質問応答に基づく日本語ゼロ代名詞同定

岩田 晟　　渡辺 太郎
奈良先端科学技術大学院大学
{iwata.sei.is6,taro}@is.naist.jp

永田 昌明
NTT コミュニケーション科学基礎研究所
masaaki.nagata.et@hco.ntt.co.jp

1 はじめに

日本語や中国では，文中の主語や目的語などの構成要素が省略されることがあり，このような省略をゼロ代名詞という．例えば，日英翻訳タスクにおいて日本語側の省略を補完する必要があり，結果として，このような省略は自然言語処理の応用タスクへと大きな影響を与える．

近年では共参照解析や NER などを質問応答 (Question Answering,QA) の一つである SQuAD 形式＼ citeSQuAD のスパン予測問題をとして捉える研究がある [1][2].

本研究 ではゼロ代名詞検出および同定タスクを QA タスク として捉えた手法を提案する. 動詞または形容詞をクエリとし，その単語 (クエリ) を主辞とする節において省略された構成要素の種類およびその節の開始位置と終了位置を予測する．句構造情報がアノテーションされた NPCMJ[3] を用いたゼロ代名詞同定タスクにおいて，BERT[4] を用いた系列ラベリングの手法と比べ，提案手法は F1 値で 2.2 ポイントの向上が確認できた.

2 関連研究

植田 ら [5] は BERT[4] を用いた述語項構造解析・共参照解析・橋渡し照応解析の同時学習を提案した. Konno ら [6] は BERT を用いた Data Augmentation を行うことで，ゼロ照応解析の精度を向上させた.

Song ら [7] は，句構造の関係を利用し，BERT を用いてゼロ代名詞検出をサブタスクにしつつ，ゼロ代名詞の同定とゼロ照応解析を同時学習を行った. この手法では自己注意機構を用いて，照応先の開始位置と終了位置を予測することでゼロ照応解析とした．Aloriani ら [8] と Aloriani ら [9] の研究では，句構造情報を踏まえて，BERT-base Multilingual の埋め込みを多層パーセプトロンの入力として用いることで，ゼロ代名詞同定とゼロ照応解析を行った. Takeno ら [10] は句構造構文情報の様々な素性を用いることで，ゼロ代名詞検出を行った.

Devlin ら [4] は BERT を利用して，QA タスクにおいて質問とそれに対応する応答見つけるために，質問と文書を入力に文書の中から応答のスパンを予測した. Li ら [2] は NER のタグに対応する質問文を生成し，その質問と文を入力とすることで entity のスパンを予測することで Nested NER の問題に対応した．Wu ら [1] は参照先をとうような質問を生成することでを共参照解析として捉えた手法を提案した.

3 スパン予測を用いたゼロ代名詞同定

本研究では，入力文における動詞や形容詞をクエリとし，クエリを主辞とする節にゼロ代名詞が存在するか，また，節に対応するスパンを予測するタスクとして定式化する．入力文 $\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}$ とすると，とその文内のクエリー x_q を与え，そのクエリを主辞とする節に対応するスパン $\{x_i, \ldots, x_j\}, 1 \leq i \leq j \leq n$ を予測する.

例えば，「宿題をすると私は決意した.」を入力文，動詞である「する」がクエリーとして与えられた場合，主語が省略されている節のスパン「宿題をする」を予測する．一方で「決意」がクエリーの場合，これを主辞とする節「~ と私は決意した.」には省略がないため，クエリを主辞とする節にはゼロ代名詞がないと予測する.

定式化したタスクに対して，BERT[4] を用いた.

3.1 スパン予測を用いたゼロ代名検出

入力はこのクエリー x_q とそのクエリーが含まれる文 $\boldsymbol{X}$ である．具体的には二つの入力を特殊文字 $[CLS]$ と $[SEP]$ を用いて，次のよう形式で BERT に与える.. $\{[CLS], x_q, [SEP], x_1, x_2, \ldots, x_n\}$ ここで $[SEP]$ を基準にこれらの入力を基に，$[SEP]$ 以降の i 番目のトークンに対して，スパンの開始確率 S_i とスパンの終了確率 E_i を予測する．$[CLS]$ の位置をゼロ代名詞がない場合に対応させ，$S_{[CLS]}$ と $E_{[CLS]}$ は，与えらた文にゼロ代名詞がない確率を表す．ス

パンの開始確率 $S_i \in \mathbb{R}$ は Devlin らの SQuAD v2.0 の手法 [4] と同様，次のように BERT の Encoder の最終層の埋め込みである $T_i \in \mathbb{R}^n$ を線形層に通して計算される．

$$S_i = softmax(T_i {W_S}^\top + b_S) \tag{1}$$

ここで重み $W_S \in \mathbb{R}^{1 \times n}$, バイアス $b_S \in \mathbb{R}$ である．また，終了確率 E_i も同様に求める．予測の開始位置と終了位置が $1 \leq i \leq j \leq n$ の時，これらの確率を用いてスパンのスコア定義は次のようにする．

$$Score_{S_i,E_j} = \log S_i + \log E_j \tag{2}$$

このスコアが最大となる区間にゼロ代名詞が存在すると予測する．一方で，文中にゼロ代名詞がないスコアは [CLS] に対応する $Score_{null} = Score_{S_{[CLS]},E_{[CLS]}}$ とする．そのため，$max_{j \geq i} Score_{S_i,E_j} < Score_{null}$ のときは，与えられた文にはゼロ代名詞がないと予測する．

学習時の損失関数はクロスエントロピー を用いる．そのため，正解の開始位置と終了位置が i', j' のとき，次のように対数尤度の和として定式化できる．

$$loss_{span} = -\log S_{i'} - \log E_{j'} \tag{3}$$

3.2 ゼロ代名詞クラス分類

3.1 節ではゼロ代名詞検出の二値分類であったが，ゼロ代名詞同定を行うためにクラス分類を行う[1)]．$[CLS]$ に対応する BERT の Encoder の最終層の埋め込み $T_{-1} \in \mathbb{R}^n$ はクエリーと文の二つの入力が考慮された埋め込みとされている．

この T_{-1} を用いて，ゼロ代名詞の num_{label} 種類のクラスに対する確率を求める．重み $W_{label} \in \mathbb{R}^{2 \times d}$ とバイアス $b_{label} \in \mathbb{R}^{num_{label}}$ を用いて次のように表せる．

$$H = dropout(T_{-1}) \tag{4}$$

$$P_{label} = softmax(H {W_{label}}^T + b_{label}) \tag{5}$$

学習ではクロスエントロピーロスと正解のゼロ代名詞クラスを用いて，損失 $loss_{label}$ 計算する．

3.3 同時学習によるゼロ代名詞同定

本研究では，3.1 節と 3.2 節のそれぞれのタスクを同時に学習することで，ゼロ代名詞の同定を行う．学習時の損失関数は次のように定義される．

$$loss_{total} = \alpha loss_{span} + (2 - \alpha) loss_{label} \tag{6}$$

1) 句構造文法の用語を借りれば，スパンは述語の最大投射，クラスは述語の下位範疇に相当する．

	文	ゼロ代名詞有 (件数)	ゼロ代名詞無
train	28764	29754 (46.9%)	34910 (54.0%)
dev	3595	3792 (46.6%)	4346 (53.4%)
test	3498	3864 (46.8%)	4370 (53.1%)
all	35956	37410	43626

表 1 データセットの内訳

ここで，α は $0 < \alpha < 2$ の間の値を取ることで，各タスクの損失関数に重みをつけるハイパーパラメータである．3.1 節のスパン予測にて省略が存在するとされたものを，3.2 節のタスクで予測したラベルのゼロ代名詞が省略されているとする．一方で，3.1 節のスパン予測にて省略がないと予測された場合は，このクエリを主辞と仮定したゼロ代名詞がないとする．

4 実験

4.1 実験データ

本実験のデータとして NINJAL Parsed Corpus of Modern Japanese（NPCMJ）[3] を用いる．NPCMJ はゼロ代名詞の情報を持つ欅ツリーバンク [11] を拡張したもので，青空文庫やニュースなどから成る約 4 万文に対して句構造がアノテーションされたコーパスである．ゼロ代名詞は，述語を主辞とする節 (inflectional phrase, IP) の最初の構成要素としてアノテーションされる．このアノテーションされた構文情報を用いて主辞が動詞 (VB) または形容詞 (ADJN) であり，ゼロ代名詞の有無にかかわらず全ての節を対象にクエリーを作成する．データセットは文単位で train : dev : test = 8:1:1 で分割を行う．データセットの各述語におけるゼロ代名詞の有無の件数と割合は表 1 のとおりである．

実験では，NPCMJ でアノテーションさているゼロ代名詞の種類のうち，$*pro*$, $*speaker*$,$*hearer*$,$*T*$ を用い，それ以外は **etc** に置き換える．構文情報のアノテーションのうち，SBJ に関するものは **SBJ**，OB1 や OB2 などの OBJ に関するものは **OBJ**，それ以外は **Other** タグに変換する．また，ゼロ代名詞がある述語 37410 件のうち，一つの述語に対して三つ以上ゼロ代名詞が含まれるものは 1% 未満である 80 件であった．そのため，今回の実験では一つのスパンに一つもしくは二つのゼロ代名詞の組み合わせをゼロ代名詞クラスとする．三つ以上の物に関しては

		Gold		
		有	無	正答率
Predict	有	2845 (34.6%)	343 (4.2%)	89.2%
	無	1019 (12.4%)	4027 (48.9%)	79.8%
	正答率	73.6%	92.2%	

表 2　ゼロ代名詞検出の混同行列 (件数)

etc に置き換える．

4.2　モデル設定

事前学習言語モデルとして，NICT-BERT[2] の BPE なし BERT$_{BASE}$ モデルを使用し，述語をクエリ，文を文脈として，述語を主辞とする節のスパンとゼロ代名詞の種類を予測するデータでファインチューニングした．

各パラメータ設定において，学習エポックは {2, 4}，バッチサイズは 16, 学習率は 3e-5，ドロップアウト確率は 0.1，式 6 のハイパーパラメータ α は 0.0 ~ 1.8 のうち 0.2 間隔で用いた．単語分割は NICT-BERT に合わせ，MeCab-Juman 辞書で単語分割を行った．

4.3　比較手法

ゼロ代名詞同定の実験において，BERT を用いた系列ラベリングを比較手法とする．Devlin ら [4] の手法を参考に，文全体を入力とし，BIOES 形式で述語にはゼロ代名詞クラス，それ以外は"O"をラベリングするように学習した．正誤判定は提案手法のクエリーに対応する述語の箇所のみに適切なラベリングを行ったかで判定した．

4.4　実験 1: ゼロ代名詞検出

3.1 節で述べた 質問応答に基づくゼロ代名詞検出に関する実験を行った．「ゼロ代名詞の有」かつ「そのゼロ代名詞の存在する節の完全一致」の時，True Positive (TP) とする．表 2 はエポック 2 における混同行列である．各セルには件数と割合を記載した．表 2 において，Gold データに対する True Negative (TN) の正答率が TP と比べて高い理由としては，TN は文中にゼロ代名詞が存在しない場合であるため，「節 (span) の一致」は考慮する必要がなく，「ゼロ代名詞の有無」の一致のみであるからと考えられる．

2)　https://alaginrc.nict.go.jp/nict-bert/index.html

Epoch	F1	Precision	Recall
2	80.6%	89.2%	73.6%
4	**81.3%**	89.1%	74.7%

表 3　ゼロ代名詞検出の F1 値

モデル	F1	precision	recall
系列ラベリング (epoch:2)	71.0%	87.6%	59.7%
系列ラベリング (epoch:4)	**74.3%**	87.8%	64.3%
zero_multi (α = 1)(epoch:2)	73.5%	74.3%	72.7%
zero_multi (α = 0.2)(epoch:2)	74.1%	74.8%	73.4%
zero_multi α = 1(epoch:4)	76.2%	76.7%	75.8%
zero_multi α = 0.4(epoch:4)	**76.5%**	77.0%	76.0%

表 4　ゼロ代名詞同定の F1 値

また，表 3 はエポック {2, 4} における F1 値，Precision,Recall を載せる．Keyaki tree bank に対しての構文的素性を用いた Takeno ら [10] の F1 値が 73.2% であり，データセットは異なるため，単純な比較はできないが，これらの結果を踏まえても，表 3 の 81.3% は十分に高いスコアと言える．

4.5　実験 2: ゼロ代名詞同定

3.3 節の同時学習を用いてゼロ代名詞の同定に関する実験を行った．TP を「ゼロ代名詞の種類」が一致した場合とする．比較手法としては 4.3 の系列ラベリングを用いる．F1 値による比較において，結果を表 4 に記す．表 4 よれば，ベースラインよりも，提案手法の方が 2.2% 優れていた．ただし提案法はクエリとなる述語の正解を与えているので，厳密な比較ではないことに注意する必要がある．

また，式 6 のハイパーパラメータ α を変更した時の F1 の変化を図 1 に記す．図 1 より，エポック 2,4 共に，ゼロ代名詞クラス分類に重心を強くして学習させるほうが，F1 値の精度が高かった．

5　エラー分析

提案手法の成果を確認するため，実際の事例を示す．下線部がゼロ代名詞の存在する節である．

(1)衝撃 や 水 濡れ に 強い の が ウリ の 機種 に 決め ました .
Gold : *pro_SBJ*　Predict : *pro_SBJ*

(2)偶然 が もたらした 大 発見 と いって よい であろう .
Gold : *T_OBJ*　Predict : *T_OBJ*

例文 (1),(2) はゼロ代名詞の種類と主格/目的格を正確に捕らえらている．
一方で誤り例を示す．

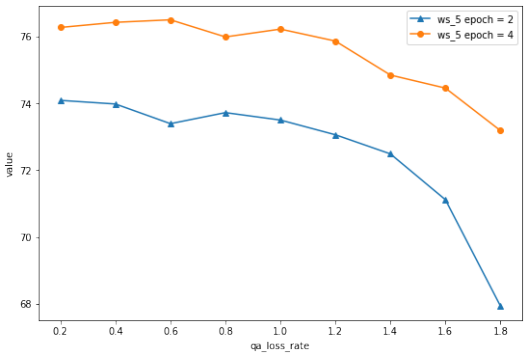

図 1　実験 2 において，α を変化させたときの F1 値

	pro-SBJ	speaker-SBJ	hearer-SBJ
正解率	51.7%	48.9%	54.8%
種類エラー率	10.1%	36.8%	38.2%

表 5　*pro* とその亜種の種類エラー率

(3)たまには おごれ よ .

Gold : *hearer_SBJ*　Predict : *speaker_SBJ*

この場合，SBJ は認識できているが，二人称 (*hearer*) であるべきところを一人称 (*speaker*) と誤っている．このように主格/目的格は正確に捕らえられるが，*pro* とその亜種である *hearer*,*speaker* に関して区別できていないことがある．*pro* と比べ，出現頻度低い *speaker* や *hearer* はこの傾向が顕著である．

表 5 は正解ラベルが SBJ に関するもの正答率と *pro* と亜種の種類エラーにおける割合である．正答率において，亜種ラベルのない *T-SBJ* が 3.7% と比べて，*pro* とその亜種である *speaker* と *hearer* は正答率が大きく下がっている．例文 (3) のような亜種に関するエラーが正答率を下げてる．このような亜種のエラーの割合は，"pro"のエラーのうち 10.1%，"speaker"と"hearer"のそのようなエラーの割合は 35% 後半もある．"pro"と比べ，"speaker"と"hearer"がエラーが多い理由として, それらの付与されたデータ数の少なさ，および，提案法は文を入力単位としており，文脈情報を利用していないことが考えられる．

6　おわりに

本研究では QA タスクにおけるスパン予測問題として定式化したゼロ代名詞の検出・種類同定を提案し，NPCMJ を用いて評価を行った．今後は推定した情報を基にゼロ照応解析や他言語のデータセットに対する評価などを行っていきたいと考えている．今回の提案手法では，一つの節に対して一つのクエリで推定を行っていたが，推定したゼロ代名詞のスパンを用いて，節と節の関係について文全体で整合性をとることで，精度の向上が期待できる．

参考文献

[1]Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6953–6963, Online, July 2020. Association for Computational Linguistics.

[2]Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5849–5859, Online, July 2020. Association for Computational Linguistics.

[3]アラステアバトラー, 吉本啓, 岸本秀樹, プラシャントパルデシ. 統語・意味解析情報付き日本語コーパスのアノテーション. 言語処理学会 第 22 回年次大会 発表論文集, 3 月 2016.

[4]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5]植田暢大, 河原大輔, 黒橋禎夫. Bert と refinement ネットワークによる 統合的照応・共参照解析. 言語処理学会 第 26 回年次大会 発表論文集, 3 月 2020 年.

[6]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 *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4956–4968, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[7]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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5429–5434, Online, July 2020. Association for Computational Linguistics.

[8]Abdulrahman Aloraini and Massimo Poesio. Anaphoric zero pronoun identification: A multilingual approach. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pp. 22–32, Barcelona, Spain (online), December 2020. Association for Computational Linguistics.

[9]Abdulrahman Aloraini and Massimo Poesio. Cross-lingual zero pronoun resolution. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 90–98, Marseille, France, May 2020. European Language Resources Association.

[10]Shunsuke Takeno, Masaaki Nagata, and Kazuhide Yamamoto. Empty category detection using path features and distributed case frames. In *Proceedings of the 2015 Confer-*

ence on Empirical Methods in Natural Language Processing, pp. 1335–1340, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[11] Alastair Butler, Tomoko Hotta, Ruiko Otomo, Kei Yoshimoto, Zhen Zhou, and Hong Zhu. Keyaki treebank : phrase structure with functional information for japanese., 2012.