BERT を利用した Zero-shot 学習による同音異義語の誤り検出

藤井 真　　新納 浩幸
茨城大学大学院 理工学研究科 情報工学専攻
{19nm727r, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

1 はじめに

本研究では，かな漢字変換や音声入力に起因する同音異義語の誤りについて，BERT[1] と Zero-shot 学習を組み合わせて検出する手法を試みる．

日本語文を電子的に入力する機会は PC やスマートフォン，タブレット端末などの普及に伴い増加，低年齢化している．近年は，小学校や中学校の授業にもタブレット端末が導入されている．このような利用環境の中で，言語の音だけを頼りにかな入力し，十分な識別認識を欠いたまま変換後の漢字を用いるといったケースも散見される．特に「追求」と「追及」のような，意味や用法の差異が小さい同音異義語について顕著である．入力を支援するシステムの教育的な重要性は高まり，誤りを検出し再確認を促すシステムの必要性も高まっている．

同音異義語の誤りを検出する既存の手法は，複合語に着目し文字連鎖を用いる手法 [2] や決定リストを用いる手法 [3]，確率的 LSA を用いる手法 [4]，Earth Mover's Distance を用いる手法 [5] がある．

本研究では，ニューラルネットワークを用いた言語モデル BERT を同音異義語の誤り検出に用いる．同時に，機械学習の学習コストを減らす取り組みである Zero-shot 学習の形式も合わせて取り入れる．

Zero-shot 学習は，学習時にアクセスできない未知のラベルを推定する手法で，主に画像処理の物体認識の分野で未知の物体を推定する際に用いられている．この推定を可能にするため，事前知識という領域を設定することが特徴となる．この事前知識にはデータ，モデル，アルゴリズムなどが想定される．これにより，未知のデータに対して頑健なモデルを得ようとする取り組みである．

本研究では，この事前知識に BERT をあてる．BERT を用いてテストデータの同音異義語部に適する単語を 10,000 語予測させ，その結果を同音異義語の誤り検出に用いる．実験では五組の同音異義語を対象とした．テストデータは毎日新聞記事の中から対象単語の載る文を抽出し，それらから人為的に誤りを含ませて作成した．実験の結果から，人が細かいニュアンスの差異を頼りに識別する同音異義語の誤りを検出できる可能性が得られた．実験精度の悪かった同音異義語の検証から，事前知識に用いたモデルの一部に過剰適合が生じている知見が得られた．

2 BERT

BERT (Bidirectional Encoder Representations from Transformers) は双方向の Transformer[6] を用い，大規模データによって事前学習された言語モデルである．大規模データにより事前学習することで強力な汎用言語モデルを作成し，そのモデルを各タスクの教師ありデータに fine-tuning する二段階のフレームワークを用いている．BERT の構造は Transformer の self-attention を中心とした多層の双方向 Transformer エンコーダである．

BERT の入力表現はラベルの無い二文をトークン化したものと特殊トークンを合わせた一つのシーケンスとなる．特殊トークンは [CLS] と [SEP] に表現され，[CLS] はシーケンス最初に設置されるトークンであり，[SEP] は文の区切りに設置されるトークンである．これらの各トークンに対し，シーケンス中の位置と二文のどちらに属するかの情報を合わせ，埋め込み表現とする．この埋め込み表現を Transformer ブロックを中心とした多層で学習し，その結果を MLM や NSP に用いて下流タスクの精度を高めている．これまでの内容を図 1 として示す．

2.1 MLM

MLM (Masked Language Model) は文献により cloze タスクとも呼ばれる処理である．BERT においては入力トークンの 12%をランダムに [MASK] トークンに置き換え，それら [MASK] されたトークンを交差エントロピー誤差を元に予測するタスクを行う．最終的な [MASK] トークンの埋め込み表現はソフト

表 1　機械学習の分類

	task T	experience E (教師あり情報)	experience E (事前知識)	performance P
教師あり学習	クラス分類	ラベルありデータ	なし	分類精度
Few-shot 学習	クラス分類	小規模ラベルありデータ	事前学習モデル	分類精度
Zero-shot 学習	クラス分類	なし	事前学習モデル	分類精度

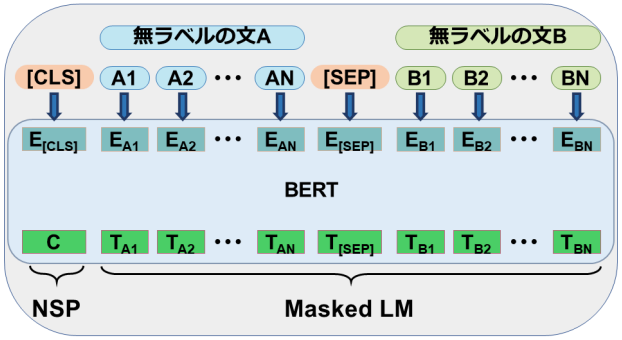

図 1　BERT モデル

マックスを通じて各単語と結びつくこととなる．このタスクにより従来の事前学習手法に比べて，文脈の前後を捉えた双方向性を獲得している．

2.2　NSP

NSP (Next Sentence Prediction) は，言語モデリングで得られない二文間の関係を BERT に理解させるタスクである．学習に用いられる単一のコーパスから，シーケンス用の文 A と文 B を選ぶ際，文 B は文 A に連続する文と非連続の文の 2 パターンが半々で選ばれる．非連続の文はにコーパス中からランダムに選ばれる．この連続・非連続に二値化された，次の文を予測するタスクが図 1 の左下の記される埋め込み表現 C を元に行われる．これにより，質問応答や自然言語推論といった文間関係の理解が重要な下流タスクの精度を高めている．

3　Zero-shot 学習による誤り検出

3.1　Zero-shot 学習

一般的な機械学習は学習コストが課題となる．学習コストはモデル作成のための計算コストや教師あり学習の教師データを作成するコストなどが挙げられる．後者に対しては半教師あり学習，教師なし学習，弱教師あり学習といった対策を生んだ．しかし，これらのいずれも基本的にはデータの量を前提にしている．一般的な機械学習のアプローチを用いると小規模データからは脆弱なモデルが生まれるという課題があるためである．

Zero-shot 学習は，これらの課題について改善を試みる取り組みとなる．Zero-shot 学習は Few-shot 学習の考え方に基づいているため，Few-shot 学習の概要から説明する．Few-shot 学習は，wang らの定義 [7] から，従来の一般的な機械学習の分類 (教師の有無など）を横断する概念となる．まず，機械学習の定義を wang らの方針を参照し設定しておく．「あるタスク T のためのプログラムが，その性能評価値 P に関して，経験 E によって改善される場合，そのプログラムは経験 E から学習している．」と設定する．few-shot はタスク固有の限られた数のラベル付きデータを意味し，経験 E にあたる．この段階では小規模データの教師あり学習と同じ様相となるが，小規模データのモデル脆弱化に対して，Few-shot 学習は事前知識という後ろ盾を用いる点が異なる．この事前知識も経験 E にあたり，その選定制約の基本は few-shot の持つ情報に触れないことである．その上で事前知識の対象には，タスク T を処理するのに適したデータ，モデル，アルゴリズムなど多様な対象が挙げられる．Zero-shot 学習はタスク固有のラベル付きデータを用いず，事前知識を中心にタスクを処理する．これまでに説明した機械学習の分類について wang らの定義に基づき表 1 に示す．

wang らは Zero-shot 学習については，E に他のモダリティを含ませる必要があることにも言及している．

3.2　提案手法

wang らの定義に基づいて提案手法を表現すると，タスク T はクラス分類としての同音異義語の誤り検出，経験 E は事前知識として事前学習モデルの BERT，性能評価値 P は分類精度となる．

まず，対象の同音異義語が含まれる文を誤り検出の対象文として入力する．入力文中から対象となる同音異義語部を BERT の [MASK] トークンに置き換える．その後，BERT を用いて [MASK] 部を予測する処理を行う．予測結果を BERT の尤度順に 10,000 語取得し，対象となる同音異義語の出現順を調べる．出現順の早い，BERT の尤度が高い方を入力文で用

いられるべき同音異義語の推測結果とする．推測結果と入力文の同音異義語が異なる場合に誤りとして検出する．

4 実験

4.1 実験設定

誤り検出に用いる同音異義語は筆者の主観により，表 2 に示す五組を対象とする．実験には 1993 年から 1999 年の毎日新聞の記事データを用いる．テストデータは，この記事データの中から対象の同音異義語が含まれる文を取り出すことで作成した．取り出す際に MeCab[1)] を経て単語の確認をしている．MeCab 用の辞書は mecab-ipadic-NEologd[8] を用いている．利用するデータ数は比較する同音異義語の文数が少ない方に合わせてランダムに抽出している．同音異義語の誤ったデータは，記事の内部に記載されている同音異義語について当該箇所だけをもう一方の同音異義語に置き換えることで作成した．誤りデータは全体の 50 %分作成する．BERT の事前学習モデルには東北大学から公開されている日本語版 BERT[2)] を用いている．この BERT の設定は「bert-base-japanese-whole-word-masking」のモデルを公開されているデフォルトのままで用いている．

二値分類として偏りの無いテストデータを用いているが，正解率や再現率を考慮するため実験の評価には F 値を用いた．

また，事前知識として用いた BERT の事前学習モデルが本実験で対象とした同音異義語について fine-tuning 無しにどの程度の識別するのかを検証するため，誤りを含めない新聞記事のままで予測させる実験も行っている．

4.2 実験結果

表 2 に，本研究で設定した内容の一部と実験の結果を示す．

誤り検出の結果を示す F 値は比較する同音異義語により様々な値となった．人が細かいニュアンスの差異を頼りに識別する同音意義語を対象としているので，「追及」と「追求」については精度が良いと言える．一方，「修行」と「修業」の精度は悪く，二値分類の結果としては脆弱となった．実験に用いたデータ数による影響は，データ数の最も多い「体制」と「態勢」，データ数の最も少ない「回答」と「解答」の結果を見る限り言及するほどに影響を与えていないと推察される．

表 2 実験結果

	記事内	抽出数	実験数	F 値	特異度
追及	9,029	2,500	5,000	94.3%	94.8%
追求	2,825	2,500			
意思	9,268	900	1,800	79.2%	82.3%
意志	964	900			
回答	10,871	500	1,000	76.6%	76.0%
解答	562	500			
体制	17,453	5,000	10,000	63.5%	64.5%
態勢	5,146	5,000			
修行	1,376	800	1,600	56.0%	58.9%
修業	888	800			

表 3 誤りを含めない識別結果

	記事内数	正	誤	正答率
追及	9,029	8,391	638	92.9%
追求	2,825	2,700	125	95.6%
意思	9,268	8,227	1,041	88.8%
意志	964	696	268	72.2%
回答	10,871	10,438	433	96.0%
解答	562	330	232	58.7%
体制	17,453	16,300	1,153	93.4%
態勢	5,146	1,677	3,469	32.6%
修行	1,376	1,278	98	92.9%
修業	888	170	718	19.1%

誤りの検出は，BERT の事前学習モデルが予測する単語を元に行っている．その影響を検証するため，誤りを含めない新聞記事のままで対象単語を予測させた結果を表 3 に示す．

検出精度の悪い同音異義語について，BERT の事前学習モデルは一方の単語に偏って予測しているという傾向が得られた．「修業」を「修行」と取り違えた文の例を表 4 に示す．表中の上二文については文が短く，文脈を想定しなければ「修業」と「修行」のどちらも正解と言って良いもので，テストデータの設定に調整が必要である．表中の下二文については，「修業」と「修行」の違いに話者の精神性や他者との関係性が含まれる点をよく理解する必要があるため，人が「修行」と誤っても責められず，より良いニュアンスとして「修業」を用いた方が良い文と言える．このようなニュアンスを優れた機械学習の

1） https://taku910.github.io/mecab/
2） https://github.com/cl-tohoku/bert-japanese

表 4 BERT が「修業」を「修行」と判定した文の例

山田さんは修業３年目だ.
まだまだ修業中の身ということを思い知らされた.
大学を卒業後，神戸のレストランでしばらく修業し，再度フランスへ.
修業感覚を欠いた現在中心の快楽志向が日本の若者の特徴.

モデルは内包しきるのか，今後も検証を進めたい.

5 おわりに

本研究では Zero-shot 学習の形式に基づき，BERT を用いた同音異義語の誤り検出を試みた．実験の結果は，現段階の提案手法ではニュアンスの差異が小さい同音異義語の誤り検出について様々な精度となることを確認した．これは Zero-shot 学習の事前知識として用いた BERT の事前学習モデルが，検出精度の悪くなる同音異義語を予測する際に，一方に偏って予測することに起因している．偏りの生じる原因は事前学習の際に用いられるデータの質と量が関わると推察する．本実験の BERT モデルは日本語版 Wikipedia コーパスを用いて事前学習されている．新聞記事よりも誤植について制約がないため同音異義語が使用される場面での質が担保されない．また，大規模データを前提とする事前学習モデルに，本実験のような単語ごとの量的な調整は行われない．このような事前学習データの質と量の問題が，機械学習の過剰適合に繋がったのではないかと推察する．本実験の様に fine-tuning を介さない Zero-shot 学習の形式をとることで，fine-tuning で見失われやすい汎用モデル自体の課題が生じた.

今後の課題として，まずは本実験の対象を拡大し結果の傾向をより詳細に分析することが挙げられる．次に，BERT の学習に用いられた日本語版 Wikipedia コーパスにおいて，本実験で対象とした同音異義語の量的な検証を行い，その影響を調べる必要もある．以上の内容を踏まえた BERT の事前学習モデルを作成し，本実験との比較も試みたい．また，本実験は Zero-shot 学習の事前知識として単一のモデルを用いたが，複数のモデルを用い集合知の形式を取り入れるなどの検証を進めたい.

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] 奥雅博, 松岡浩司. 文字連鎖を用いた複合語同音異義語誤りの検出手法とその評価. 自然言語処理, Vol. 4, No. 3, pp. 83–99, 1997.

[3] 新納浩幸ほか. 複合語からの証拠に重みをつけた決定リストによる同音異義語判別. 情報処理学会論文誌, Vol. 39, No. 12, pp. 3200–3206, 1998.

[4] 三品拓也, 貞光九月, 山本幹雄ほか. 確率的 lsa を用いた日本語同音異義語誤りの検出・訂正. 情報処理学会論文誌, Vol. 45, No. 9, pp. 2168–2176, 2004.

[5] 河原直人, 梅澤猛, 大澤範高ほか. E-016 earth mover's distance を用いた同音異義語判別 (e 分野: 自然言語・音声・音楽). 情報科学技術フォーラム講演論文集, Vol. 12, No. 2, pp. 217–218, 2013.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[7] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, Vol. 53, No. 3, pp. 1–34, 2020.

[8] 奥村学佐藤敏紀. 単語分かち書き辞書 mecab-ipadic-neologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会 (NLP2017), pp. NLP2017–B6–1. 言語処理学会, 2017.