BERT を用いた二つの辞書の対応付け

河野稜斗
東京農工大学工学部情報工学科
s209679x@st.go.tuat.ac.jp

平林照雄
東京農工大学生物システム応用科学府
s213645z@st.go.tuat.ac.jp

古宮嘉那子
東京農工大学工学研究院
kkomiya@go.tuat.ac.jp

概要

辞書間における対応付けとは，複数の辞書にタグ付けされたコーパスを利用し，辞書同士の対応をとることである．本研究では，事前学習モデルの BERT を用いて，『岩波国語辞典』の語義と『分類語彙表』の分類番号の対応付けを行う．BERT を用いた対応付け方法として，BERT を使用した教師なしの手法と，BERT の fine-tuning を使用した手法の 2 つを提案する．実験では，2 つの提案手法と比較手法として最多出現語義の F 値を算出し対応の性能を評価した．実験結果として，提案手法の両方で比較手法の性能を上回る結果となった．

1 はじめに

世の中には国語辞典だけでも多くの種類の辞書が存在するが，それぞれの辞書で語義やその粒度が一致しているとは限らない．辞書同士で語義の対応をとることにより，語義を整理することができると考える．また，辞書同士の対応を人の手作業で行った場合，辞書の数が多いほどかなりの時間と人手が必要になる．そのため，自動的に対応をとることが可能であれば，人員や時間の削減につながる．

そこで本研究では，対応付けのアプローチ方法に事前学習モデルの BERT を採用し，『岩波国語辞典』と『分類語彙表』によりタグ付けされた『現代日本語書き言葉均衡コーパス』[1, 2, 3] を利用して，語義タグと分類番号の対応をとる．BERT を用いた手法として，BERT を使用した教師なしの手法と BERT を fine-tuning した手法を提案する．また，実験では F 値を対応付けの性能の指標とし，『岩波国語辞典』の語義と『分類語彙表』の分類番号の対応付けの評価を行う．

2 関連研究

近年，事前学習モデルが自然言語処理タスクで最先端の性能を達成している．その中でも，2019 年に Devlin ら [4] によって考案された Bidirectional Encoder Representations from Transformers（BERT）という事前学習モデルが自然言語処理のタスク 11 種類で最高の性能を達成し，広く使われている．BERT は Vaswani ら [5] によって提案された Transformer を双方向に接続することで構成されている．Transformer は従来の一般的なニューラルモデルで使われる Encoder-Decoder 形式の畳み込みニューラルネットワークや再帰ニューラルネットワークを用いず，注意機構のみで構成されたモデルである．現在では，この BERT をもとにした ELECTRA[6] や RoBERTa[7]，ALBERT[8] などのモデルも多く提案されている．

対応付けに関する研究として，Bilingual Word Embeddings（BWE）の単一言語マッピングの手法を用いた研究がある．単一言語マッピングの手法は，あらかじめ 2 言語それぞれで単語の分散表現を作成しておき，言語間で意味が似ている分散表現が近づくように共通ベクトル空間に変換することで対応をとる手法である．Mikolov[9] らは，幾何学的関係が言語間で類似していることを主張し，線形射影によってある言語のベクトル空間を別の言語のベクトル空間に変換することの可能性を示唆した．さらに，平林ら [10] は，BWE の単一言語マッピングの手法を用いて，本研究と同じ岩波国語辞典の語義タグと分類語彙表の分類番号の対応付けを行った．語義タグのベクトル空間から分類番号のベクトル空間への線形変換を学習した後，語義タグの分散表現に学習した線形変換を適用し，分類番号のベクトル空

間に変換することで対応をとった．また，単語の分散表現の作成には言語ごとに学習した word2vec を利用している．

3 提案手法

本研究では，『岩波国語辞典』の語義 (以下語義タグと略す) と『分類語彙表』の分類番号 (以下分類番号と略す) を対応付けするために BERT を用いて，「暫定の対応」をとる．そして，「暫定の対応」を利用することで「確定の対応」をとる．

3.1 岩波国語辞典の語義タグ

『岩波国語辞典』では，単語の語義ごとに表 1 に示すような語義タグが付与されている．語義タグは，[見出し ID]-[複合語 ID]-[大分類 ID]-[中分類 ID]-[小分類 ID] で構成されている．例として，「手」という単語は表 1 のような意味ごとに異なる語義タグが付与されている．また，本研究において複合語を無視しているため複合語 ID の値は必ず 0 である．

表 1 岩波国語辞典における「手」

語義タグ	意味
34626-0-1-1-0	<1>人体の，方から先にある部分.
34626-0-1-1-1	<ア>人（擬人的に動物）の両肩から出た部分の全体.「―を広げる」
34626-0-1-2-0	手 (1) に関係する次のもの.
34626-0-1-2-1	方法．手段や策略.
⋮	⋮

3.2 分類語彙表の分類番号

『分類語彙表』とは，国立国語研究所によって作成された，「語を意味によって分類や整理した類義語集」である．『分類語彙表』のレコード (1 つのデータ) を構成する項目は，「レコード ID 番号/見出し番号/レコード種別/類/部門/中項目/分類項目/分類番号/段落番号/小段落番号/語番号/見出し/見出し本体/読み/逆読み」となっている．分類番号は，レコードを構成する項目のうち「類/部門/中項目/分類項目」によって表される．

1 つの単語が複数の分類項目に分類されることがあり，その単語は複数の分類番号を持つため多義語であるとみなせる．例えば．「手」という単語は表 2 に示すように「人間」としての意味，「手足・指」としての意味，「方法」としての意味などといった複数の分類項目に分類されるため多義語の 1 つである．

表 2 『分類語彙表』における「手」

文例	類	中項目	分類項目	分類番号
手が足りない	主体	人間	人間	1.2000
この**手**がある	活動	心	方法	1.3081
手を伸ばす	自然	身体	手足・指	1.5603

3.3 BERT を用いた暫定の対応付け

「暫定の対応」付けでは，BERT に語義タグもしくは分類番号が付与されている単語を含む文を入力し，タグが付いている単語の分散表現を取得し，対応付けに利用する．BERT を使用した教師なしの手法では，語義タグが付与された単語の分散表現と分類番号が付与された単語の分散表現として，BERT の出力ベクトルを使用し，コサイン類似度による語彙の対応をとる．BERT の fine-tuning の手法では，BERT にタグが付いた 2 文を入力し，(1) 語義タグが付与された単語と (2) 分類番号が付与された単語の BERT の出力ベクトルを連結して最終層の入力とし，(1)(2) が同じ意味かどうかの二値分類を行って fine-tuning を行った．

対応付けにおいて語義タグと分類番号の粒度の違いから，1 つの語義タグに対して複数の分類番号が対応する場合が存在する．対応付けを行っていくにあたり，BERT を使用した教師なしの手法と BERT の fine-tuning を用いた手法で得られる対応は「暫定の対応」となる．

3.4 確定の対応

「暫定の対応」を確定させるために「全部列挙」と「多数決」という 2 つの対応方法を用いる．

(I)全部列挙
語義タグに対応する可能性のある全ての分類番号を対応するものとして扱うことで，語義番号と分類番号の対応をとる方法である．

(II)多数決
語義タグに対応する可能性のある分類番号の中で，最も多く出現した分類番号を対応するものとして扱うことで，語義番号と分類番号の対応をとる方法である．また，「多数決」の手法においては，最も多く出現した分類番号が複数存在する (同率 1 位) 場合がある．その時は，最も多く出現した複数の分類番号全てを語義タグに対応するものとして扱う．

4 実験

本研究では，コーパスとして『現代日本語書き言葉均衡コーパス』(BCCWJ) を使用して実験を行う．BCCWJ には岩波国語辞典の語義と分類語彙表の分類番号が付与されているが，付与されている部分はごく一部を除き異なっている．また BERT は日本語の Wikipedia のデータを用いて学習された，東北大学が公開している BERT を用いる．実験では，東北大学の BERT を用い，前述の教師なしの手法と fine-tuning を利用した手法を用いて語義タグと分類番号の対応付けを行う．そして，比較手法として単語に最も付与された『分類語彙表』の分類番号を単語の各語義の分類番号とする最多出現語義 (MFS) を採用し，2 つの提案手法と比較する．

4.1 BERT の教師なしの手法の対応付け

BERT を使用した教師なしの手法を用いた対応付けでは，以下のステップに従って対応付けを行う．ここで用いる BERT は fine-tuning していない既存のモデルである．

1. 語義タグが付いている BCCWJ と，分類番号が付いている BCCWJ に分けて次の処理を行う．
   (a) BCCWJ の中から，語義タグ/分類番号が付いている単語を含む文を取得する．
   (b) 取得した文を分かち書きして BERT に入力し，語義タグ/分類番号が付いている単語の分散表現を取得する．
2. 語義タグが付いている単語と分類番号が付いている単語のうち，同じ単語を含む文の集合を作り，文ごとに語義タグが付いている単語を基準に分類番号が付いている単語に対して分散表現のコサイン類似度を求める．そして，コサイン類似度の値が最も大きい分類番号を，基準とした文の単語に付いている語義タグに対応する分類番号 (暫定) とする．
3. 2 で求めた対応は，1 つの文に含まれる語義タグの付いている単語に対しての対応であるため，語義タグが付いている単語を含む文の数だけその単語に対する対応が得られる．そのため，2 で求めた対応は暫定としている．
4. 最後に，「暫定の対応」を確定させるために「全部列挙」と「多数決」の方法を用いる．

4.2 fine-tuning した BERT による対応付け

BERT の fine-tuning の手法では，語義タグが付いている BCCWJ を用いて訓練データを作成し，BERT の fine-tuning を行い，テストデータで語義タグと分類番号が対応しているか否かの二値分類のタスクを解くことで対応付けを行う．この際，同じ単語に対して語義タグが付いた 2 つの文の組み合わせを訓練データとして fine-tuning を行った．学習時のパラメータは事前実験によって決定し，学習率 0.0001，epoch 数 30，最適化関数は SGD，損失関数はクロスエントロピー誤差を用いた．また，開発データおよびテストデータには，同じ単語に対して語義タグが付いた 1 文と分類番号が付いた 1 文の計 2 文の組み合わせを用いた．それに加えて，語義タグと分類番号の対応の正解データとして，『分類語彙表』と『岩波国語辞典第五版タグ付きコーパス 2004』の対応表 [11] を利用した．

訓練データ，テストデータ，開発データの BERT に入力する 2 文の組み合わせやデータの件数，2 つの文のタグが対応するかどうかの二値の内訳を表 3 に示す．タグ構成は，1 文目と 2 文目に含まれるタグが語義タグか分類番号なのかを表している．訓練データには語義タグ同士が同じ語義かどうかの情報のみを利用し，語義タグと分類番号の対応は利用していないことに注意されたい．また，表 3 において，入力する 2 つの語の意味が同じ場合は対応あり，異なる場合は対応なしとした．訓練データはある単語に付与された語義タグの種類ごとに最大 2 文の代表文をランダムに選択し，その組み合わせで作成する．付録の図 1 に「一緒」を例とした訓練データの作成例を示す．テストデータは単語ごとで単語に付与されている語義タグと分類番号の全ての組み合わせで作成する．

表 3 fine-tuning におけるデータの詳細

	タグ構成	対応あり	対応なし	合計
訓練	語義タグ/語義タグ	6886	14334	21220
テスト	語義タグ/分類番号	76549	114954	191503
開発	語義タグ/分類番号	3922	6078	10000

BERT がテストデータに対して出力する二値分類の値は，語義タグと分類番号の「暫定の対応」であるため，最終的な対応は BERT を使用した教師なしの手法と同様に「全部列挙」手法と「多数決」手法で求める．

4.3 対応付けの評価

本研究では，語義タグを基準にして対応する分類番号を定めているため，語義タグが付与された形態素のみに対しての「確定の対応」に限定して評価を行う．したがって．語義タグを付与された形態素が分類番号を付与された語彙素に存在しない場合，その語義タグに対する分類番号の対応は「存在しない」として扱うこととした．

評価には F 値を用いる．F 値は予測した結果の評価尺度の 1 つであり，適合率 (precision) と再現率 (recall) から算出される．

5 実験結果

BERT を使用した教師なしの手法と BERT の fine-tuning を用いた手法では，どちらも「暫定の対応」を取った後に「全部列挙」と「多数決」の方法で「確定の対応」を求め，それに対して F 値を算出した．BERT を使用した教師なしの手法の「全部列挙」と「多数決」に加えて，比較手法の MFS の対応付けの評価結果を表 4 に示す．そして，BERT の fine-tuning を用いた手法の「全部列挙」と「多数決」，および比較手法の MFS の対応付けの評価結果を表 5 に示す．表 4 と表 5 の太字の値の部分は，適合率，再現率，F 値それぞれで，最も高い値のものを示している．BERT を使用した教師なしの手法で最も高い F 値は「全部列挙」方法での 0.60，BERT fine-tuning を使用した手法で最も高い F 値は「多数決」方法での 0.63 という結果となった．それに対して比較手法の F 値は 0.50 であった．

表 4 BERT の教師なし手法での対応付けの結果

	全部列挙	多数決	MFS
適合率	0.60	**0.65**	0.53
再現率	**0.61**	0.54	0.47
F 値	**0.60**	0.59	0.50

表 5 BERT の fine-tuning 手法での対応付けの結果

	全部列挙	多数決	MFS
適合率	0.51	**0.58**	0.53
再現率	**0.77**	0.69	0.47
F 値	0.62	**0.63**	0.50

6 考察

実験の結果から，BERT を使用した教師なしの手法と BERT の fine-tuning を用いた手法の両方で，比較手法の MFS の結果を上回った．なお，それぞれの提案手法の「全部列挙」と「多数決」の F 値の差に大きな違いはなかった．

提案手法のどちらの結果も，適合率は「確定の対応」を「多数決」でとる方が「全部列挙」でとるよりも高くなっており，それに対して再現率は「確定の対応」を「全部列挙」でとる方が「多数決」でとるよりも高くなっている．その要因として，「全部列挙」では語義タグに対して「暫定の対応」に出現した分類番号全てを対応するものとして扱っているため対応する分類番号の種類が多い．また「多数決」では語義タグに対して「暫定の対応」に出現した分類番号の中で最も出現した分類番号を対応するものとして扱っており，対応する分類番号の種類は少ないことが挙げられる．そのため，「全部列挙」と「多数決」で求める対応は両極端になると考えられる．

したがって，「全部列挙」と「多数決」の中間をとる方法，例えば分類番号の出現回数に閾値を設け，一定回数出現した分類番号を対応するものとして扱うなどの方法で「確定の対応」をとることにより，さらに対応付けの性能が高まると考える．

また，本研究で扱った BERT よりもパラメータや次元が大きい BERT-large モデルや，BERT よりも多くのデータで事前学習し，かつ事前学習の回数を増やした RoBERTa モデルを使用することで対応付けの性能向上が見込まれる．

7 おわりに

本研究では，対応付けのアプローチ方法に事前学習モデルの BERT を採用し，『岩波国語辞典』と『分類語彙表』によりタグ付けされた『現代日本語書き言葉均衡コーパス』を利用して，語義タグと分類番号の対応をとった．BERT を用いた方法では，既存のモデルを使用する，BERT を使用した教師なしの手法と既存のモデルを fine-tuning した，BERT の fine-tuning の手法を提案し実験を行い評価を行った．実験結果から，BERT の fine-tuning が最も高い性能となり，提案手法は比較手法の MFS を上回る結果だった．以上のことから『岩波国語辞典』の語義と『分類語彙表』の分類番号の対応付けにおける，BERT の有効性を示した．

謝辞

本研究は JSPS 科研費 17KK0002，18K11421 の助成を受けたものです．

参考文献

[1] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written japanese. **Language resorces and evaluation**, Vol. 48, No. 2, pp. 345–371, 2014.

[2] Okumura Manabu, Shirai Kiyoaki, Komiya Kanako, and Yokono Hikaru. On semeval-2010 japanese wsd task. 自然言語処理, Vol. 18, No. 3, pp. 293–307, 2011.

[3] Sachi Kato, Masayuki Asahara, and Makoto Yamazaki. Annotation of 'word list by semantic principles' labels for the balanced corpus of contemporary written japanese. **Proceedings of PACLIC 2018**, pp. 247–253, 2018.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **In North American Association for Computational Linguistics(NAACL)**.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **In Advances in Neural Information Processing Systems**, pp. 6000–6010, 2017.

[6] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. **In International Conference on Learning Representations (ICLR)**.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. **ArXiv, abs/1907.11692**.

[8] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. **In International Conference on Learning Representations (ICLR)**.

[9] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. **arXiv preprint arXiv:1309.4168**, 2013.

[10] Teruo Hirabayashi, Kanako Komiya, Masayuki Asahara, and Hiroyuki Shinnou. Automatic creation of correspondence table of meaning tags from two dictionaries in one language using bilingual word embedding. **13th BUCC Workshop at LREC 2020**, pp. 22–28, 2020.

[11] 呉佩, 近藤森音, 森山奈々美, 荻原亜彩美, 加藤祥, 浅原正幸. 『分類語彙表』と『岩波国語辞典第五版タグ付きコーパス 2004』の対応表. 言語資源活用ワークショップ 2019, pp. 337–342, 9 2019.

付録

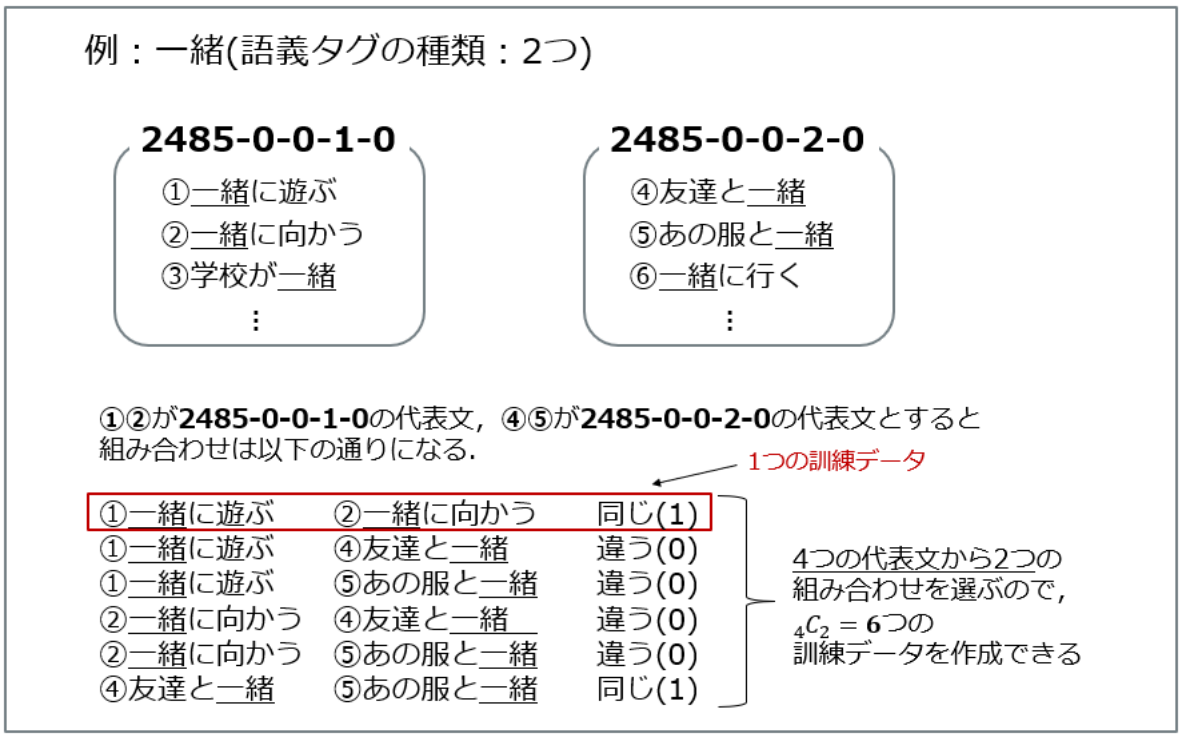

図 1　訓練データの作成例