NWJC-BERT: 多義語に対するヒトと文脈化単語埋め込みの類似性判断の対照分析

浅原 正幸 *　　西内 沙恵　　加藤 祥
国立国語研究所　国立国語研究所・筑波大学　国立国語研究所

1. はじめに

語義を扱う基本的な言語資源として単語埋め込み (word embedding) がある。単語埋め込みは各単語の意味をベクトル表現化する技術で、コサインなどの類似度により 2 単語の意味的な近さを計算したり、ベクトルに対する加減算により語義の合成・分解が可能になる。古くは、分布意味論において、単語の共起をモデル化する単語文書行列を構成したうえで、特異値分解により、疎な単語共起ベクトルから密な単語共起ベクトルを構成する手法が用いられてきた。2010 年代に入り、人工神経回路に基づく単語埋め込みの研究が進展し、様々なモデルが提案された (Mikolov et al. 2013, Pennington et al. 2014, Bojanowski et al. 2017)(1)。2018 年には文脈を考慮した単語埋め込み (contextual embedding) を学習モデルとして ELMo (Peters et al. 2018)(2)が提案された。また、言語処理の事前学習 (pre-training) モデルである BERT が提案された (Devlin et al. 2019)。BERT は「空白穴埋め問題」(masked language model) と「文の隣接課題」(next sentence prediction) を解く事前学習済みモデルを、他の課題に転移学習することができる。BERT の公開サイト(3)に多言語訓練済みが公開されているほか、日本語の訓練済み事前学習モデルが公開されている (Kikuta 2019, 柴田ほか 2019)。これらのモデルを用いて，文脈化された単語埋め込みも得ることができる。本研究では『国語研日本語ウェブコーパス』(以下 NWJC) (Asahara et al. 2014) により事前学習した BERT モデルについて紹介する。このモデルは、多層文脈化単語埋め込みをテキストに付与することを想定し、以下のような言語学研究での利用を目的として構築した：

- 多義語の語義をベクトル空間上に表現し、近さや連続性について評価したい
- 既存の辞書にない語義を検出したい
- 換喩・提喩を含めた比喩表現に見られる語義の転換を定量的に評価したい
- 節の境界性の強弱を連続的に評価したい

以下では、NWJC-BERT モデルの構築手法とともに、クラウドソーシングによる多義語の類似性判断結果と NWJC-BERT によって得られる文脈化単語埋め込みの文単位・単語単位の類似度との対照分析結果について示す。

2. NWJC-BERT モデルの構築作業

本節ではモデルの構築作業について概説する。訓練データは、NWJC の 6 単語以上の文 1,287,504,831 文 (22,653,063,443 語) を用いる。テキストは MeCab-UniDic-2.1.2 で解析された「語彙素」(形態論情報の 0-origin で 7 番目の要素) の列に変換し、語彙素列に対するモデルを構築した(4)。NWJC は文単位のコーパスであり、2 文以上の前後文脈については保存されていない。そこで、文を任意の位置で分割し、文の前件と後件とを推定するタスクとして学習させた。BCCWJ-ToriClause と

* masayu-a@ninjal.ac.jp

(1) word2vec: https://github.com/tmikolov/word2vec, GloVe: https://nlp.stanford.edu/projects/glove/, fastText: https://fasttext.cc/ など

(2) https://github.com/allenai/allennlp/blob/master/tutorials/how_to/elmo.md

(3) https://github.com/google-research/bert

(4) tokenization.py の text = self_tokenize_chinese_chars(text) をコメントアウトし --do_lower_case=false

表 1　既存の日本語 BERT モデルとの比較

	BERT with SentencePiece (Kikuta 2019)	hottoSNS-bert	@mkt3	京大 BERT (柴田ほか 2019)	NWJC-BERT
vocab_size	32,000	32,000	32,000	32,000	48,914
コーパス	日本語 Wikipedia	日本語 Twitter	日本語ビジネスニュース記事	日本語 Wikipedia	NWJC
		8592 万ポスト 14 億 Tokens	300 万記事	1800 万文 +whole word masking model	12.8 億文 226 億 Tokens
global_step	1,400,000	1,000,000			2,000,000
loss	1.377				1.442
masked_lm_accuracy	0.681		0.7 弱		0.709
masked_lm_loss	1.421				1.442
next_sent_accuracy	0.985		0.995 以上		0.965
next_sent_loss	0.595				0.082

東北大 BERT: mecab-ipadic-bpe-32k (MeCab による分割 +subword, vocab:32000, 日本語 Wikipedia, 1,000,000 iter)
東北大 BERT: mecab-ipadic-char-4k (MeCab による分割 + 文字単位, vocab: 4000, 日本語 wikipedia, 1,000,000 iter)
東北大 BERT はいずれも whole word masking model あり。

対照させて節の境界性を推定することを想定するが、分割箇所は節境界とはかぎらない。

単語リスト (`vocab.txt`) は UniDic の機能語全て 154 語彙素と UniDic-分類語彙表番号対応表 (WLSP2UniDic) に出現する 48,790 語彙素と制御語 5 種からなる。なお、 WLSP2UniDic には一部機能語と同表記のものが含まれることから、合計 48,914 語となる。このような単語リストを構築することにより、UniDic のエントリの 54.6% (468,460/872,831 エントリ) を被覆するほか、分類語彙表番号によるモデルの評価を行うことができる。しかしながら、語彙素に基づくモデルのため、サブワードによる性能向上は期待できない。訓練は NVIDIA DGX-1 の上の GPU 2 枚で行った(5)。

図 1 に BERT 訓練時の単語穴埋め問題 (Masked Language Model) と隣接文推定問題 (Next Sentence Prediction) の学習曲線を示す。global_step が 500,000 回でおおよそ収束した。以下の議論では、global_step 2,000,000 回のモデルを用いる。表 1 に既存の日本語 BERT モデルとの比較を示す。既存手法との大きな違いは、語義の選択手法にある。既存のものは、表層形を用いたうえで、空間計算量を節約するために 32,000 語彙を頻度などに基づいて取捨選択している。NWJC-BERT は、UniDic の語彙素に基づき 48,914 語彙を選択した。訓練元データとして　NWJC を用い、日本語 wikipedia の約 70 倍の文数から学習した。NWJC は前後の文の情報を保持していないため、本研究では隣接文推定を「1 文を 2 分割したものの推定」におきかえた。このため、隣接文推定のタスクが他のものより難しい設定となり、精度がやや低い。

3.　多義語に対するヒトの類似性判断と BERT による類似度の対照分析

西内ほか (2020) は、多義形容詞の複数の例文間の類似度を一対比較法によりクラウドソーシングを用いて類似性判断を収集したデータベースを構築した(6)。そのなかで時空間的な間隔を評価する「短い」(36 文)「長い」(36 文)「近い」(43 文)「遠い」(34 文) を用いて、ヒトの類似性判断と BERT が出力する文脈化単語埋め込みにおける cosine 類似度の対照分析を行った。文脈化単語埋め込みの出力においては、最終層の文単位のものと単語単位のものを出力した。表 2 に Spearman 順位相関係数を示す。いずれの語もヒトの類似性判断と NWJC-BERT が出力する類似度と強い相関があるとは言えないことがわかった。

次に、定性的な分析について示す。図 3 に「短い」の BERT による単語単位のベクトルの cosine 類

(5) 細かなオプションは次のとおり： `--train_batch_size=32 --max_seq_length=128 --max_predictions_per_seq=20 --num_train_steps=XXXXXXXX --num_warmup_steps=10000 --save_checkpoints_steps=10000 --max_eval_steps=100 --learning_rate=2e-5`

(6) データの詳細については、西内ほか (2020)・加藤 (2019)・加藤ほか (2019) を参照されたい。

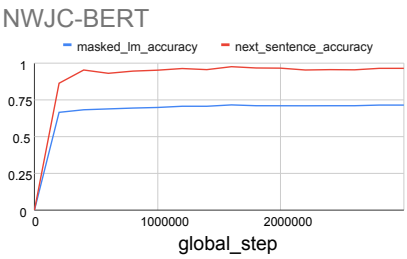

図 1　NWJC-BERT の学習曲線

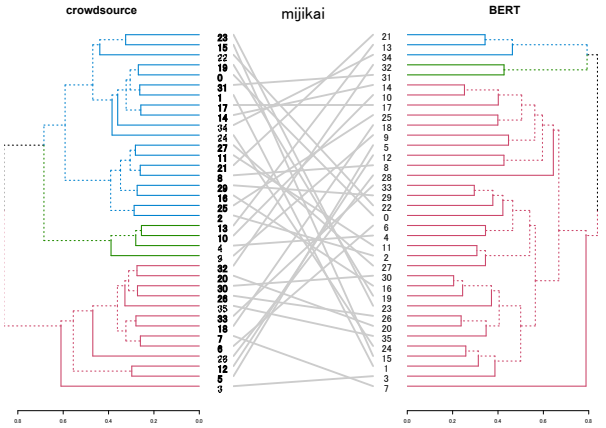

図 2　ヒトの類似性判断と BERT による類似度（単語単位）

表 2　ヒトの類似性判断と NWJC-BERT による類似度の順位相関

多義形容詞	短い	長い	近い	遠い
例文数	36	36	43	34
ヒト (crowd) と 文単位 (NWJC-BERT) の順位相関係数	0.1384	0.1556	0.0488	0.1556
ヒト (crowd) と 単語単位 (NWJC-BERT) の順位相関係数	0.0232	0.2170	0.1740	0.0948

似度（距離に変換したもの）を用いた最遠隣法によるデンドログラムを示す。各クラスタに対する特徴を赤の BOX で示す。また、時間が近いことを表す用例については ＜時間＞を付与した。ある程度、時間を意味する用例が固まって出現する一方、形式や文法に基づくグループ化が確認できた。

図 2 に、「短い」のヒトの類似性判断（図中左）と NWJC-BERT による単語単位の類似度（図中右）を対照したタングルグラムを示す [(7)]。タングルグラム中、ヒトの評価と NWJC-BERT による評価が異なる例について確認する用例 (6) と (7) は、ヒトの評価ではデンドログラム上の同じ山に位置している一方、NWJC-BERT における類似度では離れたところに位置している。いずれも「人生の短さ」を判断しており、ヒトの類似性評価では平均 3.85/5 と高い値が得られている。しかし、NWJC-BERT の類似度では (6) は過去形により既に終わった時間を、(7) は連体修飾により残りの時間を表現しており、表現の形式が異なるために cosine が 0.29 (似ていない 0 < cosine <1 似ている) と低い値であった。

(6) 道元禅師の生涯は、あまりに【短かっ】た。

(7) 私たちのように年を取った先の【短い】者は、じゃあどうしたらいいの、と言いたくなる気もします。

これらの結果から、NWJC-BERT によるモデル化は分布意味論的な類似度を評価しているが、クラウドソーシングの結果は認知意味論的な類似性を評価しており、乖離がみられる。別の観点から説明すると、 NWJC-BERT は言語の生成過程の成果物である記号列（テキスト）からモデル化しており、前後文脈の形式の類似度に依存している。しかし、ヒトの評価は前後文脈の形式が異なっていても、語義が指し示す概念が似ている場合に類似性を認めており、この差異が 2 つの評定値の順位相関の低さを反映しているのではないかと考える。

4.　おわりに

本稿では、言語研究を目的とした事前学習モデル BERT の訓練済みデータ NWJC-BERT について紹介した。クラウドソーシングによる語義調査と対照し、NWJC-BERT で評価できる形式の類似性と

(7) 「短い」のヒトの類似性判断の詳細な結果については、西内ほか (2020) を参照されたい。

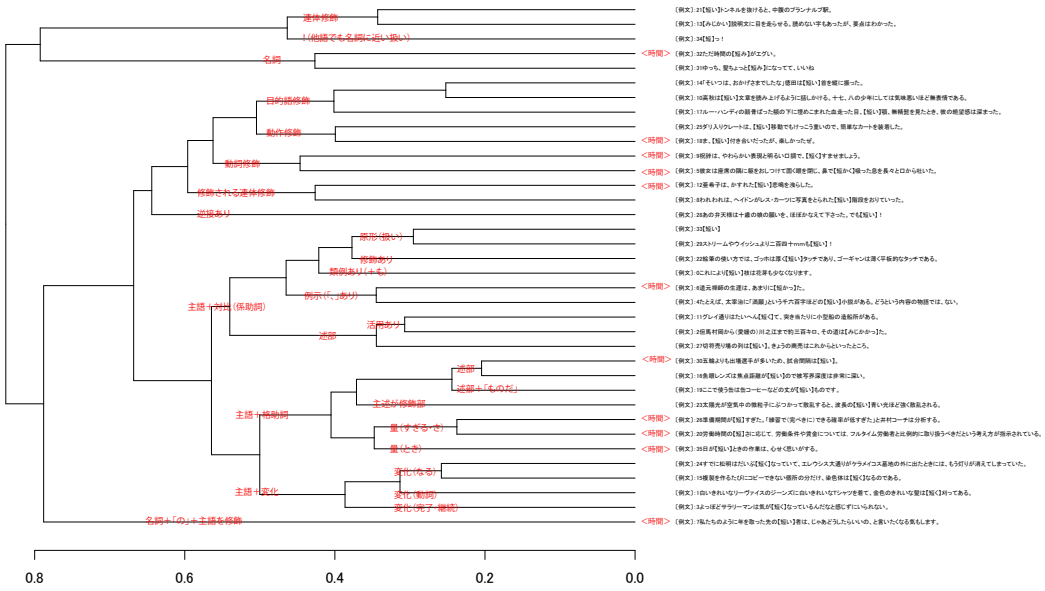

図 3　BERT による単語単位の類似度評価「短い」(値は cosine の逆関数)

評価できない語義の類似性について明らかにした。今回作成したモデルは 2020 年度中に公開予定である。

謝　辞

本研究は国立国語研究所コーパス開発センター共同研究プロジェクトおよび科研費 JP17H00917, JP18H05521, JP18K18519, JP19K00591, JP19K00655 によるものです。

文　献

T. Mikolov, K. Chen, G. Corrado, and J. Dean (2013). “Efficient Estimation of Word Representations in Vector Space.” *Proceedings of Workshops at ICLR*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning (2014). “GloVe: Global Vectors for Word Representation.” *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). “Enriching Word Vectors with Subword Information.” *Transactions of the Association for Computational Linguistics*, 5, pp. 135–146.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). “Deep Contextualized Word Representations.” *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *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.

Yohei Kikuta (2019). *BERT Pretrained model Trained On Japanese Wikipedia Articles*. `https://github.com/yoheikikuta/bert-japanese`.

柴田知秀・河原大輔・黒橋禎夫 (2019).「BERT による日本語構文解析の精度向上」　言語処理学会第 25 回年次大会 (NLP2019) 発表論文集, pp. 205–208.

Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi (2014). “Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan.” *Alexandria: The Journal of National and International Library and Information Issues*, 25:1–2, pp. 129–148.

西内沙恵・加藤祥・浅原正幸 (2020).「ヒトによる多義的形容詞に対する類似性の評価データベース構築：「長い」と「短い」の事例から」　言語処理学会第 26 回年次大会発表論文集.

加藤祥 (2019).「クラウドソーシングによる語義調査」　日本言語学会第 158 回大会予稿集, pp. 373–378.

加藤祥・西内沙恵・浅原正幸 (2019).「多義語用例の類似度による語義の分類 – 「遠い」と「近い」を例に–」　日本認知言語学会第 20 回大会発表論文集.