入力として NPCMJ-PT に登録されている文と述語の位置を BERT+CRF に入力し，PropBank 形式の意味役割を識別させる．意味役割ラベルはトークン毎に IOB2 形式で付与する．これにより，現状，どの程度推定が可能かを評価する．下記に詳細を記述する．

*6 途中「星」の部分で文の一部を省略している．
*7 https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/ (2022/3/3 アクセス)
*8 cl-tohoku/bert-base-japanese-whole-word-masking を利用した．

4.1 実験条件

実験データは NPCMJ-PT の文から照応や痕跡を処理した 5587 文を利用する．1 文に 1 概念フレームが付与されており，照応に関する意味役割は削除している．述語は 1856 種類，概念フレームの種類は 523 種類である．意味役割は全て含まれている．

学習データ，開発データ，テストデータを 8:1:1 に分割した．各々 4469 件，559 件，559 件である．最大 50 epoch の学習として Early stopping を実施する．Early stopping は開発データの loss を参照して 8 回連続で下がらない場合に学習を停止する．batch サイズは 32 とした．

評価として項の単位で意味役割が一致するかどうかを測定する．つまり BIO チャンクが項と一致し，かつ，意味役割ラベルが一致したときのみ正解として数える．O タグはカウントしない．チャンクの適合率 (pre)，再現率 (rec)，調和平均 ($f1$) を利用する．

$$pre = \frac{\text{項のタグがチャンクと一致した数}}{\text{推定した項数 (推定したチャンク数)}} \quad (5)$$

$$rec = \frac{\text{項のタグがチャンクと一致した数}}{\text{正解の項数 (正解のチャンク数)}} \quad (6)$$

$$f1 = \frac{2 \times pre \times rec}{pre + rec} \quad (7)$$

4.2 実験結果と考察

学習を実施したところ 35 回で Early stopping が発生した．学習したパラメータを元に，テストデータに対して意味役割を出力させたところ表 5 のような結果となった．

表 5 意味役割ラベル推定結果

内容	数値
文数 (述語数)	559
推定した項数	335
正解意味役割の項数	993
一致した項数	156
適合率 (pre)	0.466
再現率 (rec)	0.157
調和平均 (f1)	0.235

表 5 に示すように適合率は 0.466 であるが，再現率が 0.157 と低い．内訳をみると項として推定した数が 335 件と少なくほとんどの項を O タグとして識別していることがわかる．一方で，一旦，項として識別した場合にはだいたい半分の確率で意味役割ラベルを推定していることがわかる．関連する先行研究 [11] では，述語項構造シソーラスの名前の意味役割を推定して 0.7 付近の精度が得られている．先行研究では項の範囲 (チャンク) を与えているため，基本的に問題の難しさが異なるが事前学習を利用しない場合において，数千件レベルの学習データで事例が十分にあれば学習できることから，他の手法で項を同定できれば意味役割ラベルの推定はより高いものなることが予測される．

正しく推定した例を示す (news_KAHOKU_37 コーパス). この例文では述語は「遅い」である．項の範囲として「〜

入力文	保護	者	から	の	提出
出力	B-arg1	I-arg1	I-arg1	I-arg1	I-arg1
	物	が	遅い	、	と ..
	I-arg1	I-arg1	O	O	O ..

が」までが正解であるが，正しく獲得できている．多く見られた誤りでは O タグが出力されて意味役割が付与されていない事例が見られた．

こうした結果に対する今後の改善点について考察する．まず出力結果を調べて気がついたこととして項の位置がずれる誤りがあまり起こっておらず，ほとんどが項として出力されるかが問題である点である．4 千事例ほどで様々な項の表現があるなかで項を取り出す部分が学習できているのは想定外であった．確実な場合のみ IOB タグを出力している．問題は再現率が低い点である．ある述語に対する項の位置のみを学習する方法を NPCMJ-PT のアノテーションデータ以外からも獲得すると精度が向上する可能性が考えられる．意味役割のタグは異なるが項やチャンクを付与した言語資源は他にも存在するため (例えば EDR 電子化辞書など [13]) こうしたデータの利用の検討も考えられる．また BERT 自体も 12 層ではなくより大きいモデルを利用した場合の学習も検討する必要がある．

今回は小さい学習データで時間を節約することで学習結果からまずどのような状況が起きるかを確認したが，今後は 7 万件の学習データを適用した場合の結果を明らかにしたい．今回の BERT+CRF モデルでは述語の概念フレーム予測は行わなかった．理由としては予備実験で精度が低いこと (1000 以上の可能性があるため)，また各述語には概念フレームはほとんど 1 つであり，辞書に基づいた制約を入れることが必須であるため，CRF の枠組とは合わないと考えたためである．今後こうた部分についても検討する予定である．

5. まとめ

本稿では国立国語研究所で開発している統語・意味解析コーパス NPCMJ（NINJAL Parsed Corpus of Modern Japanese) に述語項構造シソーラス (Predicate-Argument Structure Thesaurus (PT)) に基づく概念フレームと意味役割を付与した NPCMJ-PT を構築し，その内容について記述した[*9]．現段階で約 7.5 万件の概念フレーム，約 15 万件の項について意味役割ラベルの付与を行った．意味役割ラベルとして PropBank 形式のものと名前の意味役割を同時に付与している．

さらに，NPCMJ-PT の一部の学習データを利用して BERT+CRF モデルによる意味役割タグの推定実験を行った．推定実験の結果 f1 スコアは 0.235 と低いことがわかった．この原因として適合率は 0.466 であるのに対して再現率が 0.157 と大きく下げており，項としての識別の部分でまず問題があることが明らかになった．今回は学習時間を短くするために学習データを小さくして適用し，識別の特徴を観察した．今後，全学習データを利用する際に，再現率を上げる手法を適用する必要があることが予測される．

謝辞 本研究は国立国語研究所の共同研究プロジェクト「NPCMJ コーパスに対する日本語意味役割分析」および科研費 (課題番号 15H03210) と (課題番号 19K00552) の助成を受けたものである．

[*9] http://www.compling.jp/dev/ (2022/3/3 アクセス)

参考文献

[1] Baker, C. F., Fillmore, C. J. and Lowe, J. B.: The Berkeley FrameNet project, *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pp. 86–90 (1998).
[2] Bonial, C., Babko-Malaya, O., Choi, J. D., Hwang, J. and Palmer, M.: PropBank Annotation Guidelines Version 3.0, *Proceedings of the Joint Conference on EMNLP and CoNLL: Shared Task*, pp. 1–40 (2012).
[3] Butler, A.: Treebank Semantics, Technical report, Hirosaki University (2021). (`https://entrees.github.io/index.html` accessed 2022/1/12).
[4] Carreras, X. and Màrquez, L.: Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling, *Proceedings of the CoNLL-2005 Shared Task* (2005).
[5] Fillmore, C. J. and Baker, C. F.: Frame Semantics for Text Understanding, *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL* (2001).
[6] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C.: Neural Architectures for Named Entity Recognition, *arXiv:1603.01360* (2016).
[7] Màrquez, L., Carreras, X., Litkowski, K. C. and Stevenson, S.: Semantic Role Labeling: An Introduction to the Special Issue, *Computational Linguistics*, Vol. 34, No. 2, pp. 145–159 (2008).
[8] Palmer, M., Gildea, D. and Kingsbury, P.: The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics*, Vol. 31, No. 1, pp. 71–105 (2005).
[9] 吉本 啓，周 振，小菅智也，大友瑠璃子，Butler, A.：日本語ツリーバンクのアノテーション方針，言語処理学会第 19 回年次大会 (2013).
[10] 竹内孔一：NPCMJ に対する PropBank 形式の意味役割と名前の意味役割付与における考察，情報処理学会研究報告, 2020-IFAT-138(4)，pp. 1–6 (2020).
[11] 岡村拓哉，竹内孔一，石原靖弘：ニューラルネットワークを利用した日本語意味役割付与モデルの構築，情報処理学会論文誌，Vol. 60, pp. 2063–2074 (2019).
[12] 深田 智，仲本康一郎：概念化と意味の世界，研究社 (2008).
[13] 日本電子化辞書研究所：EDR 電子化辞書使用説明書 (第 2 版) (1995).