事前学習と finetuning の類似性に基づくゼロ照応解析

今野颯人 [1]　清野舜 [2,1]　松林優一郎 [1,2]　大内啓樹 [2]　乾健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所
{ryuto, inui}@ecei.tohoku.ac.jp　y.m@tohoku.ac.jp
{shun.kiyono, hiroki.ouchi}@riken.jp

1 はじめに

日本語や中国語では，述語の項が頻繁に省略される．**ゼロ照応解析**（zero anaphora resolution; **ZAR**）はそのような項の省略を解析するタスクであり，文の意味理解のために重要な役割を担っている．図 1 の日本語における ZAR の例では，「来なかった」という述語の主語である「友人」が省略されている．

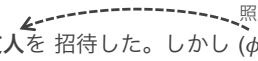

図 1 日本語におけるゼロ照応解析の例．省略された項は**ゼロ代名詞**と呼ばれ，ϕ で表される．

この省略の解析にはゼロ代名詞と照応先周辺にある文脈的つながりを理解するための常識的知識（**照応的知識**と呼ぶ）が必要である．例えば図 1 では，「招待された友人」が「来なかった人」になりやすいといった知識である．照応的知識を大規模コーパスから獲得する研究はこれまでにも行われてきたが [1, 2]，従来は複数の述語と項のペアで表されるイベント間の関係（スクリプト知識）に焦点を絞っており，ZAR への効果は限定的であった．

そこで本研究では，この照応的知識獲得の問題に対して二つの提案を行う．一つは，より広範な文脈表現を扱うための学習方法の提案（**知識獲得法の改善**）であり，もう一つは，学習した照応的知識を適切に利用する方法の提案（**知識適用法の改善**）である．第一に，照応的知識の獲得に焦点を当てた新たな事前学習タスクである**擬似ゼロ代名詞解析**（pseudo zero pronoun resolution; **PZERO**）を提案する（図 2）．PZERO は，生テキストに 2 回以上登場する名詞句の一つをマスクし，マスクに入る語を文脈から選択するタスクである．これは，これまでに照応的知識を必要とするタスクの性能に貢献することが報告されてきた [3, 4, 5] 従来の MLM の事前学習方法（***cloze* タスク** [6]）を，より照応関係を直接学習するように設計し直すもので，広範な文脈表現の間

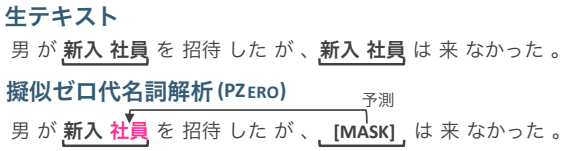

図 2 擬似ゼロ代名詞解析（PZERO）の概要

の関係をモデルに与えられる．

第二に，新たな事前学習で獲得した知識を適切に解析へ適用するため，事前学習とタスク形式をそろえた ZAR モデル，**擬似ゼロ代名詞に基づく項選択**（argument selection as PZERO; **AS-PZERO**）**モデル**を提案する．これにより，事前学習と fine-tuning の隔りを緩和し，事前学習で得た照応的知識を ZAR へ適用することを狙う．提案手法の全体像を図 3 に示す．

実験の結果，我々が提案する事前学習タスクとモデルを組み合わせることで，日本語 ZAR の性能を大幅に向上させられることがわかった．

2 日本語ゼロ照応解析

日本語 ZAR は，述語の項を同定する述語項構造解析タスクの一部として定式化されている．述語項構造解析では，ZAR に加え，述語と直接係り受け関係にある項（DEP）も解析対象であり，それぞれの述語に対してガ格，ヲ格，ニ格を同定する．また，ゼロ照応は述語とその項（ゼロ代名詞の照応先）との位置関係によって以下の三つに分類される．

- 文内ゼロ（*intra*）：項が述語と同じ文内にある．
- 文間ゼロ（*inter*）：述語の文より前方にある．
- 外界ゼロ（*exophora*）：項が文書内に出現しない．

本研究では上記三つと DEP を解析対象とする．

3 事前学習：擬似ゼロ代名詞解析

3.1 定式化とモチベーション

擬似ゼロ代名詞解析（PZERO）は，文章中に 2 回以上出現する名詞句のうち 1 箇所をマスクした文章

を受け取り，マスクに入る名詞句の主辞（サブワード）を入力テキスト中から一つ選択するタスクとして定式化する．図 2 の例では，2 回出現している「新入社員」が解析対象となる．このタスクの狙いは，同じ文字列の名詞句が照応関係にあるという強い仮定を置き，マスクされた名詞句を擬似的にゼロ代名詞とみなすことで，生文書から得る大量の訓練事例を用いて照応関係を直接事前学習し，ZAR に必要な照応的知識を獲得することである．

モデルは，一つのマスクトークン [MASK] を含む長さ T の系列 $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ を入力として受け取り，[MASK] に該当する名詞句の末尾のトークンを入力系列から選択する．ここで，$\boldsymbol{x} \in \mathbb{R}^{|\mathcal{V}|}$ は one-hot ベクトル，$\mathcal{V}$ は語彙である．また，マスクされた名詞句と同じ表層形をもつ名詞句は全て正解とみなす．

3.2 擬似データ作成方法

PZERO の訓練事例の作成方法を述べる．まず，生文書から連続した n 個の文を取得し，サブワード系列へと変換する．このとき，MLM の訓練事例作成方法 [6] に従って，系列の先頭に [CLS]，文の境界に [SEP] を挿入する．また，系列の長さがモデルの最大系列長 T_{max} を超えないように系列の先頭部分を削って調節する．次に，最後の文の中から同じ文字列が入力系列中に 2 回以上出現する名詞句を選択し，これを一つの [MASK] へと置き換える．

3.3 事前学習方法

Transformer ベースの MLM [6] をモデルに用いる．まず，入力系列 $\boldsymbol{X}$ を受け取り，各 $\boldsymbol{x}_t$ に対応する D 次元の埋め込み表現 $\boldsymbol{e}_t \in \mathbb{R}^D$ を入力層から得る．

$$\boldsymbol{e}_t = \boldsymbol{e}_t^{\text{token}} + \boldsymbol{e}_t^{\text{position}}. \tag{1}$$

ここで，$\boldsymbol{e}_t^{\text{token}} \in \mathbb{R}^D$ は各トークンを表す単語埋め込み表現，$\boldsymbol{e}_t^{\text{position}} \in \mathbb{R}^D$ は位置を表す位置埋め込み表現である．次に，得られた埋め込み表現の系列 $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_T)$ を，transformer 層によって最終隠れ層の系列 $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T)$ へとエンコードする．その後，各最終隠れ層 $\boldsymbol{h}_t \in \mathbb{R}^D$ について，そのトークンが [MASK] に入るかどうかを表すスコア $s_t \in \mathbb{R}$ を，[MASK] の最終隠れ層 $\boldsymbol{h}_{\text{mask}}$ との計算によって得る．

$$s_t = (\boldsymbol{W}_1 \boldsymbol{h}_t + \boldsymbol{b}_1)^\top \cdot (\boldsymbol{W}_2 \boldsymbol{h}_{\text{mask}} + \boldsymbol{b}_2). \tag{2}$$

$\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{D \times D}$ と $\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{R}^D$ は学習パラメータである．これよりスコア系列 $\boldsymbol{s} = (s_1, \ldots, s_T)$ を得る．

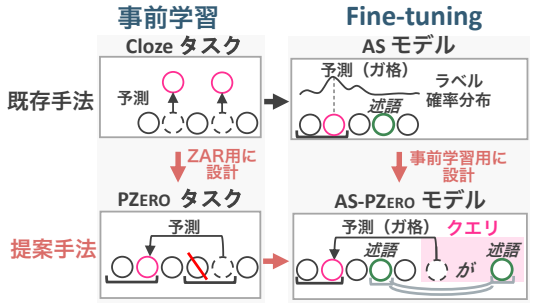

図 3 提案手法の概要

訓練では，正解となるトークンのスコアが最大となるようモデルを学習する．損失関数にはカルバック・ライブラー情報量 $\mathcal{L} = \text{KL}(Y || \text{softmax}(\boldsymbol{s}))$ を用いる．$Y \in \mathbb{R}^T$ は正解の位置を表す確率分布であり，正解となるトークンが n 個存在する場合，正解の位置には $1/n$ が，それ以外には 0 が割り当てられる．

4 Fine-tuning：ゼロ照応解析モデル

4.1 ラベル確率に基づく項選択モデル

我々のベースラインである**ラベル確率に基づく項選択**（argument selection with label probability; **AS**）**モデル**は，Kurita ら [7] のモデルをベースとしており，事前学習済みモデルの上に分類層を追加したものである．モデルは，系列 $\boldsymbol{X}$ と述語の区間を表す p_{start}，p_{end} を入力として受け取り，述語の l 格の項となる単語を $\boldsymbol{X}$ から一つ選択する．ここで，l はガ，ヲ，ニのいずれかを表す．事前学習時の入力と同様，入力系列 $\boldsymbol{X}$ は [CLS] と [SEP] を含んだ複数の文から構成され，最大系列長は T_{max} である．対象述語は常に末尾の文に存在する．また，述語の項が入力系列 $\boldsymbol{X}$ に存在しない場合にはモデルに [CLS] を選択させる．

まず，入力系列 X を受け取り，各 $\boldsymbol{x}_t \in \{0, 1\}^{|\mathcal{V}|}$ に対応する埋め込み表現 $\boldsymbol{e}_t \in \mathbb{R}^D$ を入力層から得る．

$$\boldsymbol{e}_t = \boldsymbol{e}_t^{\text{token}} + \boldsymbol{e}_t^{\text{position}} + \boldsymbol{e}_t^{\text{predicate}}. \tag{3}$$

ここで，$\boldsymbol{e}_t^{\text{token}}$ と $\boldsymbol{e}_t^{\text{position}}$ は（1）式と同様であり，新たに導入した $\boldsymbol{e}_t^{\text{predicate}} \in \mathbb{R}^D$ は，t 番目のトークンが述語かどうかを表す埋め込み表現である．入力層の操作を図 4 に示す．次に，各埋め込み表現 $\boldsymbol{e}_t$ から最終隠れ層 $\boldsymbol{h}_t \in \mathbb{R}^D$ を事前学習済みの transformer 層を用いて得る．その後，入力系列に対するラベル l の確率分布 $\boldsymbol{o}_l = (o_{l,1}, \ldots o_{l,T}) \in \mathbb{R}^T$ を分類層から得る．

$$o_{l,t} = \frac{\exp(\boldsymbol{w}_l^\top \boldsymbol{h}_t + b_l)}{\sum_t \exp(\boldsymbol{w}_l^\top \boldsymbol{h}_t + b_l)}. \tag{4}$$

ここで，$\boldsymbol{w}_l \in \mathbb{R}^D$ と $b_l \in \mathbb{R}$ はモデルパラメータであ

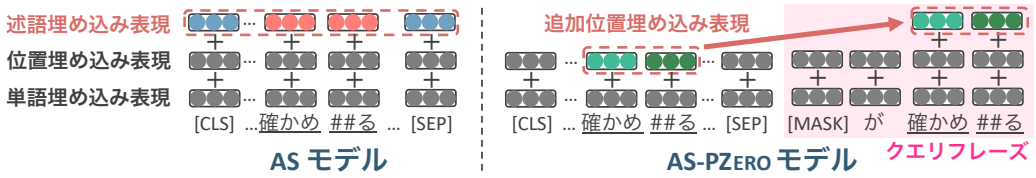

図 4　AS と AS-PZERO における入力層．AS-PZERO にはクエリフレーズが存在．対象述語の位置の与え方が異なる．

表 1　NTC 1.5 の評価セットにおける，**入力が 1 文のみ**の設定での F_1 値．

ID	Method	ZAR-*intra*	DEP	All
(a)	Matsubayashi ら [8]	55.55	90.26	83.94±0.12
(b)	Konno ら [5]	64.15	92.46	86.98±0.13
(c)	AS	69.32	93.65	88.87±0.12
(d)	AS-PZERO	**69.91**	**93.83**	**89.06**±0.11

る．最後に，確率分布 $\boldsymbol{o}_l$ に従って，確率が最大となるトークンを述語の項 l として一つ選択する．

モデルが [CLS] を選択した場合，さらに項を 4 つのカテゴリ $z \in \{\mathtt{author}, \mathtt{reader}, \mathtt{general}, \mathtt{none}\}$ へ分類する．ここで，author, reader, general は外界ゼロの細分類を，none は項が存在しないことを表す．各カテゴリに対する確率分布 $\boldsymbol{o}_l^{\mathrm{exo}} = (o_{l,\mathtt{author}}^{\mathrm{exo}}, o_{l,\mathtt{reader}}^{\mathrm{exo}}, o_{l,\mathtt{general}}^{\mathrm{exo}}, o_{l,\mathtt{none}}^{\mathrm{exo}}) \in \mathbb{R}^4$ は，分類層により [CLS] の最終隠れ層 $\boldsymbol{h}_1$ から得る．

$$o_{l,z}^{\mathrm{exo}} = \frac{\exp(\boldsymbol{w}_{l,z}^{\top}\boldsymbol{h}_1 + b_{l,z})}{\sum_z \exp(\boldsymbol{w}_{l,z}^{\top}\boldsymbol{h}_1 + b_{l,z})}. \tag{5}$$

$\boldsymbol{w}_{l,z} \in \mathbb{R}^D$ と $b_{l,z} \in \mathbb{R}$ はモデルパラメータである．

学習時は，項を構成する末尾のトークンに正解ラベルを割り当てる．共参照関係によって正解の項が複数存在する場合は，全ての正解となる項にラベルを割り当てる．また，3.3 節と同様に，正解を表す確率分布 $\boldsymbol{Y} \in \mathbb{R}^T$ を作成し，正解となるトークンの確率が高くなるようにモデルを学習する．

4.2　擬似ゼロ代名詞に基づく項選択モデル

AS モデルは，事前学習で得られた（2）式のパラメータ（$\boldsymbol{w}_1, \boldsymbol{w}_2, b_1, b_2$）を使わず，（4）式により新たなパラメータ（$\boldsymbol{w}_l, b_l$）を用いて学習するため，事前学習によって獲得した照応的知識を効果的に使えていない可能性がある．我々が提案する**擬似ゼロ代名詞に基づく項選択**（argument selection as PZERO; **AS-PZERO**）**モデル**は，PZERO で訓練されたパラメータを使い，ZAR を PZERO として解析する．具体的には，ZAR の入力系列 $\boldsymbol{X}$ に [MASK] を含んだフレーズを挿入し，PZERO と同様の形式で述語の項を解析する．このモデルは事前学習と finetuning の隔りを緩和する既存研究 [9, 10] から着想を得たものである．

AS-PZERO モデルの入力系列を $\boldsymbol{X}'$ とする．図 4 に示すように，$\boldsymbol{X}'$ は $\boldsymbol{X}$ の末尾にクエリフレーズを挿入することで得られる．クエリフレーズによって，モデルは [MASK] に該当する単語を入力系列から選択する PZERO と同様の形式で ZAR を解析できる．クエリフレーズは（1）[MASK]，（2）項を表す格助詞（が・を・に），（3）対象述語から構成される．対象述語を構成するトークン数を $T_{\mathrm{predicate}}$ とすると，$\boldsymbol{X}'$ の系列長は $T + 2 + T_{\mathrm{predicate}}$ である [1)]．

まず，入力系列 $\boldsymbol{X}'$ を受け取り，各 $\boldsymbol{x}_t \in \{0,1\}^{|V|}$ に対応する埋め込み表現 $\boldsymbol{e}_t \in \mathbb{R}^D$ を得る．

$$\boldsymbol{e}_t = \boldsymbol{e}_t^{\mathrm{token}} + \boldsymbol{e}_t^{\mathrm{position}} + \boldsymbol{e}_t^{\mathrm{addposi}}. \tag{6}$$

ここで，$\boldsymbol{e}_t^{\mathrm{addposi}}$ は追加位置埋め込み表現であり，対象述語の位置をモデルへ与えるために新たに用意したものである．$\boldsymbol{e}_t^{\mathrm{addposi}}$ は $1 \le t \le T+2$ では $\boldsymbol{0} \in \mathbb{R}^D$ を，それ以外では $\boldsymbol{e}_{T+3+m}^{\mathrm{addposi}} = \boldsymbol{e}_{p_{\mathrm{start}}+m}^{\mathrm{position}}$ をとる．ここで，$m \in \mathbb{N}$ は $0 \le m < T_{\mathrm{predicate}}$ を満たす．例えば図 4 では，“確かめ” と “##る” の 2 つから構成される対象述語について，それぞれのトークンの位置埋め込み表現がクエリフレーズのそれぞれのトークンへと加算されている．この操作により，対象述語の位置をモデルへ与えることができる．

埋め込み表現を得た後の計算は 3.3 節と同様である．[CLS] トークン（$\boldsymbol{x}_1$）のスコアが最も高かった場合，モデルは外界ゼロについて，4.1 節における式（5）と同様に計算を行う．

5　実験設定

PZERO データセット　日本語 Wikipedia を PZERO の訓練データとして使用した．コーパス中の全名詞句を解析対象とし，係り受け解析器 *Cabocha* [11] の解析結果から品詞を用いたルール [2)] によって名詞句を同定した．入力系列に用いる文数 n の最大値は 4 とした．事例数は約 1740 万となり，うち約 3000 を開発セット，残りを訓練セットとした．

ZAR データセット　NAIST Text Corpus（NTC）1.5 [12,

1）　$\boldsymbol{X}$ は事前に最大系列長 $T_{\max}$ を超えないように調整する．
2）　名詞句同定のルールについては付録 A に詳細を示す．

表 2 NTC 1.5 の評価セットにおける，**入力が複数文**の設定での F_1 値．**太字**の値は各列における最高性能を示す．(h) から (k) における F_1 の性能向上は，全ての ZAR のカテゴリで統計的に有意差が見られた．

	事前学習 1	事前学習 2		finetuning		ZAR				DEP	All
ID	Cloze	Cloze	PZERO	AS	PZERO	All	*intra*	*inter*	*exophora*		
(f)	✔			✔		62.27±0.42	71.55±0.32	44.30±0.79	64.04±0.63	94.44	82.97
(g)	✔				✔	62.47±0.53	71.09±0.59	45.20±0.51	64.41±0.94	94.46	83.03
(h)	✔	✔		✔		62.54±0.47	71.82±0.21	44.98±1.05	63.94±0.73	**94.51**	83.10
(i)	✔	✔			✔	62.85±0.19	71.52±0.22	45.97±0.42	64.55±0.71	94.49	83.18
(j)	✔		✔	✔		63.06±0.19	71.96±0.38	46.37±0.34	64.42±0.46	94.43	83.26
(k)	✔		✔		✔	**64.18**±0.23	**72.67**±0.32	**48.41**±0.35	**65.40**±0.36	94.50	**83.65**

13] を使用した[3]．評価スクリプトには Matsubayashi ら [8] と同様のものを使用した．

モデル　モデルの実装には `Transformer` ライブラリ [14] を使用し，パラメータ初期値に `bert-base-japanese` を用いた[4]．実験では，5 つの異なるシード値の平均で比較を行った．

6　実験結果・分析

実験では（1）PZERO による事前学習の効果と（2）AS-PZERO による finetuning の効果を調べる．以下の 2 つの設定で実験を行った．

1. **入力が 1 文のみ**: 既存研究の多くはこの設定で実験を行っている [8, 15, 5]．既存研究と我々のモデルを同様の設定において比較する．*intra* と DEP が評価の対象となる．
2. **入力が複数文**: 述語を含む文とその前方文が入力となる．*intra*，*inter*，*exophora*，DEP が評価の対象となる．

入力が 1 文のみ　事前学習済みモデルから finetuning した AS モデルと AS-PZERO モデルの結果を表 1 に示す．結果より，我々のモデルは *intra* と DEP の両方で既存の最高性能 [5] を大きく上回っており，性能が十分高いことを示している．

入力が複数文　PZERO による事前学習の効果を調べるため，3 つの事前学習済みモデルを用意する．まず，事前学習 1 によって，事前学習済みモデル（モデル 1）を用意する．そこからさらに，事前学習 2 によって，*cloze* タスクと PZERO，それぞれ同じ更新回数で事前学習する（モデル 2&3）．3 つの事前学習済みモデルから，AS と AS-PZERO でそれぞれ finetuning し，合計 6 つのモデルを用意した．その結果を表 2 に示す．

（1）文脈を与えることで intra と DEP の性能は向上するか？　（f），（g）のモデルは表 1 の（d），（e）とそれぞれ同じモデルであり，入力文の数だけが異なる．これら 4 つのモデルにおける *intra* と DEP について比較すると，たとえ解析対象が文内のみに限ったとしても，文脈を与えることで性能を向上させられることがわかる．これは既存研究 [16, 17] の結果とも一致する．

（2）AS の性能は PZERO による事前学習で向上するか？　PZERO で事前学習したモデル（j）は *cloze* タスクで事前学習したモデル（h）よりも ZAR の性能で上回っており，特に文間ゼロ（*inter*）で大きな向上を示している（44.98 → 46.37）．文間ゼロは項の候補が多いうえ，統語的手がかりが少なく，解析に意味的な理解が必要である．文間ゼロの性能向上は，モデルが PZERO による事前学習によって照応的知識をより獲得していることを示唆している．

(3) PZERO と AS-PZERO の組み合わせで性能は向上するか？　finetuning の方法だけが異なる，モデル（j）と（k）を比較する．（k）は全てのカテゴリで（j）を上回っており，さらに DEP を除いた全てのカテゴリで最も性能が高かった．この結果は，AS-PZERO と PZERO を組み合わせることで事前学習と finetuning の隔りをうまく緩和でき，事前学習で獲得した照応的知識を ZAR へうまく適用できることを示唆している．

7　おわりに

本稿では，日本語における ZAR に取り組んだ．ZAR に必要な照応的知識を生コーパスから獲得するための新たな事前学習タスクである PZERO と，事前学習で得た知識をうまく ZAR へ適用するための AS-PZERO モデルを提案した．実験結果より，両者を組み合わせることで世界最高性能を達成した．

謝辞

本研究は JSPS 科研費 JP19H04425，JP19K12112 の助成を受けたものです．

3）　使用したデータセットの統計は付録 B に示す．
4）　使用したハイパーパラメータは付録 C に示す．

参考文献

[1]Nathanael Chambers and Dan Jurafsky. Unsupervised Learning of Narrative Schemas and their Participants. In *ACL and IJCNLP*, pages 602–610, 2009.

[2]Mark Granroth-Wilding and Stephen Clark. What Happens Next? Event Prediction Using a Compositional Neural Network Model. In *AAAI*, pages 2727–2733, 2016.

[3]Abdulrahman Aloraini and Massimo Poesio. Cross-lingual Zero Pronoun Resolution. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *LREC*, pages 90–98, 2020.

[4]Linfeng Song, Kun Xu, Yue Zhang, Jianshu Chen, and Dong Yu. ZPR2: Joint zero pronoun recovery and resolution using multi-task learning and BERT. In *ACL*, pages 5429–5434, July 2020.

[5]Ryuto Konno, Yuichiroh Matsubayashi, Shun Kiyono, Hiroki Ouchi, Ryo Takahashi, and Kentaro Inui. An Empirical Study of Contextual Data Augmentation for Japanese Zero Anaphora Resolution. In *COLING*, pages 4956–4968, December 2020.

[6]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, pages 4171–4186, 2019.

[7]Shuhei Kurita, Daisuke Kawahara, and Sadao Kurohashi. Neural Adversarial Training for Semi-supervised Japanese Predicate-argument Structure Analysis. In *ACL*, pages 474–484, 2018.

[8]Yuichiroh Matsubayashi and Kentaro Inui. Distance-Free Modeling of Multi-Predicate Interactions in End-to-End Japanese Predicate-Argument Structure Analysis. In *COLING*, pages 94–106, 2018.

[9]Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *NeurIPS*, pages 5754–5764, 2019.

[10]Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *ACL*, pages 8342–8360, 2020.

[11]Taku Kudo and Yuji Matsumoto. Japanese Dependency Analysis using Cascaded Chunking. In *CoNLL*, 2002.

[12]Ryu Iida, Mamoru Komachi, Naoya Inoue, Kentaro Inui, and Yuji Matsumoto. Annotating Predicate-Argument Relations and Anaphoric Relations: Findings from the Building of the NAIST Text Corpus. *Natural Language Processing*, 17(2):25–50, 2010.

[13]Ryu Iida, Mamoru Komachi, Naoya Inoue, Kentaro Inui, and Yuji Matsumoto. NAIST Text Corpus: Annotating Predicate-Argument and Coreference Relations in Japanese. In *Handbook of Linguistic Annotation*, pages 1177–1196. Springer, 2017.

[14]Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *EMNLP demo*, pages 38–45, 2020.

[15]Hikaru Omori and Mamoru Komachi. Multi-Task Learning for Japanese Predicate Argument Structure Analysis. In *NAACL*, pages 3404–3414, 2019.

[16]Chaoyu Guan, Yuhao Cheng, and Hai Zhao. Semantic Role Labeling with Associated Memory Network. In *NAACL*, pages 3361–3371, June 2019.

[17]Tomohide Shibata and Sadao Kurohashi. Entity-Centric Joint Modeling of Japanese Coreference Resolution and Predicate Argument Structure Analysis. In *ACL*, pages 579–589, 2018.

[18]Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. A Japanese Predicate Argument Structure Analysis using Decision Lists. In *EMNLP*, pages 523–532, 2008.

[19]Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.

表 3　NAIST Text Corpus 1.5 の統計情報

データセット		*dep*	*intra*	*inter*	*exophora*
訓練	ガ格	36934	12219	7843	11511
	ヲ格	24654	2136	948	128
	ニ格	5744	465	294	60
開発	ガ格	7424	2665	1812	1917
	ヲ格	5055	445	177	32
	ニ格	1612	138	101	28
評価	ガ格	14003	4993	3565	3717
	ヲ格	9407	906	371	55
	ニ格	2493	260	145	54

A　名詞句同定のルール

本稿では，PZERO の解析対象をルールによって同定した名詞句とした．名詞句の同定には，係り受け解析器である *Cabocha* [11] で解析を行い，以下の品詞タグによるルールを用いた．

1. cabocha で文節に区切る．
2. 名詞は存在するが動詞は存在しない文節を選択する．ここで，名詞句の始点を q_{start}，終点を q_{end} とする．
3. 文節の終点から始点へと順に単語を見ていき，名詞または名詞性名詞接尾辞に属する単語を発見した段階で，その単語の位置を q_{end} と定める．
4. 文節の始点から終点へと順に単語を見ていき，記号以外の単語を発見した段階で，その単語の位置を q_{start} と定める．
5. q_{start} から q_{end} までの単語に “始まりの括弧” が含まれていた場合，q_{end} をその括弧の一つ前の単語の位置とする．
6. q_{start} から q_{end} までの単語が名詞句候補となる．名詞句の候補が，記号・英語・数字のみで構成されていた場合は名詞句候補から除外する．また，名詞句候補が “もの”，“こと”，“ため”，“何”，“誰” のいずれかであった場合も名詞句候補から除外する．

B　データセットの統計情報

本稿では，NAIST Text Corpus（NTC）1.5 [12, 13] を使用した．また，Taira らの分割法 [18] に従って，訓練・開発・評価セットを作成した．その統計情報を表 3 に示す．

C　ハイパーパラメータ

表 4 に実験のハイパーパラメータの設定を示す．

表 4　ハイパーパラメータ

項目名	値
最適化アルゴリズム	Adam [19] ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-8}$)
事前学習	
ミニバッチ事例数	2,048
最大学習率	1.0×10^{-4}（Cloze タスク），2.0×10^{-5}（PZERO）
Learning Rate Schedule	Inverse square root decay
Warmup Steps	5,000
学習更新回数	30,000
損失関数	Cross entropy（Cloze タスク），KL divergence（PZERO）
finetuning	
ミニバッチ事例数	256
最大学習率	1.0×10^{-4}（AS），1.0×10^{-5}（AS-PZERO）
Learning Rate Schedule	Matsubayashi ら [8] の Appendix A と同様の設定
最大エポック数	150
学習停止アルゴリズム	Matsubayashi ら [8] の Appendix A と同様の設定
損失関数	KL divergence（AS，AS-PZERO），cross entropy（*exophora*）