JDD-PAS: 規範的な日本語日常対話コーパスへの意味役割ラベル・述語項構造付与

[1,2,3] 吉野 幸一郎, [3,2] 李 相明, [2] 波部 英子, [4] 大村 舞,
[4] 浅原 正幸, [5] 若狭 絢, [5,6] 赤間 怜奈, [5,6] 鈴木 潤
[1] 東京科学大学 情報理工学院, [2] 理化学研究所 GRP, [3] 奈良先端科学技術大学院大学
[4] 国立国語研究所, [5] 東北大学 言語 AI 研究センター, [6] 理化学研究所 AIP
koichiro@c.titech.ac.jp, masayu-a@ninjal.ac.jp, akama@tohoku.ac.jp

概要

人間同士の対話・コミュニケーションのモデル化する上で、対話中にやりとりされた発話が持つ構文・意味構造や談話構造は重要な役割を持つ。本研究はこうした対話分析に活用可能で、また既存の意味役割解析が対話においてどの程度利活用可能かを測るテストベッドを作成する目的で、日本語日常対話コーパスへの意味役割ラベル・述語項構造のアノテーションを行った。構築したアノテーションスキーマ・フレームワークを用いることで、1 人月あたり 300 対話程度のアノテーションが可能であり、今後の対話研究に本枠組みを用いることが出来そうな見通しが立ったことを報告する。

1 はじめに

人間同士の対話・コミュニケーションのモデル化は自然言語処理における重要な研究目標の一つである。自然言語発話においてやりとりされる内容と、それらの内容に対する話者の理解（心内世界との接地）や外界と紐づき（外界世界との接地）を明らかにすることは、こうしたモデル化の一助となる。意味役割ラベル（Semantic Role Labeling; SRL）[1, 2, 3]、述語項構造（Predicate-Argument Structure; PAS）[4, 5]、抽象的意味表現（Abstract Meaning Representation; AMR）[6] などの表現は、文が持つ意味構造を明らかにしようとするもので、対話中の発話理解としてもこれまで様々な取り組みが行われてきた [7, 8, 9][1)]。

これら文中の意味構造を明らかにする方法論は、主として新聞などの書き言葉を対象に議論されてきた [10, 11]。日本語においては現代日本語書き言葉均衡コーパス（BCCWJ）[12] に対する述語項構造アノテーション [13, 14] の一部にブログなど話し言葉調の発話へのアノテーションの試みがある。また、著者らは実世界参照・照応が生じる対話を対象に述語項構造に加えて外界との参照関係を付与した対話データセットのアノテーションを行った [15]。ただし、対話を対象として大規模に述語項構造を付与したデータは存在しない。

これは話し言葉、自然対話に対する意味構造のアノテーションが容易でないことが大きい。自然対話においてはしばしば実世界の事物に対する外界照応が生じ、アノテーション負荷が非常に大きくなる [11]。また、対話中の話し言葉には項だけでなく述語の省略や順序の入れ替わり、要素の言い直しなど、書き言葉においては記述時に内省・校正が行われる部分がそのまま表れ、アノテーションを複雑にする。実際に述語項構造解析を用いた対話システムにおいては、こうした点が大きな問題として報告されてきた [16]。また、対話においては実際には発話レベルだけでなく談話レベルでの発話間関係も重要である [17]。こうした対話に含まれる発話内の局所的構造と発話間の大域的構造を統一的に扱ったデータセットの構築と分析が求められる。

これらの問題に対処するため、今回の研究では日本語日常対話コーパス（Japanese Daily Dialogue; JDD）[18][2)] を対象とした述語項構造を中心とするアノテーションを行う。JDD では可能な限り規範的な表現形式の対話が収録されており、その作成プロセス上で対話中に生じうる言い直し、言い淀みや順序の入れ替わり、項・述語の省略が可能な限り修正・

1） 意味役割ラベルと述語項構造は先行研究によってしばしば同一視される。ただし原義としての意味役割ラベルは、述語に対して係り受け関係を持つ格要素候補が持つ意味役割を識別するものである。本研究では係り受け関係にない文内外の格要素候補と述語の関係をゼロ照応によって扱う場合を述語項構造として区別する。

2） https://github.com/jqk09a/japanese-daily-dialogue: 本拡張は JDD の Web サイトで公開予定

表 1　JDD Topic 3（旅行）における対話例

ID	発話
A1:	おはようございます。高原の朝は冷えますね。
B1:	おはようございます。本当ですね。羽織るものが欲しいです。
A2:	朝食の前に散歩でもいかがですか？
B2:	良いですね。どこを歩きましょうか？
A3:	湖の周りを歩きましょう。林道の先に湖があるそうですよ。
B3:	それでは、湖を一周しましょう。
A4:	一周するのにどのくらい時間がかかるでしょうか？
B4:	ロッジのオーナーに聞いてみましょう。

補完されている。こうしたコーパスを対象とすることでアノテーションの負荷を軽減し、ある程度規模の大きいアノテーションを構築することを目指す。また、対話中の発話間に含まれる文間の関係を述語と格要素の文間ゼロ照応 [19] としてアノテーションする。これにより、対話全体における構文・意味構造の現れ方を明らかにし、また既存の述語項構造解析を対話に適用する場合の問題点を明らかにしようとする。

今回は JDD に対する述語項構造のアノテーションスキーマ・フレームワークの構築を行い、全 5,261 対話のうち約 6%程度にあたる 309 対話へのアノテーションを試行した結果について報告する。実際には約 600 対話までアノテーションが終了している[3)]。また現行のマニュアルおよび KWJA[20] を用いた自動解析の併用により、300 対話あたり 1 人月程度でアノテーション可能であることを確認した。

2　述語項構造のアノテーション

意味役割ラベル・述語項構造のアノテーションを行うにあたり、これまで著者らが実世界との接地を想定して構築してきた対話データセット（J-CRe3）で構築してきたアノテーションスキーマ [15] を用いる。J-CRe3 は対話参与者の一人称視点での動画像がある状態での対話に対するアノテーションを行ったもので、フレーズグラウンディング [21] と述語項構造を統合的に扱い実世界での対話における意味を取り扱うことを指向している。

今回対象とする JDD は疑似対話のデータセットであり、実世界での参照先が存在しない。その代わりに、発話中で参照される様々な事物の参照先や意味が明確となるよう、規範的な表現へ統制が行われている。JDD における対話の例を表 1 に示す。この例の発話 A4 は「(一周するのに）どのくらい（時間が）かかるでしょうか？」の括弧内のように、実際の発話で省略されうる内容ができる限り丁寧に補足されている。こうした性質は対話に対するアノテーションの負荷を大きく軽減することが期待される。

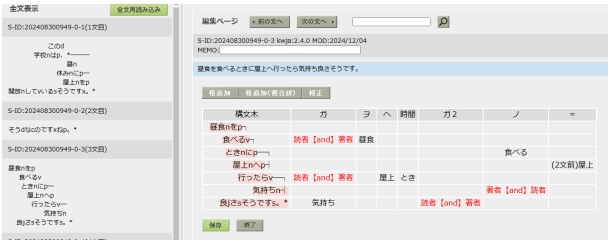

図 1　アノテーション画面

2.1　基礎アノテーションとの整合

JDD にはこれまでに既にいくつかの基礎アノテーションが施されている [22]。具体的には、形態論情報として UniDic に基づく短単位・長単位形態論情報、構文情報として文節境界、文節係り受け情報、および単語単位の Universal Dependencies [23] が付与されている。今回我々がベースとする J-CRe3 のアノテーションスキーマは文節単位の京大コーパスに基づくため、これらのうち文節係り受け情報を用いて、文節単位での述語項構造情報を付与していく。具体的には、対話に対する KWJA[20] の自動解析結果と文節単位での係り受け情報を突合し、係り受け情報のアノテーション情報を優先しながら京大コーパスアノテーションツールキット[4)] 上での重ね合わせを行い、これを修正する形でアノテーションを実施する。図 1 に、実際に KWJA の自動解析結果を重ね合わせたアノテーション画面を示す。

2.2　今回取り扱う範囲

今回、アノテーションの対象とする述語は述語および事態性名詞とする。述語は動詞だけでなくイ形容詞、ナ形容詞、判定詞によって用言化する場合を含む。事態性名詞は末尾に「する」「なる」を付与して動詞化する名詞句を認定対象とする。

格は、ガ、ヲ、ニ、デ、ト、ヘ、カラ、ヨリ、マデ、ガ 2 を付与する。時間表現については、KWJA の自動解析結果には含まれるが修正の対象としな

3）　2025 年 1 月 10 日現在。

4）　https://github.com/ku-nlp/KyotoCorpusAnnotationTool

い。また、意味役割としては表層格に限定して付与し、原則として深層格、格交替は扱わない。例えば、「(コーヒーは) ブラックでお願いします。」のような場合、「ブラック」は「お願いします」のデ格として付与し、ヲ格には変換しない。また、ガ格に関してのみ必要に応じてガ 2 格を用いる。例えば「(聞き手は) このジャケットはどうでしょうか？」という発話の場合、「どうでしょうか」に対して「ジャケット」にガ格を、「聞き手」にガ 2 格を付与する。

同じ助詞を持つが役割が異なる要素が複数存在する場合は、表層の助詞と異なる格を付与することがある。例えば「昼食に屋上に行く。」の場合、「行く」に対して「昼食」をニ格として、「屋上」をヘ格として付与する。ただし、複数の事物が同じ意味役割を持つ場合、例えば図 1 の「著者」と「読者」のようなケースは [and] によってこれらを結んで付与する。この [and] は後述する照応の関係には用いない。

2.3 対話内外への照応

JDD では、対話の自然さを損なわない範囲で対話中の省略がある程度補われている。ただしそれでも、発話の意味を正しく理解するために対話履歴中の発話内容を参照する必要がある場合が存在する。例えば表 1 中 A4 の「一周する」に対して、実際には一文前 (B3) の「湖を」が省略されている。このとき、アノテーションの上では A4 中の「一周する」に対してヲ格として「(1 文前) 湖」のように文脈から補われる格要素を付与する。この補完は同じ対話の文脈全てを対象として行う。補完候補が複数存在する場合、述語が生じた箇所から最も近い候補を格要素として付与する。

これまでの述語項構造アノテーションの研究の多くでは、直接係り受け関係にない述語と格要素の間に関係を認定する場合、述語に直接係るゼロ代名詞が存在するとみなして、ゼロ代名詞と格要素候補の間に照応関係を認定してきた [19]。今回はこのように係り受け関係に捉われることなく、それまでの対話履歴に存在するあらゆる格要素候補から述語への関係を直接認定する。ただしこれまでの慣例にならい、これらのうち直接係り受けの関係にないが文内にあるものを文内ゼロ、所属する文が異なるものの関係を文間ゼロと呼ぶ。

照応の問題を厳密に考えようとする場合、照応関係にあるもの同士の実体が同一 (＝の関係) か異なる (≒の関係) かが問題となる。例えば、「兄が本を買った。私はそれを借りた。」と「兄が本を買った。私もそれを買った。」の例では、「本」と「それ」は前者が＝、後者が≒の関係になる。JDD においては対話の場面として想定される実世界の文脈が存在しないため、＝と≒の区別に実用上の意義がない。そこで今回のアノテーションにおいては＝と≒の関係を同一視し、照応関係の認定は後述する「話し手・聞き手」を除いて行わない[5]。それぞれの述語の格要素としては発話中に現れた代表表記を付与する。ただし、一般常識を話す場合など、特定の事物ではなく総称としての人などに言及する場合は、外界照応先として「不特定：人」などを付与する。

表 2 JDD-SRL の Topic 1 における格情報の統計

格	係受	ゼロ (文内)	ゼロ (文間)	外界	合計
ガ	1,338	966	781	2,479	5,564
ヲ	756	604	387	53	1,800
ニ	461	344	162	573	1,540
デ	177	146	16	0	339
ト	175	50	23	22	270
ヘ	28	26	16	1	71
カラ	54	20	7	0	81
ヨリ	6	2	0	0	8
マデ	14	17	3	0	34
ガ 2	143	105	49	612	909

表 3 話し手・聞き手に着目した統計

格	話し手	聞き手	合計
ガ	1,643	767	2,410
ヲ	4	14	18
ニ	63	455	518
デ	0	0	0
ト	11	6	17
ヘ	0	0	0
カラ	0	0	0
ヨリ	0	0	0
マデ	0	0	0
ガ 2	441	146	587

2.4 話し手・聞き手

対話が書き言葉に対して特徴的なのは、対話参与者である話し手と聞き手が発話の中では明示されずに格要素の候補となりうる点である。この点に関しては我々は J-CRe3 のアノテーションにおいて、著者・読者 [24] を拡張した話し手・聞き手の外界照応を設定しており、これを用いる。KWJA の自動解析においては著者・読者のアノテーションが自動で付与されるため、これを話し手・聞き手に読み替えて

5) KWJA は照応関係の自動解析を行うが、この自動解析結果に対しては「話し手・聞き手」を除いて修正は加えない。

表 4　KWJA 出力との一致率

格	係受		ゼロ（文内）		ゼロ（文間）		外界		合計	
ガ	0.913	(1,221/1,338)	0.836	(808/966)	0.444	(347/781)	0.234	(579/2,479)	0.531	(2,955/5,564)
ヲ	0.880	(665/756)	0.838	(506/604)	0.543	(210/387)	0.170	(9/53)	0.772	(1,390/1,800)
ニ	0.764	(352/461)	0.663	(228/344)	0.463	(75/162)	0.092	(53/573)	0.460	(0,000/1,540)
デ	0.847	(150/177)	0.781	(114/146)	0.250	(4/16)	N/A	(0/0)	0.791	268/339
ト	0.937	(164/175)	0.84	(42/50)	0.739	(17/23)	0.136	(3/22)	0.837	(226/270)
ヘ	0.0	(0/28)	0.0	(0/26)	0.0	(0/16)	0.0	(0/1)	0.0	(0/71)
カラ	0.0	(0/54)	0.0	(0/20)	0.0	(0/7)	N/A	(0/0)	0.0	(0/81)
ヨリ	0.0	(0/6)	0.0	(0/2)	N/A	(0/0)	N/A	(0/0)	0.0	(0/8)
マデ	0.0	(0/14)	0.0	(0/17)	0.0	(0/3)	N/A	(0/0)	0.0	(0/34)
ガ２	0.755	(108/143)	0.657	(69/105)	0.163	(8/49)	0.144	(88/612)	0.300	(273/909)

利用する。話し手・聞き手は発話者に対応した相対関係として付与する。

ただし、発話の中で「あなたは」など明示的に話し手・聞き手を指している言及がある場合はそちらに優先してアノテーションを付与する。また照応に関する例外として、これら一人称・二人称表現に対しては話し手・聞き手に対する＝を付与する。

3　構築したコーパスの統計

実際に JDD の Topic 1（Dailylife）の一部（309 対話）に付与した各ラベルの統計について表 2 に示す。アノテーションの対象となる文数は 3,305、文節数は 11,968、述語数は 7,933 であった。まず、一般的な述語項構造のアノテーションと同様にガヲニ、続いてガ２格が多く存在ことがわかる。また、今回の特徴である話し手・聞き手がどの程度の頻度でどの格に現れたかを表 3 に示す。今回は話し手・聞き手は全て外界としてカウントされている。例えば、外界ガ格 2,479 件中、1,643 件が話し手、767 件が聞き手で、この 2 つで外界ガ格の 97.2%である。話し手・聞き手を除いた場合のガ格外界照応は 58 件、ニ格は 54 件である。著者・読者表現と比較した場合、特に聞き手（読者）表現への参照が大幅に増えている。これは今回のデータが二者間の対話で、相手を非明示的に対話中で参照することが多いことに起因すると考えられる。また、話し手・聞き手表現への参照はガヲニ以外ではト格での参照がいくつかあった。これは「一緒に○○しましょう」などの Commisive[25] の行為に付随する述語に生じたものが多かった。

4　自動解析の精度・修正点

アノテーションのベースにした KWJA の自動解析結果の正解率を表 4 に示す。ヘ、カラ、ヨリ、マデに関しては頻度が少なく、KWJA の自動解析を行わなかった。KWJA の自動解析結果はガ、ヲ、ニ、デ、ト、ガ２の係り受け、文内ゼロに関しては高い精度を計測している。また文間に関しても、対話用にチューニングがされているわけではないモデルとして高い精度を計測している。今回付与された外界に関しては多くが新しく導入した話し手・聞き手であったため、スコアが低くなっている。スコア算出の際、KWJA の解析結果からは著者・読者を話者との相対関係により読み替えて用いた。この結果から、今回付与した話し手・聞き手のラベルは従来の著者・読者と異なる性質を持ち、対話用にチューニングしたモデルが必要であることがわかる。

5　おわりに

人間同士の対話・コミュニケーション中に行われる意味のやりとりをモデル化することを指向して、日本語日常対話コーパスへの述語項構造アノテーションを行った。対話における述語項構造アノテーションを困難にする言い淀み、順序の入れ替わり、省略などが修正・補完されたデータセットを用いることで、今後対話における意味のやりとりの分析に大きく寄与するデータセットが構築されることが期待される。また、対話において生じる対話内外への照応や話し手・聞き手を考慮したアノテーションスキーマを構築し活用した。

まず 300 対話程度にアノテーションを行ったところ、対話コーパスに特有の現象がいくつか確認され、本アノテーションの意義が確認された。また、今回構築したアノテーションスキーマ・フレームワークにより 1 人月で 300 対話程度のアノテーションが可能であることが確認された。

今後は、まず本アノテーションを最低限 JDD の 30%、1500 対話程度まで拡張する。その後必要に応じて残りの部分のアノテーションを検討し、またコーパスを用いた対話分析を行う。

謝辞

本研究の一部は JSPS 科研費 JP19K13195、JP22K17943、JP23K24910 およびムーンショット型研究開発事業 JPMJMS2011-35 (fundamental research) の支援を受けたものです。

参考文献

[1] Xavier Carreras and Lluís Màrquez. Introduction to the conll-2005 shared task: Semantic role labeling. In **Proc. CoNLL-2005: Shared Task**, pp. 152–164, 2005.

[2] Jan Hajic, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, M Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In **Proc. CoNLL 2009: Shared Task**, pp. 1–18, 2009.

[3] Anders Björkelund, Love Hafdell, and Pierre Nugues. Multilingual semantic role labeling. In **Proc. CoNLL2009: Shared Task**, pp. 43–48, 2009.

[4] Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. In **Proc. ConLL2008: Shared Task**, pp. 159–177, 2008.

[5] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In **Proc. EMNLP-CoNLL2012: Shared Task**, pp. 1–40, 2012.

[6] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In **Proc. LAW-7 & ID**, pp. 178–186, 2013.

[7] Koichiro Yoshino, Shinsuke Mori, and Tatsuya Kawahara. Spoken dialogue system based on information extraction using similarity of predicate argument structures. In **Proc. SIGDIAL2011**, pp. 59–66, 2011.

[8] Yun-Nung Chen, William Yang Wang, and Alexander I Rudnicky. Learning semantic hierarchy with distributed representations for unsupervised spoken language understanding. In **Proc. INTERSPEECH**, pp. 1869–1873, 2015.

[9] Dian Yu, Mingqiu Wang, Yuan Cao, Izhak Shafran, Laurent Shafey, and Hagen Soltau. Unsupervised slot schema induction for task-oriented dialog. In **Proc. NAACL-HLT2022**, pp. 1174–1193, 2022.

[10] 黒橋禎夫. 京都大学テキストコーパス プロジェクト. 言語処理学会第 3 回年次大会論文集, 1997.

[11] 飯田龍, 小町守, 井之上直也, 乾健太郎, 松本裕治. 述語項構造と照応関係のアノテーション: NAIST テキストコーパス構築の経験から. 自然言語処理, Vol. 17, No. 2, pp. 2_25–2_50, 2010.

[12] Kikuo Maekawa. Kotonoha and bccwj: development of a balanced corpus of contemporary written japanese. In **Corpora and Language Research: Proc. ICKLLC**, pp. 158–177, 2007.

[13] 小町守, 飯田龍. Bccw」に対する述語項構造と照応関係のアノテーション. 日本語コーパス平成 22 年度公開ワークショップ, pp. 325–330, 2011.

[14] Mizuki Sango, Hitoshi Nishikawa, and Takenobu Tokunaga. Effectiveness of domain adaptation in japanese predicate-argument structure analysis. In **Proc. PACLIC**, 2018.

[15] 植田暢大, 波部英子, 松井陽子, 湯口彰重, 河野誠也, 川西康友, 黒橋禎夫, 吉野幸一郎. J-cre3: 実世界における参照関係解決のための 日本語対話データセット. 自然言語処理, Vol. 31, No. 3, pp. 1107–1139, 2024.

[16] 吉野幸一郎, 森信介, 河原達也ほか. 述語項の類似度に基づく情報抽出・推薦を行う音声対話システム. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3386–3397, 2011.

[17] Seiya Kawano, Koichiro Yoshino, David Traum, and Satoshi Nakamura. End-to-end dialogue structure parsing on multi-floor dialogue based on multi-task learning. **Frontiers in Robotics and AI**, Vol. 10, p. 949600, 2023.

[18] 赤間怜奈, 磯部順子, 鈴木潤, 乾健太郎. 日本語日常対話コーパスの構築. 言語処理学会第 29 回年次大会発表論文集, 2023.

[19] 笹野遼平, 黒橋禎夫ほか. 大規模格フレームを用いた識別モデルに基づく日本語ゼロ照応解析. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3328–3337, 2011.

[20] Nobuhiro Ueda, Kazumasa Omura, Takashi Kodama, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. Kwja: A unified japanese analyzer based on foundation models. In **Proc. ACL2023**, pp. 538–548, 2023.

[21] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In **Proc. ECCV2020**, pp. 752–768. Springer, 2020.

[22] 赤間怜奈, 浅原正幸, 若狭絢, 大村舞, 鈴木潤. 日本語日常対話コーパスへの基礎解析アノテーション. 言語処理学会第 30 回年次大会発表論文集, 2024.

[23] Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal dependencies. **Computational Linguistics**, Vol. 47, No. 2, pp. 255–308, 07 2021.

[24] Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. Japanese zero reference resolution considering exophora and author/reader mentions. In **Proc. EMNLP2013**, pp. 924–934, 2013.

[25] Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. Dialogue act annotation with the iso 24617-2 standard. **Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything**, pp. 109–135, 2017.