Transformer デコーダモデルを利用した日本語意味役割において，特徴量抽出位置および Attention Mask の形状が与える影響

曽和晃太郎　竹内孔一
岡山大学大学院 環境生命自然科学研究科
pty01m90@s.okayama-u.ac.jp, takeuc-k@okayama-u.ac.jp

概要

2022 年の ChatGPT 登場以降，多くの大規模言語モデルが提供されているが，それらの多くが Transformer のデコーダモデルを採用している．一方，意味役割付与タスクでは依然として Transformer エンコーダモデルが主導している現状がある．本研究では，Transformer デコーダモデルを利用した日本語意味役割付与タスクにおいて，精度向上に寄与する特徴量の位置および Attention Mask の形状を検討し，ベースラインモデルとの比較実験を行った．EOS，LA，PS，PS-Attn. の 4 つの手法を提案し，PS と PS-Attn. でベースラインモデルと比較して F1 値における大幅な精度の向上を示した．

1　はじめに

2018 年に Transformer が登場して以来，自然言語処理は飛躍的な進歩を遂げてきた．この Transformer は，エンコーダとデコーダの二つの構造を有しており，それぞれ異なるタスクで活用されている．エンコーダを基盤としたモデルは，感情分析 [1] や名前付きエンティティ認識（NER）[2] といった分類タスクに広く利用されている一方，デコーダを基盤とするモデルは，対話システム [3] や QA システム [4] などの生成タスクにおいて活用されている．

2022 年には，デコーダモデルを基盤とする対話システム「ChatGPT」が登場し，それを契機に大規模言語モデル（LLM: Large Language Model）の開発が加速した．これらのモデルの多くは，GPT-NeoX や Llama などのデコーダアーキテクチャを採用しており，対話や QA システムのみならず，画像生成 [5] や RAG（Retrieval-Augmented Generation）[6] といった様々な応用分野でも成果を上げている．

一方，意味役割付与（SRL: Semantic Role Labeling）の領域では，依然としてエンコーダモデルが主要な役割を果たしている．意味役割付与とは，文中の述語を中心に，「いつ」「どこで」「誰が」「何を」といった関連性のあるトークン列を推測し，それに適切なラベルを付与するタスクである．この分野では，BERT[7, 8] や RoBERTa[9] など，エンコーダモデルを用いた先行研究が数多く存在する．現在，豊富な知識を内包する LLM を活用することで，より高精度な意味役割付与が期待されている．本研究では，デコーダモデルを用いた意味役割付与の精度向上を目指し，特徴量の位置および Attention Mask の形状が与える影響について検討する．

2　CRF と BIO tagging による SRL

本研究では SRL を系列ラベリング問題と捉え，トークン列全体に尤もらしいラベルを付与することを考える．そこで，本章では CRF(Conditional Random Field)[10] と BIO tagging を用いた系列ラベリングとしての日本語意味役割について説明する．

2.1　CRF

CRF は，与えられた系列データに基づいてラベル列の条件付確率を計算し，それを最大化することで，最適なラベル列を予測するモデルである．入力を $\mathbf{X} = \{x_0, x_1, ..., x_T\}$，ラベル列を $\mathbf{Y} = \{y_0, y_1, ..., y_T\}$ とした場合，条件付確率は以下のように表される．

$$p(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(X)} \exp\left(\sum_{t=0}^{T}\sum_{k} \lambda_k f_k(y_{t-1}, y_t, \mathbf{X})\right) \quad (1)$$

ここで λ_k, f_k はそれぞれ重みパラメータと特徴関数を表す．また，$Z(X)$ は正規化のために導入される定数であり，すべての可能なラベル系列 $\mathbf{Y}'$ の和として以下のように表される．

$$Z(\mathbf{X}) = \sum_{y' \in \mathbf{Y}'} \exp\left(\sum_{t=0}^{T}\sum_{k} \lambda_k f_k(y'_{t-1}, y'_t, \mathbf{X})\right) \quad (2)$$

損失関数は式 1 に対して負の対数尤度を用いて以下のように計算される．

$$\mathcal{L} = -\sum_{i=0}^{N} \log P\left(\mathbf{Y}^{(\mathbf{i})} | \mathbf{X}^{(\mathbf{i})}\right) \tag{3}$$

ここで入力系列 **X** は LLM の最終層の隠れ状態であることに注意されたい．

2.2 BIO tagging を用いたスパン推定

ラベル推定において，複数のトークンにラベルを適用する必要がある場合，その範囲を表現する手法として BIO tagging を用いる．BIO tagging では，ラベルとタグを組み合わせた形式で表記し，例えば **B-Arg0** はラベル Arg0 の開始を示す．ここで，**I-Arg0** から **I-Arg1** への遷移のように，異なるラベル間での I タグの直接的な遷移は禁止される．これにより，ラベルの一貫性を維持しつつ，明確な範囲指定が可能となる．

2.3 BIO 拡張による Byte Fallback 対応

本研究で使用する LLM のトークナイズアルゴリズムは Byte-level BPE である．このアルゴリズムでは，対象のトークンが語彙に存在しない場合，そのトークンをバイト列として表現する Byte Fallback が発生する．Byte Fallback が発生しているトークンに対してトークン ID への変換を行った場合，一つのトークンから複数のトークン ID が生成されるという問題が生じる．これにより，入力のトークン ID 列と BIO タグ系列のサイズが一致しなくなり，システムが正常に動作しなくなる可能性がある．そこで本研究では前処理段階において BIO 拡張という処理を行う．これによって複数のトークン ID を出力するトークンの位置情報を保存し，BIO タグ系列の対応した位置に I タグもしくは O タグを増加した分だけ挿入することでトークン ID 列と BIO タグ系列のサイズを同一に保つことができる．

3 提案手法

3.1 ベースラインと比較モデル

今回使用するモデルのネットワークを図 1 に示す．Unidic に合わせて区切ったテキストをトークン ID に変換したのちに LLM へと渡し，得られた特徴量を CRF 層へと渡す．学習層は LLM の最終層 1 層のみとする．

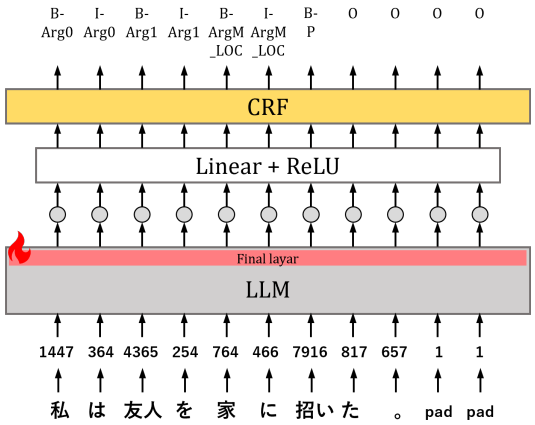


図 1: ネットワークの概観

3.1.1 ベースラインモデル

図 2(a) に本研究で使用するベースラインモデルを示す．ベースラインモデルでは，LLM から得られた特徴量をそのまま CRF 層へと渡す．また，LLM に渡される Attention Mask は各シーケンスの長さとする．

本研究ではこのベースラインモデルを基準として以下に示す 4 つの手法との比較を行う．

- EOS: End Of String
- LA: Last_hidden_state Average
- PS: Predicate Start
- PS-Attn.: Predicate Start Attention

3.1.2 EOS モデル

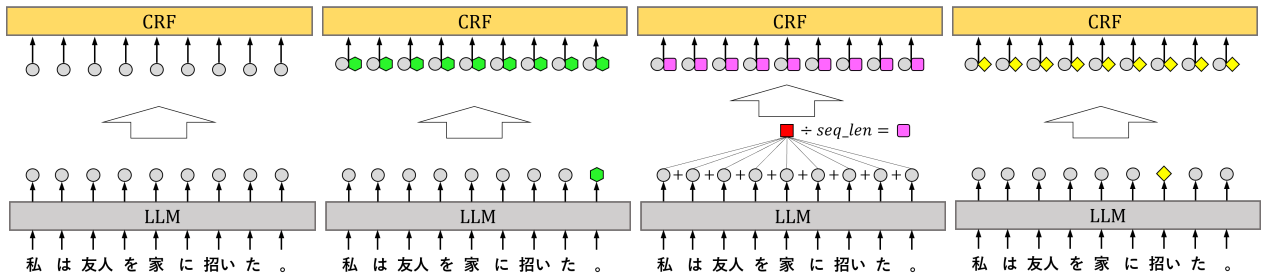
図 2(b) に示す EOS モデルでは，ベースラインモデルで使用した特徴量と，文末のトークンの特徴量を取り出し，結合したものを CRF 層に渡す．Attention Mask はベースラインモデルと同じシーケンス長とする．

3.1.3 LA モデル

図 2(c) に示す LA モデルでは，ベースラインモデルで使用した特徴量と，全トークンの特徴量を合計したものをシーケンスの長さで除したものを CRF 層に渡す．Attention Mask はベースラインモデルと同じシーケンス長とする．

3.1.4 PS モデル

図 2(d) に示す PS モデルでは述語の先頭のトークンとベースラインモデル使用した特徴量を結合したものを CRF 層へと渡す．Attention Mask はベースラ



(a) ベースラインモデル (b) EOS モデル (c) LA モデル (d) PS モデル

図 2: 各手法における CRF 層へ渡す特徴量

インモデルと同じシーケンス長とする．

3.1.5 PS-Attn.

私	は	友人	を	家	に	招い	た	。
1	1	1	1	1	1	1	0	0

図 3: PS-Attn.

PS-Attn. モデルでは図 3 で示すように，Attention Mask を文頭トークンから述語の開始位置までに制限したものを LLM へ渡す．取り出す特徴量はベースラインモデルと同じ最終層の隠れ状態のみとする．

4 実験・結果

本実験では以下に示すモデルを使用する．

- OpenCALM-7b
- CALM2-7b

OpenCALM は GPT-NeoX をベースアーキテクチャとしており．CALM2 は Llama2 をベースアーキテクチャとしている．GPT-NeoX，Llama2 はともに Transformer のデコーダベースの構造を持つ．

4.1 データセット

本実験では国立国語研究所から公開されている日本語コーパスである NPCMJ に対して，Propbank 形式の意味役割ラベルを人手で付与した NPCMJ-PT[11] を用いる．NPCMJ-PT のデータ 54,127 件を学習，検証，テストに 8:1:1 の割合で使用する．

4.2 評価方法

評価方法として，式 (4)(5)(6) で示される Precision，Recall，F1 値を用いる．ここで，s は項の範囲，l は項のラベルを示し，(s, l) で示された範囲とラベルのペアに対して，予測ペア集合と正解ペア集合の両方に存在するペア，すなわち項および意味役割ラベルが正解と一致した数を用いて評価する．

$$\text{Precision} = \frac{\left|\{(s,l) \in \text{予測} \wedge (s,l) \in \text{正解}\}\right|}{\left|\text{予測}\right|} \times 100 \quad (4)$$

$$\text{Recall} = \frac{\left|\{(s,l) \in \text{予測} \wedge (s,l) \in \text{正解}\}\right|}{\left|\text{正解}\right|} \times 100 \quad (5)$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

また，予測系列をすべて O タグで返す，すなわちラベルの予測を行わなかったものの数を All-O-tags として定義する．All-O-tags は直接的にモデルの性能を評価することはできないが，モデルが述語とその対応した項に注目を向けられたかどうかを測る．値が小さいほど予測すべき系列に注目できていることを示す．

4.3 実験結果

表 1 に本実験の結果を示す．OpenCALM，CALM2 の両方において PS が最も高い F1 値を得た．また，それぞれのモデルでベースラインモデルよりも PS で 22 ポイント以上，PS-Attn. では OpenCALM で 4 ポイント程度，CALM2 では 20 ポイント以上の精度の向上が見られた．一方で，EOS ではベースラインモデルよりも精度が低下した．

5 考察

4.3 節で示した実験結果の要因にデータセットの形式が考えられる．今回使用した NPCMJ-PT はデータ一つに対し，述語一つとなっている．そのため，一つの文章に対し複数の述語が含まれている場合，その述語の数だけデータが存在することとなる．したがって，同一の文章に対し別の意味役割ラベルを与える学習を行った場合，余分な情報を CRF 層へ

表 1: 各手法の結果とベースラインモデルとの差分

model	method	F1	Δ	Precision	Δ	Recall	Δ	All-O-tags
OpenCALM-7B	base	39.19	-	39.53	-	39.42	-	1942 / 5408
	eos	34.43	-4.76	35.03	-4.50	34.34	-5.08	2247 / 5408
	la	38.30	-0.89	39.10	-0.43	38.21	-1.21	1831 / 5408
	ps	**63.46**	**+24.27**	**65.83**	**+26.30**	**62.95**	**+23.53**	0 / 5408
	ps-attn.	43.01	+3.82	43.81	+4.28	43.33	+3.91	1187 / 5408
CALM2-7B	base	27.79	-	28.40	-	27.59	-	2567 / 5408
	eos	26.91	-0.88	27.12	-1.28	27.25	-0.34	723 / 5408
	la	40.95	+13.16	41.54	+13.14	40.98	+13.39	1765 / 5408
	ps	**51.20**	**+23.41**	**53.45**	**+25.05**	**50.47**	**+22.88**	52 / 5408
	ps-attn.	47.98	+20.19	48.86	+20.46	48.62	+21.03	846 / 5408

渡している可能性が考えられる．EOS モデルなどでは文章の前半に登場する意味役割に対して，文末のトークンの特徴量を参考に予測せざるを得ず，注目すべき項以外の情報が格納された結果，O タグを出力するような学習を行ってしまう．ここで，表 1 の All-O-tags では，PS が極端に低い値を示している．これは，明示的に述語位置を与えることで，考えるべき項の位置を指示している．

また，Transformer のデコーダの構造と日本語の語順も影響していると考える．Transformer のデコーダモデルは，直前までのトークン列を入力として，事前学習で得た確率分布に基づき次のトークンを生成する．この生成プロセスによって得られるトークンの特徴量はそのトークンまでの情報を格納している．日本語は SOV 型（主語-目的語-述語）の語順を持ち，述語が登場する時点で，多くの意味役割を持つ項がすでに出揃っている．

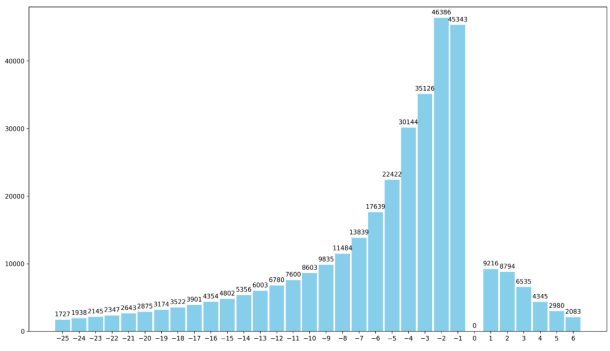


図 4: 述語と項の相対位置の分布

今回使用したデータセット 54,127 件に対して，述語と項の相対位置の分布を示すグラフを図 4 に示す．述語の位置を 0 として，項が登場する相対位置を示している．視認性の観点から登場回数が全体の 3%(1,624 回) 以下の分布は削除している．この図より，80%以上の項が述語の手前で登場し，述語の後ろに項が登場するものは最大でも 9,216 件で全体の 20%に満たない．すなわち述語の先頭のトークンの特徴量を取得することで，対応する項の情報がおおよそ得られていると考える．したがって，EOS モデルや LA モデルのように文章の全体の情報を取得する必要はなく，述語の周辺に LLM の注意を向けることが精度向上に寄与したのではないかと考える．

6 おわりに

本研究では，日本語意味役割付与において，有効な特徴量の位置および Attention Mask の形状について検討し，実験を行った．ベースラインモデルと比較して，述語の開始位置の特徴量を結合した PS モデルでは OpenCALM，CALM2 の両方で 20 ポイント以上の精度の向上が見られた．また，Attention Mask を述語の開始位置までに限定する PS-Attn. でも OpenCALM，CALM2 の両方で精度が向上した．一方で，文末のトークンを結合する EOS モデルと最終層の隠れ状態を平均して結合する LA モデルでは精度の向上はみられなかった．これらの結果より，文全体の特徴よりも述語位置周辺の情報が日本語意味役割付与では精度向上につながることがわかった．

今後は意味役割の登場する分布を事前分布として扱うことで高精度の意味役割付与を行うことを考えている．また，述語の大まかな意味を推測する概念フレーム付与タスクと日本語意味役割付与タスクにおける，高精度のマルチタスク学習モデルの構築を目指す．

参考文献

[1] Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. Aspect-based sentiment analysis using BERT. In Mareike Hartmann and Barbara Plank, editors, **Proceedings of the 22nd Nordic Conference on Computational Linguistics**, pp. 187–196, Turku, Finland, September–October 2019. Linköping University Electronic Press.

[2] Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. MELM: Data augmentation with masked entity language modeling for low-resource NER. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2251–2262, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[3] Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. Evaluating attribution in dialogue systems: The BEGIN benchmark. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 1066–1083, 2022.

[4] Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. Self-prompting large language models for zero-shot open-domain QA. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 296–310, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[5] Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. DALLE-2 is seeing double: Flaws in word-to-concept mapping in Text2Image models. In Jasmijn Bastings, Yonatan Belinkov, Yanai Elazar, Dieuwke Hupkes, Naomi Saphra, and Sarah Wiegreffe, editors, **Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP**, pp. 335–345, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[6] Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. RULE: Reliable multimodal RAG for factuality in medical vision language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 1081–1093, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[7] Wei Liu, Songlin Yang, and Kewei Tu. Structured mean-field variational inference for higher-order span-based semantic role labeling. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 918–931, Toronto, Canada, July 2023. Association for Computational Linguistics.

[8] Yu Zhang, Qingrong Xia, Shilin Zhou, Yong Jiang, Guohong Fu, and Min Zhang. Semantic role labeling as dependency parsing: Exploring latent tree structures inside arguments. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 4212–4227, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[9] Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 549–559, Online, August 2021. Association for Computational Linguistics.

[10] Charles Sutton and Andrew McCallum. An introduction to conditional random fields, 2010.

[11] Koichi Takeuchi, Alastair Butler, Iku Nagasaki, Takuya Okamura, and Prashant Pardeshi. Constructing Web-Accessible Semantic Role Labels and Frames for Japanese as Additions to the NPCMJ Parsed Corpus. In **Proceedings of The 12th Language Resources and Evaluation Conference (LREC2020)**, 2020.