前後段落を用いて生成した単語分散表現による日本語語義曖昧性解消の検証

前原太陽 [1]　竹中要一 [2]
[1] 関西大学大学院　総合情報学研究科　[2] 関西大学　総合情報学部
suisougakuperc@gmail.com　takenaka@kansai-u.ac.jp

概要

近年、言葉の意味をベクトルで表現する分散表現を用いることで、コンピュータが言葉の意味を扱いやすくなった。しかし、多義語の語義曖昧性解消という問題が依然として存在している。語義曖昧性解消とは、多義語の文章中での語義を判定することである。これはコンピュータが言語の意味を処理するために重要な作業である。本研究では日本語の語義曖昧性解消を目指し、異なる語義のクラスタ間の分散は大きく、クラスタ内の分散は小さくなるように、分散表現を生成する方法を提案、検証する。提案するモデルは、対象段落の前後の段落のデータを用いる。既存手法と提案手法の両方で分散表現を生成し、クラスタ間分散とクラスタ内分散、総合評価を行う。

1　はじめに

言語には同じ単語に複数の意味が紐づく多義語が存在する。例として「頭」という文字を岩波国語辞典 第五版で調べてみる。「頭」という漢字の読みとしては「あたま」「かしら」「ず」「がしら」の 4 種類が存在しており、それぞれに語義が定義されている。「あたま」という読みには「動物の脳（や目・口・耳・鼻）がある部分。かしら。こうべ。」や「頭と関係が深い次のもの。」「頭に似たもの。」「あたまかず。人数。」といった 4 語義が存在する。さらに、4 語義の中でも「頭と関係が深い次のもの。」と「頭に似たもの。」にはより細かい語義が定義されており、一番細かく分類されている語義数で「頭」の語義を数えると、合計 13 語義が存在する (表 1)。

「頭が追い付かない」という文章の場合、文章に出てきた「頭」という文字が「動物の脳がある部分」として使われているわけではなく、「脳の動き」という語義として使われている。人間にとっては文章中に使われている単語の語義を、文脈から判断することは比較的容易であるが、コンピュータにとってはまだ難しい課題となっている。

表 1　「頭」の読みと語義

読み	語義
あたま	動物の、脳（や目・口・耳・鼻）がある部分。かしら。こうべ
	髪
	脳の動き。考え方。心。
	物の上部。てっぺん。
	上に立つ人。首脳。かしら。
	うわまえ。
	最初。
	あたまかず。人数。
かしら	あたま。文語的。
	一族の長。首領。 特に、仕事師・街火消の棟梁
がしら	... したとたん。
	一番... した人。
ず	あたま

※「頭」の漢字に関する項目は省いた。

自然言語処理における語義曖昧性解消 (Word Sense Disambiguation, WSD) とは文章中で使用されている多義語の文章中での語義を判定するを正しく判別することをいう。WSD はコンピュータによる文章の意味理解の精度を向上させるために重要なタスクであると言える。

本研究では WSD を目的とした分散表現の生成方法を提案する。分散表現の生成方法としては「文脈埋め込みの分散表現」である BERT を用いる。Transformers[1] で提供されている基本的なモデルである BertModel を継承した提案モデルを作成し、新たに学習をするわけではなく、事前学習モデルを利用した上で、分散表現を生成したい多義語を含む段落の前後段落を提案モデルに入力し、分散表現を生

成する。

分散表現を利用して WSD を目指した研究は過去にもいくつか報告されている。菅原ら [2] は Word2Vec によって生成された分散表現を用い、多義語の素性を前後の分散表現を並べたものとした。提案した素性を用いて語義曖昧性解消タスクを行ったところ、従来の単語の素性を用いた場合に比べ高い精度を達成した。また、曹ら [3] は BERT が文脈埋め込みの分散表現であることに注目し、生成された分散表現を教師あり学習を行った分類器で WSD を行ったところ、高い精度であったことを報告した。

本研究では多義語が持つ語義をそれぞれ一つのクラスタとし、文中に出現する単語を要素とする。単語に分散表現を紐づけることで、クラスタ内分散やクラスタ間分散の計算が可能となる。

また、目標としては語義のクラスタ内分散を小さく、同時に語義同士のクラスタ間分散を大きくすることで、語義の判別を行いやすくすることである。提案手法と従来の BertModel で分散表現を生成したのち、語義のクラスタに関して Pseudo F の値を計算し、比較する。

2 手法

2.1 BERT

BERT [4] は 2018 年に Jacob らによって提案された Transformer[5] を活用した自然言語処理の深層学習モデルである。主に翻訳や文章分類、質問回答などの分野で活用されており、各タスクに特化したライブラリも公開されている。

BERT はラベルなしの大量の文章データを元に行う「事前学習」と、ラベル付きデータを用いて特定分野や特定タスクに特化した形にモデルのパラメータを調整する「ファインチューニング」の 2 段階で学習を行う。

また、事前学習は「マスク付き言語モデル」と「Next Sentence Prediction」の 2 つの方法によって、対象言語の法則を学習させる。「マスク付き言語モデル」では文章中のトークンを [MASK] という特殊トークンに置換する。全トークンのうち、15 % を [MASK] に置き換え、残りの 75 %のトークンから [MASK] が元々どのようなトークンだったかを予測するタスクによって学習を行う。「Next Sentence Prediction」は入力した 2 つの文章が連続している文章かどうかを判断するタスクによって学習する。

2.2 BertModel

BertModel は Transformers のクラスのひとつであり、文章を入力として、トークンの分散表現を出力することができる。また、Transformers では提供されていないような特定のタスクに特化したクラスを作成する際の継承元のクラスとしても利用されている。BertModel を使う際は、事前にトークナイザを用いて、文章をトークン ID に変換する必要がある。

トークナイザからの出力を BertModel に入力することによって、それぞれの単語に対して 768 次元の分散表現を生成する。BertModel で分散表現を生成する流れを図 1に示す。なお、図中で embedding 層に入力するトークン列の長さは 256 に設定している。

2.3 Sentence-BERT

Sentence-BERT(SBERT)[6] は 2019 年に Reimers らによって発表された、BERT を改良したモデルである。元々の BERT も文章分類や FAQ の自動生成といった文章に関する精度が高いものとして知られているが、データ量が増えると計算時間が膨大になってしまうという課題が存在した。SBERT は BERT が抱える計算効率の低さを改善しただけではなく、文章を扱うタスクにおいて精度が高いモデルとして報告されている。

2.4 提案モデル

提案モデルを図 2に示す。BertModel への入力としてトークナイザからの出力に加え、トークナイザの入力として利用した段落の前後の段落の分散表現を、追加情報として入力する。前処理として前後段落の分散表現を SBERT を用いて生成する。

提案モデルによる分散表現の生成は、BertModel に入力したトークン ID が embedding 層を通過するまでは従来の BertModel と同様の動作とする。

相違点としては embedding 層の出力を、前処理で生成した SBERT の出力と結合させたのちに encoder 層に入力する点である。encoder 層からの出力から SBERT を結合した領域を削除し、最終的な単語の分散表現とする。

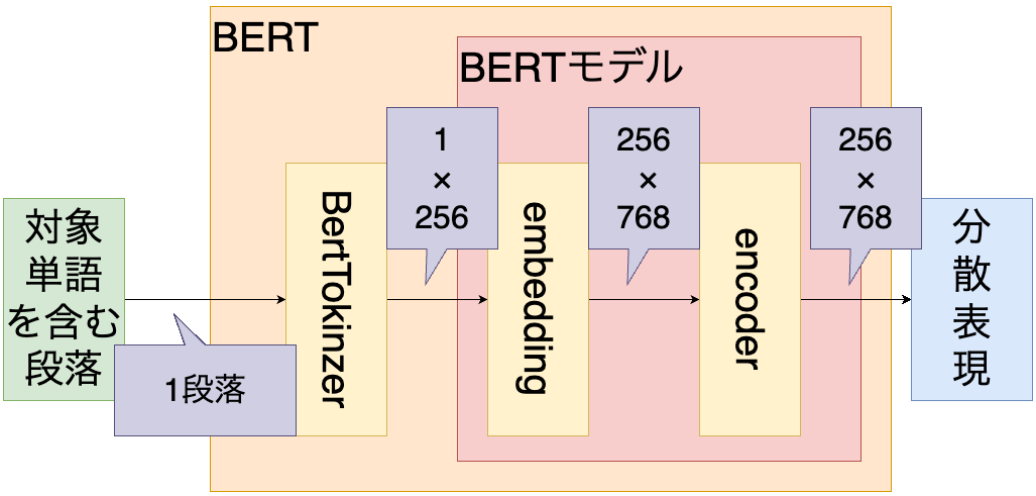

図 1 BertModel

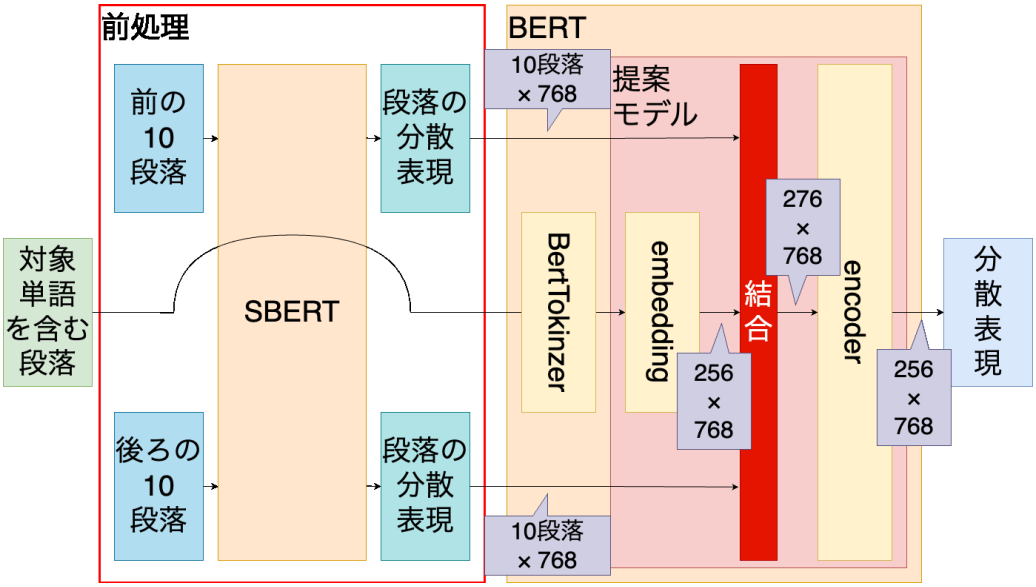

図 2 提案モデル

3 実験

3.1 条件

提案手法において生成した分散表現が語義曖昧性解消に寄与することを検証するために実験を行った。分散表現を生成する文章として、奥村らが作成した Semeval-2010 Japanese WSD Task[7] のデータセットを用いた。SemEval-2010 Japanese WSD Task は書籍や白書、新聞といったテキストデータを分かち書きし、それぞれの単語に語義 ID を付与した合計 1980 件のデータセットである。

本研究では学習済みモデルとして乾らが公開している cl-tohoku/bert-base-japanese-whole-word-masking [8] を利用する。提案モデルに使用する前後段落数は 10 段落とし、embedding 層へ入力するトークン列の長さは 256 とした。なお、SemEval-2010 Japanese WSD Task ではすでに分かち書きがされているため、BertJapanseTokenizer をトークナイザとしては利用せず、トークン ID への変換のみを行った。

実験では語義曖昧性解消の評価対象として、活用のない名詞のみを対象とした。

3.2 Pseudo F

BertModel と提案モデルで生成した分散表現において、語義ごとのクラスタの変化量を評価する値として、Calinski らによって提案された Pseudo F[9] を用いる。

Pseudo F の定義を式 (1) に示す。k はクラスタ数を表し、E は要素 n の集合である。W_k, B_k は式 (2)、

式 (3) でそれぞれ表され、W_k はクラスタ内分散の和を、B_k はクラスタ間分散を表す。また，c_q はクラスタ q における中心、C_q はクラスタ q 内の要素の集合である。

$$Pesudo\ F = \frac{trace(B_k)}{trace(W_k)} \times \frac{n_E - k}{k - 1} \tag{1}$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T \tag{2}$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T \tag{3}$$

Pseudo F はクラスタ間の距離をクラスタ内分散の合計で割ったものであるため、クラスタ間の距離が大きく、クラスタ内分散が小さほど Pseudo F の値が大きくなる。したがって、Pseudo F の値が大きいほど本研究の目標の目的である、語義ごとのクラスタが判別が行いやすくなっていると言える。

4 結果

BertModel と提案モデルを比較するために提案モデルにおける Pseudo F、クラスタ内分散、クラスタ間分散の値を BertModel の値で割った値を用いる。式 (4) で示す値が 1 を超えていれば、BertModel の値よりも上昇しており、反対に 1 を下回れば BertModel の値よりも減少していることになる。なお、Pseudo F は各単語ごとに計算する。

$$\text{値の上昇率} = \frac{\text{提案モデルの値}}{BertModel\ \text{の値}} \tag{4}$$

全単語における Pseudo F の値の上昇率をバイオリンプロットした結果を図3に示す。Pseudo F の値の上昇率が 1 周辺になっている単語が多いことがわかる。また、各単語における Pseudo F、クラスタ内分散、クラスタ間分散の値についての平均値と中央値を表 2に示す。

表 2 平均値と中央値

値	平均	中央値
Pseudo F	1.033	1.010
クラスタ内分散	0.976	0.980
クラスタ間分散	0.996	0.993

Pseudo F の値において、平均、中央値 1 を超えており、BertModel よりも値が向上していることがわかる。中央値が 1 を超えているという点から、対象単語の半分以上 (52%) の単語にて値が向上している。

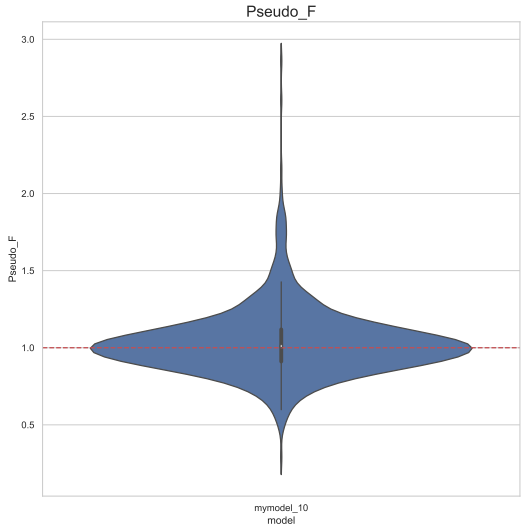

図 3 Pseudo F

5 終わりに

本研究では BertModel における embedding 層への入力に、従来のトークン ID の配列などに加え、トークン ID の基となった段落の前後の段落の分散表現を追加情報にすることで、語義曖昧性解消を目的とした分散表現の生成を目指した。BertModel と提案モデルの両方で作成した分散表現について Pseudo F の値を計算し、比較したところ半分以上の名詞において Pseudo F の値が向上した。クラスタ間分散およびクラスタ内分散については両方減少したが、分子であるクラスタ間分散の方が減少率が少なかったため、Pseudo F の値が上昇したと考えられる。

また、本研究ではファインチューニングなどを行わない状態で、各単語の語義のクラスタに着目して実験し、評価を行った。今後はファインチューニングを行ったうえで、実際に語義を推定するような実験を行い、正答率の向上を確認する必要があると考える。

参考文献

[1] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. pp. 38–45. Association for Computational Linguistics, October 2020.

[2] 菅原拓夢, 笹野遼平, 高村大也, 奥村学. 単語の分散表現を用いた語義曖昧性解消. 言語処理学会 第 21 回年次大会 発表論文集, pp. 648–651, 2015.

[3] 曹鋭, 田中裕隆, 白静, 馬ブン, 新納浩幸. Bert を利用した教師あり学習による語義曖昧性解消. 言語資源活用ワークショップ発表論文集= Proceedings of Language Resources Workshop, 第 4 巻, pp. 273–279. 国立国語研究所, 2019.

[4] 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.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.

[6] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

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

[8] 東北大学自然言語処理研究グループ. Bert-base japanese whole-word masking. https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking, 2022.

[9] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. **Communications in Statistics-theory and Methods**, Vol. 3, No. 1, pp. 1–27, 1974.