

日本語述語項構造解析のための統語パターン分析

中山 周 松林 優一郎 乾 健太郎

東北大学

{nakayamas, y-matsu, inui}@ecei.tohoku.ac.jp

1 はじめに

述語項構造解析の研究ではこれまでに様々な解析モデルが提案されてきたが、日本語の項構造解析における本質的な課題である項のゼロ照応解析については、解析対象を文内に限っても新聞ドメインで50%未満と低い精度にとどまり続けており、日本語の項構造解析の高い難易度を物語っている。

述語項構造解析において、ゼロ照応問題に対して解析の手がかりとするための情報は、これまでにさまざま考えられてきた。具体的に推定モデルに組み込まれた例としては、一般的な統語係り受けパス情報の他に、名詞、格助詞、述語の共起に関する統計値を用いる手法 [3, 4, 5, 9, 10]、二つの述語間で、それぞれの格スロットに出現する項の類似度を用いる手法 [1]、項の共有に関する情報として、支援動詞辞書を用いる手法や、スクリプト的な知識を学習する手法 [6, 2, 8] などがあるが、そもそも、ゼロ照応問題にどのような現象がどの程度あらわれるのか、あるいは、特定の解析モデルがフォーカスしている課題について、どの程度の割合を解くことが出来たかといった定量的な分析はこれまでになされておらず、今後具体的にどのような種類の問題を中心に取り組みればよいか不明瞭な状態となっている。

我々は、この難解な項のゼロ照応問題へ適切にアプローチするために、現象の特徴を出来る限り詳細に分析し把握することを試みる。本稿では、手始めにこの残り50%の文内ゼロ照応問題への挑戦的を絞り、それぞれの現象がどのような部分問題（あるいは局所的なパターン）から成立していると解釈できるのか、またそれら部分問題の分布を分析し、報告する。

2 分析対象・分析方法

述語項構造解析における「項のゼロ照応」とは、述語の特定の格について、文中に直接的な統語係り受け関係のある句として項が出現していないものの、実際には意味的に存在し、省略されている現象をさす。例えば、図1の文において、述語（Pで表記）「勉強」に対するガ格の項（Aで表記）は「太郎」であるが、これらは直接係り受け関係にない。

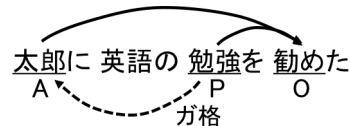


図1 文内の直接係り受け関係にない述語と項の例

このような事例について、既存の解析器の精度はF値で50%未満と十分とは言い難い。そこで、我々は問題の性質や現象の分布を把握するため、手始めに、このような項のゼロ照応のうち、実際の項に対応する表現（照応先）が同一文中に現れているものに対象を絞って現象の分布や既存システムでの解析精度を分析した。この理由は、照応先が文中に現れている事例では、探索のスコープが比較的短い上に、様々な統語的なパターンが観測でき、またこれらを機械的に分類できるため、初期の分析として見通しよく始めることができることによる。

我々はまず、文内ゼロ照応となっている述語と項のペア200事例に対して事前に人手による手がかり分析を実施し、文内ゼロ照応の多くの場合は並列構造や複数の述語をまたいだ構文的なパターン、述語間の意味的繋がりによって照応先を推定できるという従来の仮説と同様の帰結を得た。次に、その結果を踏まえ、分析対象コーパス中に現れる文内ゼロ照応関係にある述語Pと項Aのペアについて、以下の7つの指標で事例を分類した。

- 対象述語（P）の品詞（動詞、サ変名詞、その他）
- Pに対する項（A）の格（ガ、ヲ、ニ）
- Aと直接的に統語係り受け関係がある語（O）の品詞（動詞、サ変名詞、その他）
- A, O, Pの出現順序
- Oが述語相当語の場合、OとPがAを項として共有しているか
- OとPがAを項として共有している場合、
 - 二つの格ラベルが一致しているか
 - PとOが並列構造で繋がっているか、PがOの項であるか、それ以外
- OとPの間の係り受け距離

ここで、対象の述語Pと項Aだけではなく、Aと直接係り受け関係にある語Oを分析の対象に含めることで、直接的に項を推定しやすいAとOの関係を、より間接的

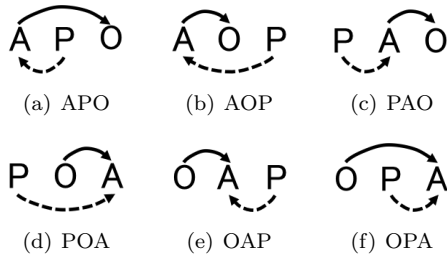


図2 A, O, P の出現順序と統語係り受け関係・述語-項関係の概観：実線は直接統語係り受け関係、破線は述語-項関係。

な関係となっている A と P の関係推定に利用できる事例が具体的にどの程度存在するか、また、既存のシステムでそのような事例をどの程度正しく推定できているかを分析する。ただし、事例によっては、文中に A として適切な共参照関係にある複数個の語が存在する場合があります。^{*1}そのような場合、(i)「述語よりも前にある語を優先する」(ii)「述語に単語位置がより近い語を優先する」の二つのルールを順に適用することで A を一意に定める。また一般に、A には、直接係り受け関係が認められる語の候補として、複数の直接係り元文節と係り先文節の中にある語が考えられるため、O を一意に定めるための方法が必要である。本稿では、A と P の関係において最も関わりが深いと思われる語 O を以下の方法で選択する。

1. A に対する係り元文節、係り先文節の主辞を O の候補として抽出する。
2. 1 で抽出した語に述語相当語が含まれる場合、述語相当語のみに候補を絞り込む。
3. 2 までで得られた候補のうち、P との統語係り受け距離がもっとも近いものを残す。
4. 3 までで得られた候補のうち、P に単語位置が最も近いものを一つだけ選ぶ。

この方法を用いれば、例えば次のような文について、 O_2 より O_1 を優先して選択することが出来る。

- (1) a. [豊か_P]で[興味深い_{O1}][世界_A]が[広がって_{O2}]いる。
 b. 手品を[した_{O2}][人_A]が[周りに[驚き_P]を[与える_{O1}]].

A, O, P の出現順序は、連体修飾などの構造的な特徴を簡潔にとらえるのに役立つ。{A, O, P} の置換として 6 通りの順序組がありうるが、これらを統語係り受け関係・述語-項関係と併記して示すと、図 2 のようになる。

分析対象のデータとして、NAIST テキストコーパス (NTC) 1.5 版を利用する。コーパスは、述語項構造解析の研究で一般的に利用されている Taira et al.[11] の分割に基づいて、訓練、開発、評価用データの区分に分割し、評価データにおいて事例カテゴリ別に出現頻度と既存システムの解析精度を測定する。また、各事例カテゴリに

^{*1} 「太郎、太郎、太郎と繰り返し手を振り呼ぶ次郎。」という文における「呼ぶ」のヲ格「太郎」のような場合。

表 1 NTC 1.5 評価データにおける文内ゼロ照応事例数及び解析精度

	Precision	Recall	F
ガ	61.36 (2928/4772)	36.99 (2928/7916)	46.15
ヲ	53.44 (202/ 378)	14.31 (202/1412)	22.57
ニ	26.79 (56/ 209)	13.18 (56/ 425)	17.67
ALL	59.45 (3186/5359)	32.67 (3186/9753)	42.17

における具体的な例文の提示や、事例ベースの分析には開発データを利用する。

既存システムの解析精度の参考値として、松林&乾 [7] のシステムを用いる。^{*2} 述語項構造解析の既存研究においては利用しているデータの違いもあり、正確に精度を比較することが難しいが、松林&乾のシステムは現状での最高精度を達成するシステムの一つである。このシステムは、統語係り受け解析結果と述語の位置を入力に取り、各単語がある述語の特定の格の項かどうかを点推定し、結果を出力する。評価は、松林&乾 [7] に記載されているのと同様の方法 (文節単位) で行った。以降、解析精度とは松林&乾のシステムの精度のことを指す。

評価データにおける文内ゼロ照応事例の統計値と解析精度は表 1 のとおりである。

3 分析結果にもとづく考察

本稿末の表 3 に分析の結果を示した。まず、格助詞ごとの分布を見ると、文内ゼロ照応の事例はガ格 81%、ヲ格 14%、ニ格 4% と、殆どの事例がガ格の省略であることが分かる。解析精度はガ格で最も高く、ヲ格、ニ格は殆ど正答が難しい状況となっている。述語側の品詞の分布は動詞 48%、サ変名詞 38%、その他 13% であり、主に動詞とサ変名詞がその大半を占める。品詞別の精度は、形容詞を除き、概ね 40% 前後となっている。

次に、A の直接係り先である O との関係を見ると、まず O の品詞は動詞が 65% と強い偏りを見せており、サ変名詞と合わせて A の直接係り先の 73% が動詞性の項構造を持ちうる品詞となっている。さらにこれらの O のうち、P と項を共有しているものの割合は約 78% と高く、文内ゼロ照応全体でみても O と P の間に項の共有がある事例が 58% と半数以上存在することが分かった。これは、並列構造解析や、スクリプト知識を利用して解ける可能性のある事例が比較的多数存在することを示している。さらに、項共有の有無によって解析精度に 20% 以上の大きな開きが見られたため、より詳細な分析として、O と P が項を共有している場合に (a) 「二つの格ラベルが一致しているか」(b) 「P と O が並列構造で繋がっ

^{*2} ただし、松林&乾の実験では学習・推定に正解の統語係り受け木を与えたのに対して、本稿では、より実用的なモデル上での精度を計測する目的で、CaboCha 0.66 の出力結果を用いて学習・推定した結果を用いる。このため、彼らの報告より解析精度が低い。分析時には、システムの述語項構造出力を正解の係り受け構造がアノテーションされた NTC (京大コーパス形式) の形態素と文節区切りにマッピングし、正解の統語係り受け木を用いて分析を行う。

表 2 A,O,P の語順、PO 間の係り受け距離別の解析精度

語順	PO 距離	Precision	Recall	F
APO	1	71.95 (1185/1647)	62.76 (1185/1888)	67.04
	2	62.44 (384/ 615)	48.36 (384/ 794)	54.51
	3	51.95 (133/ 256)	38.11 (133/ 349)	43.97
	4	36.25 (29/ 80)	22.66 (29/ 128)	27.88
	5 以上	49.12 (28/ 57)	24.56 (28/ 114)	32.75
	ALL	66.25 (1759/2655)	53.74 (1759/3273)	59.35
AOP	1	60.52 (305/ 504)	28.77 (305/1060)	39.00
	2	52.04 (115/ 221)	15.78 (115/ 729)	24.21
	3	45.75 (70/ 153)	14.74 (70/ 475)	22.29
	4	32.38 (34/ 105)	10.73 (34/ 317)	16.11
	5 以上	35.04 (41/ 117)	10.82 (41/ 379)	16.53
	ALL	51.36 (565/1100)	19.09 (565/2960)	27.83
OAP	1	0.00 (0/ 0)	0.00 (0/ 0)	0.00
	2	0.00 (0/ 0)	0.00 (0/ 0)	0.00
	3	66.45 (307/ 462)	39.31 (307/ 781)	49.40
	4	63.07 (111/ 176)	27.96 (111/ 397)	38.74
	5 以上	39.42 (41/ 104)	11.29 (41/ 363)	17.56
	ALL	61.86 (459/ 742)	29.79 (459/1541)	40.21
PAO	1	30.71 (78/ 254)	29.89 (78/ 261)	30.29
	2	29.11 (23/ 79)	12.11 (23/ 190)	17.10
	3	57.41 (62/ 108)	16.19 (62/ 383)	25.25
	4	37.14 (13/ 35)	8.23 (13/ 158)	13.47
	5 以上	26.67 (8/ 30)	5.93 (8/ 135)	9.70
	ALL	36.36 (184/ 506)	16.33 (184/1127)	22.54
POA	1	70.97 (176/ 248)	33.02 (176/ 533)	45.07
	2	46.51 (20/ 43)	17.70 (20/ 113)	25.64
	3	44.19 (19/ 43)	17.92 (19/ 106)	25.50
	4	18.18 (2/ 11)	4.65 (2/ 43)	7.41
	5 以上	25.00 (1/ 4)	2.13 (1/ 47)	3.92
	ALL	62.46 (218/ 349)	25.89 (218/ 842)	36.61

ているか、あるいは P が O の項であるか、それ以外か」という二つの指標で事例を分類した。

まず、二つの述語で格ラベルが一致しているものは項を共有している事例の 78% を占めており、更に、格ラベルが異なる場合と比べて 30% ほど精度が良いことが分かった。次に、P と Q が並列構造で繋がっているものは項を共有する事例の 13%、P が O の項である事例 (機能動詞構文や制御動詞構文などの事例を含む) は 10% と比較的少量に留まっており、その他の事例が大多数であることが分かった。その他の事例には、スクリプト的知識を要するものなどが含まれていると想定されるが、詳細な分析は今後の課題である。一方、より明確な手がかりがある並列構造や、O, P が述語-項関係になっているものは、松林&乾のシステムでは解析モデルとしてこれらの構造を明示的に扱ってはいないながらも、解析精度が 67% 前後とそれ以外の事例に比べて高い数値を示す結果となった。これは、これらの事例がゼロ照応解析の有望な手がかりとなっていることを示す証拠であり、明示的な並列構造解析や機能動詞・制御動詞の辞書的な取り扱いによって精度が更に向上する可能性を示唆している。

A, O, P の語順に関しては、主要部終端型である日本語では APO, AOP の割合が多く、この形が全体の 77% を占めている。続いて、述語が項を挟む形の PAO・OAP、更に、P を含む句が連体修飾節を修飾する形の POA と続く。解析精度は APO の語順で最も高く、F 値で 58% を達成している。これは前述の並列構造や機能動詞構文のほとんどが APO の語順を取っているためと考えられるが、一方で、二番目に多く、同様に O と P の並列構造

を含むと考えられる AOP の語順では、F 値 27% と解析精度に大きな開きが見られるのが興味深く、またその他の語順とも解析精度に開きが見られた。

P と O の間の係り受け距離を見ると 1 が 46%、2 以下が 68%、比較的統語的な手がかりが希薄となる 3 以上が 32% となだらかに分布している。一般に係り受け距離が遠くなるほど解析精度は低下していくが、P と O の間に直接係り受け関係が見られる事例では 53%、特にこのうち AOP の語順を取るものについては 67% と比較的高い精度で解析出来ていることが分かった。

4 おわりに

我々は、述語項構造解析における中心的な課題である項のゼロ照応問題へ適切にアプローチするために、現象の特徴を出来る限り詳細に分析し把握することを試みた。具体的には、人手による手がかり分析の事前調査にもとづき、文内ゼロ照応関係にある述語と項のペアを、統語情報と述語-項関係の情報を用いて機械的に分類可能な 7 つの指標の組み合わせで分類した。分析内容として、各事例カテゴリにおける事例数の分布を示したほか、最先端のシステムの精度を測り、各指標における解析精度の偏りを示した。特に、対象述語と、項と直接係り受け関係にある述語との間で項を共有している事例の割合が全体の 58% 存在することが分かったほか、並列構造や機能動詞構文の他にも項の共有情報を利用して解ける可能性のある事例が比較的多数存在することが示された。今後は、これらの残された手がかりの詳細を更に分析する予定である。また、今回利用した分析スクリプトは他の解析システムの結果を分析できる形で公開する予定である。

参考文献

- [1] Yuta Hayashibe, Mamoru Komachi, and Yuji Matsumoto. Japanese predicate argument structure analysis exploiting argument position and type. In *IJCNLP*, pp. 201–209, 2011.
- [2] 飯田龍, 徳永健伸. 述語対の項共有情報を利用した文間ゼロ照応解析. 言語処理学会第 16 回年次大会発表論文集, pp. 804–807, 2010.
- [3] Ryu Iida, Kentaro Inui, and Yuji Matsumoto. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *COLING-ACL 2006*, pp. 625–632. Association for Computational Linguistics, 2006.
- [4] Ryu Iida and Massimo Poesio. A cross-lingual ILP solution to zero anaphora resolution. In *ACL 2011*, pp. 804–813, 2011.
- [5] Kenji Imamura, Kuniko Saito, and Tomoko Izumi. Discriminative approach to predicate-argument structure analysis with zero-anaphora resolution. In *ACL-IJCNLP 2009 Short Papers*, pp. 85–88. Association for Computational Linguistics, 2009.
- [6] 小町守, 飯田龍, 乾健太郎, 松本裕治. 名詞句の語彙統語パターンを用いた事象性名詞の項構造解析. 自然言語処理, Vol. 17, No. 1, pp. 141–159, 2006.
- [7] 松林優一郎, 乾健太郎. 統計的日本語述語項構造解析のための素性設計再考. 言語処理学会第 20 回年次大会発表論文集, pp. 360–363, 2014.
- [8] 大内啓樹, 進藤裕之, Duh Kevin, 松本裕治. 述語対の項共有情報を利用した文間ゼロ照応解析. 情報処理学会 第 220 回自然言語処理研究会 研究報告, pp. 1–6, 2015.
- [9] Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. A fully-lexicalized probabilistic model for japanese zero anaphora resolution. In *COLING 2008 Volume 1*, pp. 769–776. Association for Computational Linguistics, 2008.
- [10] Ryohei Sasano and Sadao Kurohashi. A discriminative approach to japanese zero anaphora resolution with large-scale lexicalized case frames. In *IJCNLP 2011*, pp. 758–766, 2011.
- [11] Hiroto Taira, Sanae Fujita, and Masaaki Nagata. A Japanese predicate argument structure analysis using decision lists. In *EMNLP 2008*, pp. 523–532, 2008.

表 3 事例カテゴリー毎の事例数と解析精度

	Precision	Recall	F	出現割合	例文 ([p]: 対象述語, [ga, wo, ni]: 対象述語に対する格, [o]: 項の直接係り先/元の語)
対象述語の品詞					
動詞	61.32 (1650/2691)	35.16 (1650/4693)	44.69	48.11 (4693/9753)	立派な [しめ飾り p] は [高価 o] なので、いつも安い小さな物で [済ませて p] きました。
サ変名詞	61.24 (1152/1881)	30.62 (1152/3762)	40.83	38.57 (3762/9753)	米市場で圧倒的な [販売 p] を [誇る o] 3 [社 ga] だけに...
形容詞	38.71 (72/ 186)	26.18 (72/ 275)	31.24	2.81 (275/9753)	特殊法人の見直し問題は [古くて p] [新しい o] [問題 a] だが、
その他	52.26 (312/ 597)	30.50 (312/1023)	38.52	10.48 (1023/9753)	阪南 [市 a] は [今後 o]、既に払った [手付金 p] の返還を地権者に求め、新たな予定地選定に入る。
対象述語の格					
力	61.41 (2928/4768)	36.99 (2928/7916)	46.17	81.16 (7916/9753)	参謀本部が [大統領 ga] 直屬に [なれ o] ば、国防相の権限は大幅に縮小 [さ p] れることになり、
ヲ	53.44 (202/ 378)	14.31 (202/1412)	22.57	14.47 (1412/9753)	流感 [学校 wo] では [欠席 p] 扱いに [なら o] ないので、会社勤めではそうもいきません。
ニ	26.79 (56/ 209)	13.18 (56/ 425)	17.67	4.35 (425/9753)	わからないとは [先生 ni] を [質問 p] 攻めに [した o] 。
O の品詞					
述語動詞	61.76 (2335/3781)	36.59 (2335/6382)	45.95	65.43 (6382/9753)	同 [市 ga] の財政 [赤字 o] の補てんに [使わ p] れる。
述語名詞	56.12 (211/ 376)	27.37 (211/ 771)	36.79	7.90 (771/9753)	豆電球四 [個 wo] を [つないで o] 点灯 [さ p] せるには、一個よりずっと力がいった。
その他	53.42 (640/1198)	24.62 (640/2600)	33.70	26.65 (2600/9753)	米側が首腦 [会談 ni] にできるだけマスコミの [注目 p] を [集め o] させないようにした
語順及び係り受け					
APO	66.36 (2030/3059)	52.16 (2030/3892)	58.41	39.90 (3892/9753)	約百八十 [人 ga] が不法駐車 [改善 p] 命令を [受けて o] あり
AOP	51.53 (674/1308)	18.72 (674/3601)	27.46	36.92 (3601/9753)	シウメイコ上陸 [議長 ga] が [会談 o] 終了後、同通信に [語った p]。
PAO	36.28 (193/ 532)	16.19 (193/1192)	22.39	12.22 (1192/9753)	川崎さんと [親しい p] 同市内の [住職 ga] に「明男です」と男の声で電話が [あり o]、
POA	64.71 (209/ 323)	26.90 (209/ 777)	38.00	7.96 (777/9753)	日本語の [勉強 p] のために身についている、と [語る p] 韓国 [人 ga] も少なくない。
OAP	60.77 (79/ 130)	28.11 (79/ 281)	38.44	2.88 (281/9753)	アルゼンチンに [大敗 o] した [日本 ni] は [アーニング p] に見舞われてしまった。
OPA	33.33 (1/ 3)	10.00 (1/ 10)	15.38	0.10 (10/9753)	(開発データに事例なし)
項共有					
あり	74.76 (2150/2876)	38.19 (2150/5630)	50.55	78.70 (5630/7153)	約百八十 [人 ga] が不法駐車 [改善 p] 命令を [受けて o] あり
なし	30.91 (396/1281)	26.00 (396/1523)	28.25	21.29 (1523/7153)	さらに委員 [会 ga] に [注文 o] したいのは、地方での公職やシンポジウム開催を積極的に [考えて p] ほしいということだ。
項の直接係り先/元の述語と対象述語の格が同じ/異なる					
同じ	77.14 (1940/2515)	43.97 (1940/4412)	56.01	78.36 (4412/5630)	[経済 ga] が発展 [し p]、成熟 [する o] に従って
異なる	58.17 (210/ 361)	17.24 (210/1218)	26.60	21.63 (1218/5630)	事あるごとに電話一本で相談 [する p] 欧米の [首脳 ga] レベルに日本はまだ [達して o] いない
項共有があるとき					
P、O が並列	88.36 (410/ 464)	53.95 (410/ 760)	66.99	13.49 (760/5630)	ドイツ「シュテリン」誌のヨヘン・ベースト [記者 ga] が就職を [受け o]、[死亡 p]。
P が O の項	83.82 (316/ 377)	56.53 (316/ 559)	67.52	9.92 (559/5630)	朝鮮民主主義人民共和国をめぐる [問題 ga] も、新たな [展開 p] を [見せよう o]。
その他	69.98 (1424/2035)	33.03 (1424/4311)	44.88	76.57 (4311/5630)	同 [市 ga] の財政 [赤字 o] の補てんに [使わ p] れる。
P から O までの係り受け距離					
1	65.73 (2012/3061)	45.16 (2012/4456)	53.54	45.67 (4455/9753)	[堤防 wo] を鉄、コンクリートで [補強 o] 後、土で [覆い p] 自然の復元を目指す手法を提唱
2	57.04 (636/1115)	29.13 (636/2183)	38.57	22.38 (2183/9753)	あまりにも性高 [すぎた p] [ため o]、韓国内でもこの政策 [転換 p] には疑問の声も出ている。
3	51.65 (344/ 666)	22.28 (344/1544)	31.13	15.83 (1544/9753)	しかし、[韓国 ga] の民主主義 [発展 o] のためには、[避けて p] 通れない道である。
4	38.60 (105/ 272)	13.74 (105/ 764)	20.27	7.83 (764/9753)	[船木 ga] を指導 [する o] コーチも「怖いなんて言ってきたら [競技 p] に出さない」と話した
5 以上	36.93 (89/ 241)	11.03 (89/ 807)	16.98	8.27 (807/9753)	被災者が [国 ga] や [自治体 o] が助けてくれたと実感できる事業の実現に [工夫 p] を求めたい