G-23

BERT を用いた指示詞の照応関係の推定手法の検討

大和 秀徳†　　岡田 真†　　森 直樹†

1. はじめに

近年，機械学習技術の大きな発展により，自然言語処理においても様々なタスクに機械学習が用いられるようになった．自然言語処理において注目されているタスクの 1 つとして対話システムがある．対話システムは人間と人工知能が対話をするタスクであり，人工知能には自然な応答が求められる．

対話システムを構築する上で課題となるタスクとして，日本語や中国語のような単語間の区切りが存在しない文章を単語に分割する形態素解析，文章内の単語間の係について解析する構文解析，代名詞や指示詞などといった照応詞の指示内容の推定やゼロ代名詞と呼ばれる省略された名詞句を補完する照応解析などが挙げられる．これらのタスクは人工知能が自然言語を理解する上で非常に重要である．形態素解析や構文解析については高い精度で解析されているが，照応解析については文章の構造や単語の表象的な情報のみならずそれらの意味や文脈情報まで考慮する必要があるため，形態素解析や構文解析と比較すると精度は高くなく，未だに難しいタスクとされている．

そこで本研究では自然言語処理における指示詞の内容理解を目的として， BERT を用いた指示詞の先行詞の推定手法の検討をする．

2. 指示詞

指示詞とは，物事を指し示す機能をもつ語のことである．益岡らは指示詞を機能ごとに名詞形態指示詞，連体詞形態指示詞，副詞形態指示詞の 3 種類に大別した［1］．名詞形態指示詞は 「これ，それ，あれ，どれ」といったそれ自体が名詞的な働きをする指示詞である．連体詞形態指示詞は 「この，その，あの，どの」といった体言に接続する指示詞である．副詞形態指示詞は 「こう，そう，ああ，どう」といった物事の程度や状態を示す指示詞である．

本研究では，物事自体を指示対象とする名詞形態指示詞および連体詞形態指示詞を先行詞推定の対象とした．また，連体詞形態指示詞については接続する体言についても指示詞の範囲とした．

3. 要素技術

3.1 BERT

Bidirectional Encoder Representations from Transformer (BERT) [2] は 2018 年に Google が発表した Transformer による双方向エンコーダによって学習された言語モデルである．文章分類，質問応答，固有表現抽出等の多様なタスクで公開当時の最高性能を達成するといった成果が報告されている．従来の言語モデルでは，特定のタスクに対して 1 つの言語モデルを用いたが， BERT は事前学習したものを転移学習によってファインチューニングすることで複数のタスクに対応することができる．本研究では，東北大学の乾研究室が公開している日本語 Wikipedia をもとに事前学習をしたモデル[1]を用いた．

3.2 形態素解析

形態素解析とは，自然言語で表現された文章を言語上で意味を持つ単語の最小構成要素である形態素に分割する技術である．自然言語処理では，形態素解析によって得られた形態素を入力として様々なタスクをする．本研究では，形態素解析ツールとして MeCab [3] を用いた．

4. 京都大学ウェブ文書リードコーパス

京都大学ウェブ文書リードコーパス（Kyoto University Web Document Leads Corpus， KWDLC）［4，5，6］は，さまざまなウェブ文書のリード（冒頭）3 文に各種言語情報を人手で付与したテキストコーパスである．コーパスの規模は約 5,000 文書で，言語情報として形態素，固有表現，構文・格関係,照応・省略関係，共参照関係，談話関係などの情報が付与されている．

本研究では，指示詞において共参照関係 (同一の物事を指し示す関係) に関する情報が付与されているものをデータセット作成に用いた．

5. 数値実験

5.1 実験 1

本研究の目的は指示詞の先行詞の推定である．実験 1 では予備実験として与えられた文章内に存在する名詞形態指示詞および連体詞形態指示詞に対して先行詞が存在するかどうかの 2 値分類をした．本研究では文章内に指示詞が 1 つだけ存在しており，かつ指示詞に対して KWDLC において共参照関係にあることを示すタグ 「“=”， “=構”，“=≒”，“=構≒”」が付与されているものが存在する場合は先行詞が存在するとして正例ラベルを，そうでない場合は先行詞が存在しないとして負例ラベルを付与したデータセットを作成した．また，正例と負例のデータ数は 1 : 1 になるように調整した．表 1 に実験で用いたデータ数を示す．

入力文章 S を形態素解析器 MeCab を用いて分かち書きし，単語列 $W = \{w_1, w_2, \ldots, w_n\}$ を得た．ここで， n は単語列の長さである．そして BERT への入力に必要な “[CLS]” トークンを W の先頭に挿入し，“[SEP]”トークンを W の末尾に挿入し W' とした．その後，各データの単語列の長さを揃えるために単語列の長さが最長のものを n_{MAX} として，単語列の長さが n_{MAX} に満たないものは 末尾に “0” をパディングした．これを BERT へ入力して得られた “[CLS]” トークンの分散表現 $E_{[\mathrm{CLS}]}$ を文章の分散表現として獲得し，識別器である 3 層 Multilayer Perceptron (MLP) へ入力することで得られた出力のうち，1 を正例，0 を負例として識別した．モデルの訓練時，BERT のファインチューニングを最終層のみにした．図 1 にモデルの概略図を，表 2 に学習パラメータを示す．また，全データを訓練データとテストデータに 9 : 1 で分割し，訓練データについて 5 分割交差検証をし，テストデータを用いてモデルの評価をした．

† 大阪府立大学

[1] https://github.com/cl-tohoku/bert-japanese

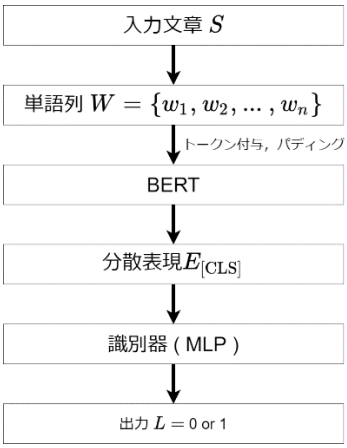

図 1:モデルの概略図 (実験 1)

表 1:データ数

正例	負例	合計
421	421	842

表 2: 学習パラメータ (実験 1)

次元数 (MLP 入力層)	768
次元数 (MLP 中間層)	512
次元数 (MLP 出力層)	2
学習率	5×10^{-5}
ドロップアウト率	0.2
最適化手法	Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)
損失関数	Cross Entropy Loss
活性化関数 (MLP 中間層)	ReLU
Epoch 数	30
バッチサイズ	64

5.2 実験 2

実験 2 では与えられた文章内に対して先行詞に対応する単語に対してラベル 1 を，そうでない単語に対してラベル 0 を出力する系列ラベリングをした．本研究では文章内に指示詞が 1 つだけ存在し，かつその指示詞に対して先行詞が存在する文章を実験の対象とする．また，先行詞に相当する単語は指示詞に対して KWDLC において共参照関係にあるという情報が与えられている単語のみとし，単語列の長さが 80 以上のものは実験データから除いた．実験で用いた文章数は 393 文章であった．

実験 1 と同様にして入力文章 S を形態素解析器 MeCab を用いて分かち書きし，単語列 $W = \{w_1, w_2, \dots, w_n\}$ を得た．そして BERT への入力に必要な “[CLS]” トークンを W の先頭に挿入し，“[SEP]” トークンを W の末尾に挿入し W' とした．このとき，各データの系列長を揃えるために単語列W'の長さが最長のものを n_{MAX} として，単語列の長さがn_{MAX} に満たないものは末尾に “0” をパディングした．これを BERT へ入力して得られた “[CLS]” トークンの分散表現 $E_{[\mathrm{CLS}]}$ を文章の分散表現として獲得し，識別器である 3 層 MLP へ入力することで，n_{MAX}と各種トークンを合わせた 82 次元の出力 $L = \{l_{[\mathrm{CLS}]}, l_1, \dots, l_{n_{\mathrm{MAX}}}, l_{[\mathrm{SEP}]}\}$ を得る．このとき，l_i ($i = [\mathrm{CLS}], 1, \dots, n_{\mathrm{MAX}}, [\mathrm{SEP}]$) は単語 w_i が指示詞の先行詞であると推測した場合に正例ラベル 1 を出力し，先行詞でないと推測した場合に負例ラベル 0 を出力する．モデルの訓練時，BERT のファインチューニングを最終層のみにした．図 2 にモデルの概略図を，表 3 に学習パラメータを示す．また，全データを訓練データとテストデータに 9 : 1 で分割し，訓練データについて 5 分割交差検証をし，テストデータを用いてモデルの評価をした．

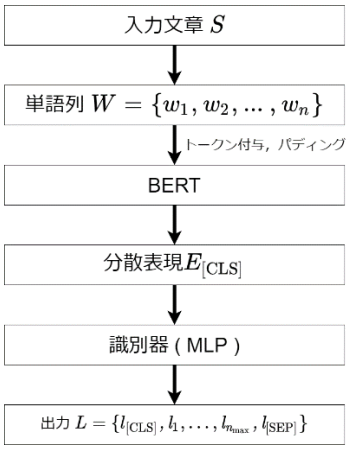

図 2:モデルの概略図 (実験 2)

表 3: 学習パラメータ (実験 2)

次元数 (MLP 入力層)	768
次元数 (MLP 中間層)	512
次元数 (MLP 出力層)	82
学習率	5×10^{-5}
ドロップアウト率	0.2
最適化手法	Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)
損失関数	BCEWithLogitsLoss
活性化関数 (MLP 中間層)	ReLU
Epoch 数	30
バッチサイズ	32

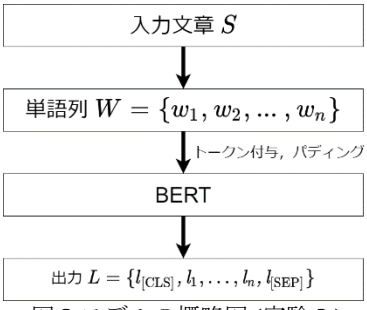

図 3:モデルの概略図 (実験 3)

表 4:学習パラメータ (実験 3)

学習率	3×10^{-5}
ドロップアウト率	0.1
最適化手法	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)
損失関数	Cross Entropy Loss
Epoch 数	10
バッチサイズ	32

表 5:各実験結果

	精度	再現率	F1 値
実験 1	0.634 ± 0.043	0.728 ± 0.045	0.676 ± 0.014
実験 2	0.076 ± 0.019	0.026 ± 0.014	0.036 ± 0.019
実験 3	0.652 ± 0.055	0.362 ± 0.040	0.464 ± 0.049

5.3 実験 3

実験 3 では 指示詞の先行詞推定を単語のトークン識別問題とみなして推定をした．モデルの実装に Transformers[2] の BertForTokenClassification を用いた．実験データは実験 2 と同様の 393 文章を用いた．BERT への入出力の形式は実験 2 と同様であるが，出力のサイズは入力文章の単語列長に各種トークンを加えたものとなる．モデルの訓練時，BERT のファインチューニングを全層にした．図 3 にモデルの概略図を，表 4 に学習パラメータを示す．また，全データを訓練データとテストデータに 9 : 1 で分割し，訓練データについて 5 分割交差検証をし，テストデータを用いてモデルの評価をした．

6. 結果と考察

6.1 実験 1

表 5 にテストデータにおける正例の精度，再現率，F1 値を示す．また， Accuracy は 0.639 ± 0.048 となった．すべて正例と推定した場合をベースラインとしたときの Accuracy である 0.5 上回った．図 4，5 に交差検証時の 全体の Accuracy および Loss の平均の推移を示す．青線が訓練時の推移を，緑色の破線が検証時の推移を示している．また，図 4 は縦軸に全体の Accuracy を，横軸に Epoch 数を示しており，図 5 は縦軸に Loss を，横軸に Epoch 数を示している．図 5 から分かるように，検証時の Loss が 10 Epoch 付近から上昇しており過学習が見られた．表 6 に最も F1 値の高かったモデルの混同行列を示す．実験 1 では出力のラベルが正例であるか負例であるかの 2 値分類をしたため，ある種の文章分類になってしまったと考えられる．また，Accuracy が訓練時とかけ離れている原因としては，指示詞に対する先行詞の有無というラベリングが非常に恣意的なもので，各データ間に普遍的な特徴があまり存在しないからであると考えられる．

6.2 実験 2

表 5 にテストデータにおけるラベル 1 の精度，再現率，F1 値を示す．図 6，7 に交差検証時の全体の Accuracy および Loss の平均の推移を示す．青線が訓練時の推移を，緑色の破線が検証時の推移を示している．また，図 6 は縦軸に全体の Accuracy を，横軸に Epoch 数を示しており，図 7 は縦軸に Loss を，横軸に Epoch 数を示している．表 7 に最も F1 値が高かったモデルの混合行列を示す．表 7 からわかるように，ラベル 1 をほとんど間違えてしまっている．これは，“[CLS]” トークンの分散表現のみを識別に用いた場合，識別に単語の意味や単語の位置といった単語自体の情報をうまく活用できていないことが原因であると考えられる．

6.3 実験 3

表 5 にテストデータにおける正例ラベル 1 の精度，再現率，F1 値を示す．図 8，9 に交差検証時の全体の Accuracy および Loss の平均の推移を示す．青線が訓練時の推移を，緑色の破線が検証時の推移を示している．また，図 8 は縦軸に全体の Accuracy を，横軸に Epoch 数を示しており，図 9 は縦軸に Loss を，横軸に Epoch 数を示している．表 8 に最も F1 値が高かったモデルの混合行列を示す．識別器に MLP を用いた実験 2 の系列ラベリングと比較すると良い結果が得られた．これは，各単語自体の情報を識別に用いることができたからであると考えられる．このことから，指示詞の先行詞推定に単語の情報を用いることの有効性を確かめることができた．

[2]https://github.com/huggingface/transformers

7. まとめと今後の課題

本研究では，指示詞の照応関係の推定として，3 つの実験をした．実験 1 で BERT から得られる分散表現を用いて与えられた文章内に存在する指示詞に対して先行詞が存在するかどうかの 2 値分類をし，ベースラインを超える Accuracy を得ることができた．実験 3 では BERT をトークン識別問題としてファインチューニングすることで指示詞の先行詞推定をし，実験 2 よりも高い精度で識別することができた．今後の課題として，単語の素性や単語間の関係を考慮した推定が挙げられる．また将来的には指示詞の内容を踏まえたより自然な応答文の生成や，画像情報と組み合わせた指示詞の先行詞推定に取り組みたいと考えている．

謝辞

なお，本研究は一部，日本学術振興会科学研究補助金基盤研究（B）（課題番号 19H04184）の補助を得て行われたものである．また，本研究は一部，日本学術振興会科学研究補助金基盤研究（C）（課題番号 20K11958）の補助を得て行われたものである．

参考文献

[1] 益岡隆志, 田窪行則. 基礎日本語文法・改訂版. くろしお出版, 1992.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[3] 工藤拓, 山本薫, 松本裕治. Conditional random fields を用いた日本語形態素解析. 情報処理学会研究報告. NL, 自然言語処理研究会報告, 第 161 巻, pp. 89-96. 一般社団法人情報処理学会, may 2004.

[4] 萩行正嗣, 河原大輔, 黒橋禎夫. 多様な文書の書き始めに対する意味関係タグ付きコーパスの構築とその分析. 自然言語処理, 第 21 巻, pp. 213-248, 2014.

[5] Daisuke Kawahara, Yuichiro Machida, Tomohide Shibata, Sadao Kurohashi, Hayato Kobayashi, and Manabu Sassano. Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. In *In Proceedings of the 25th International Conference on Computational Linguistics*, pp. 269–278, 2014.

[6] Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. Building a diverse document leads corpus annotated with semantic relations. In *In Proceedings of the 26th Pacific Asia Conference on Language Information and Computing*, pp. 535–544, 2012.

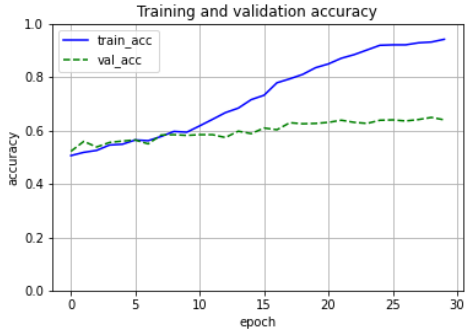

図 4: Accuracy の推移 (実験 1)

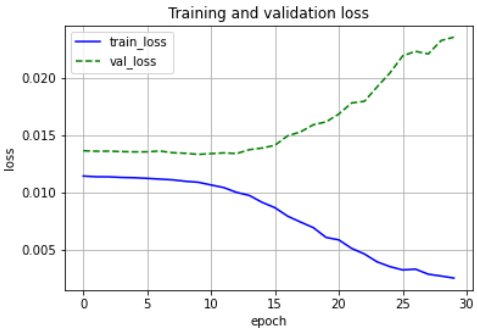

図 5: Loss の推移 (実験 1)

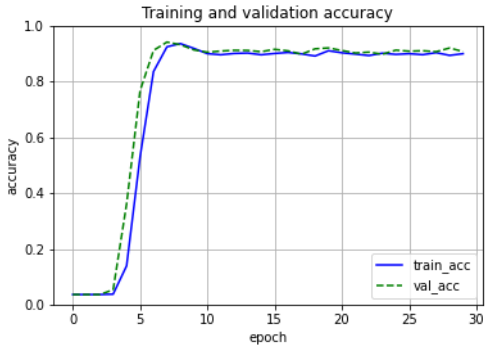

図 6: Accuracy の推移 (実験 2)

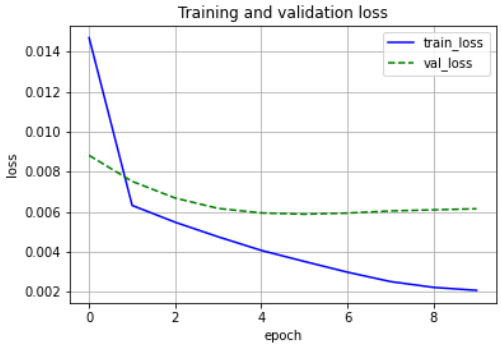

図 7: Loss の推移 (実験 2)

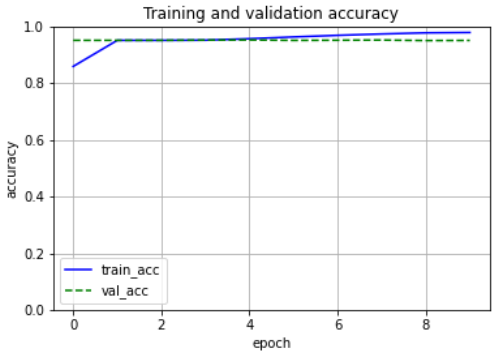

図 8: Accuracy の推移 (実験 3)

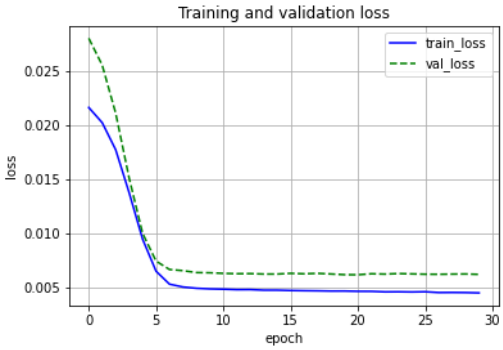

図 9: Loss の推移 (実験 3)

表 6 : 混同行列 (実験 1)

		予測	
		負例	正例
正解	負例	30	13
	正例	13	19

表 7 : 混同行列 (実験 2)

		予測	
		負例	正例
正解	負例	3058	70
	正例	146	6

表 8 : 混同行列 (実験 3)

		予測	
		負例	正例
正解	負例	2165	23
	正例	91	61