国立国語研究所学術情報リポジトリ

Clustering of sentences including a target word by using BERT

メタデータ	言語: jpn 出版者: 公開日: 2020-02-06 キーワード (Ja): キーワード (En): 作成者: 馬, ブン, 田中, 裕隆, 曹, 鋭, 白, 静, 新納, 浩幸, Ma, Wen, Tanaka, Hirotaka, Cao, Rui, Bai, Jing メールアドレス: 所属:
URL	https://doi.org/10.15084/00002586

BERT を利用した単語用例のクラスタリング

馬ブン (茨城大学大学院理工学研究科情報工学専攻) *
田中裕隆 (茨城大学工学部情報工学科) †
曹鋭 (茨城大学大学院理工学研究科情報工学専攻) ‡
白静 (茨城大学大学院理工学研究科情報工学専攻) §
新納浩幸 (茨城大学大学院理工学研究科情報工学専攻) ¶

Clustering of sentences including a target word by using BERT

Ma Wen (Graduate School of Science and Engineering, Ibaraki University)
Hirotaka Tanaka (Department of Computer and Information Sciences, Ibaraki University)
Cao Rui (Graduate School of Science and Engineering, Ibaraki University)
Bai Jing (Graduate School of Science and Engineering, Ibaraki University)
Hiroyuki Shinnou (Graduate School of Science and Engineering, Ibaraki University)

要旨

事前学習モデルである BERT は入力文中の単語に対する埋め込み表現を出力するが，その埋め込み表現はその単語の文脈に依存した形となっている．つまり BERT から得られる単語の埋め込み表現はその単語の意味を表現していると考えられる．本論文では，この点を確認するために BERT から得られる単語の埋め込み表現を利用して，その単語の用例のクラスタリングを行う．実験では日本語版 BERT 事前学習モデルを利用して，単語「意味」の用例クラスタリングを行った．語義曖昧性解消のための標準的な特徴ベクトルや分散表現から構築した特徴ベクトルからクラスタリングを行う場合と比較することで，BERT から得られる単語の埋め込み表現が，より適切に意味を表現できていることを示す．

1. はじめに

本論文では BERT (Devlin et al. (2018)) から出力される単語の埋め込み表現がその単語の意味を表現していることを確認するために，BERT から得られる単語の埋め込み表現を用いて，単語用例のクラスタリングを行う．

事前学習モデルである BERT は入力文中の単語に対する埋め込み表現を出力するが，その埋め込み表現はその単語の文脈に依存した形となっている．つまり BERT から得られる単語の埋め込み表現はその単語の意味を表現していると考えられる．この点を確認するために，単

* 19ND302H@vc.ibaraki.ac.jp
† 16T4032N@vc.ibaraki.ac.jp
‡ 18ND305G@vc.ibaraki.ac.jp
§ 19ND301R@vc.ibaraki.ac.jp
¶ hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

語 w の用例を収集し，その用例を BERT に入力し，単語 w に対応する埋め込み表現の集合をクラスタリングする．埋め込み表現が意味を表現していれば，語義毎のクラスタが形成されると考えられる．

実験では SemEval-2 の日本語の語義曖昧性解消 (Word Sense Disambiguation, 以下 WSD) のタスク (Okumura et al. (2011)) で対象単語の 1 つとなった単語「意味」を取りあげる．そこでの訓練用例とテスト用例を合わせた 99 用例を対象に，日本語版 BERT 事前学習モデルを利用して，単語「意味」の用例クラスタリングを行った．WSD の標準特徴ベクトルや分散表現から構築した特徴ベクトルからクラスタリングを行う場合と比較することで，BERT から得られる単語の埋め込み表現が，より適切に意味を表現できていることを示す．

2. 関連研究

本研究は事前学習モデルである BERT を Word Sense Induction (WSI) に応用したと捉えられる．ここでは WSI と事前学習モデルの 2 つの観点から関連研究を述べる．

WSI は入力文内の対象単語の意味をコーパスから推定するタスクである (Manandhar et al. (2010),Navigli and Vannella (2013))．推定対象となる意味の表現と推定する手法にバリエーションはあるが，標準的には，対象単語の用例をクラスタリングし，そのクラスタを意味と考えて，対象単語の入力文がどのクラスタに属するかで意味を推定する．そのため WSI は本質的には対象単語用例のクラスタリングと見なせる．クラスタリングのポイントは用例中の対象単語をどのような特徴ベクトルとして表現するかである．古典的には WSD で利用される対象単語に対する特徴ベクトルを用いればよい．近年は分散表現を利用した特徴ベクトルの作成法も提案されている (Kågebäck et al. (2015))．

事前学習モデルは大規模なコーパスからあらかじめ学習させた状態の言語のモデルである．様々なモデルの形態があるが，標準的にはそのモデルを利用して，入力単語列をその埋め込み表現列へ変換する．利用法としてはその変換された埋め込み表現列を直接タスクに利用する feature based の利用法と，タスクに対するネットワークを学習すると同時にそのモデル自体も学習する fine-tuning の利用法がある．OpenAI GPT (Radford et al. (2018)) はニューラルネット翻訳の Transformer (Vaswani et al. (2017)) の decoder 部分を利用した言語モデルである(1)．個別のタスクを解くネットワークをそのモデルに連結して利用する．ネットワークのパラメータを学習する際に，連結された言語モデルのパラメータも同時に更新する fine tuning を行うことで，転移学習が行える．言語モデルをタスクに応じて fine tuning するという観点では ULMFiT (Howard and Ruder (2018)) も知られている．ただし ULMFiT はネットワークの構造を提案したものではなく，言語モデルの fine tuning による転移学習に特化した学習方法を提案している．ELMo (Peters et al. (2018)) は文脈を考慮した単語の分散表現を導くモデルである．実体は 2 層の双方向 LSTM であり，大規模コーパスを利用して言語モデルを学習する．これが事前学習モデルとなり feature based の形で利用できる．BERT は OpenAI や ELMo の改良版と位置づけられる事前学習モデルであり，多くの自然言語処理

(1) 言語モデルは一種の事前学習モデルである．

のデータセットに対して SOTA を実現している．

3. BERT を用いた単語用例のクラスタリング

3.1 BERT

BERT の基本のパーツは Multi-head attention である．Multi-head attention は n 単語埋め込み表現列を入力として，各埋め込み表現をより適切なものに変換して出力する．つまり出力は変換された n 単語埋め込み表現列である．

Multi-head attention の概略を述べる．基本は self attention なので Q, K, V の 3 組が入力である．今，単語埋め込み表現が m 次元であったとする．Multi-head attention では m 次元ベクトルを $d_k(= m/k)$ 次元に圧縮する線形変換器を Q, K, V それぞれに対して用意する．Q, K, V の実体は $d_k \times d_k$ の線形変換行列である．Multi-head attention の入力は n 個の m 次元ベクトルであるが，これが先の圧縮機で $n \times d_k$ の行列 X に変換され，Q, K, V に渡され $n \times d_k$ の行列 XQ, XK, XV ができる．これらを Q', K', V' とおき，以下の式(2)により self attention を行う．

$$\texttt{softmax}\left(\frac{Q'K'^T}{\sqrt{d_k}}\right)V'$$

これは $n \times d_k$ の行列である．上記の処理を k 個並行して行うと，$n \times d_k$ の行列が k 個作成され，これらを横に連結することで，$n \times m$ の行列が作成できる．これを更に同次元に線形変換することで Multi-head attention の出力が作られる．

BERT はこの Multi-head attention を 12 層（あるいは 24 層）重ねたモデルである．結局，BERT は n 単語埋め込み表現列を入力とし，それをより文脈に合った n 単語埋め込み表現列に変換していると捉えることができる．

3.2 単語用例のクラスタリング

単語用例をクラスタリングするには，用例中の対象単語に対する特徴ベクトルを構築すればよい．そのためここでは，用例を BERT に入力したときに，出力される対象単語に対する埋め込みベクトルを用例の特徴ベクトルと見なすことにする (図 1 参照)．

なお日本語版 BERT 事前学習モデルとして，ここでは以下で公開されているモデルを利用することにする．

`https://drive.google.com/drive/folders/1iDlmhGgJ54rkVBtZvgMlgbuNwtFQ5OV-`

このモデルの大きな特徴は tokenizer として MeCab と NEologd (Toshinori Sato and Okumura (2016)) を利用していることである．本論文では以下このモデルを MeCab 版 BERT と呼ぶことにする．

(2) Scaled Dot-Product Attention

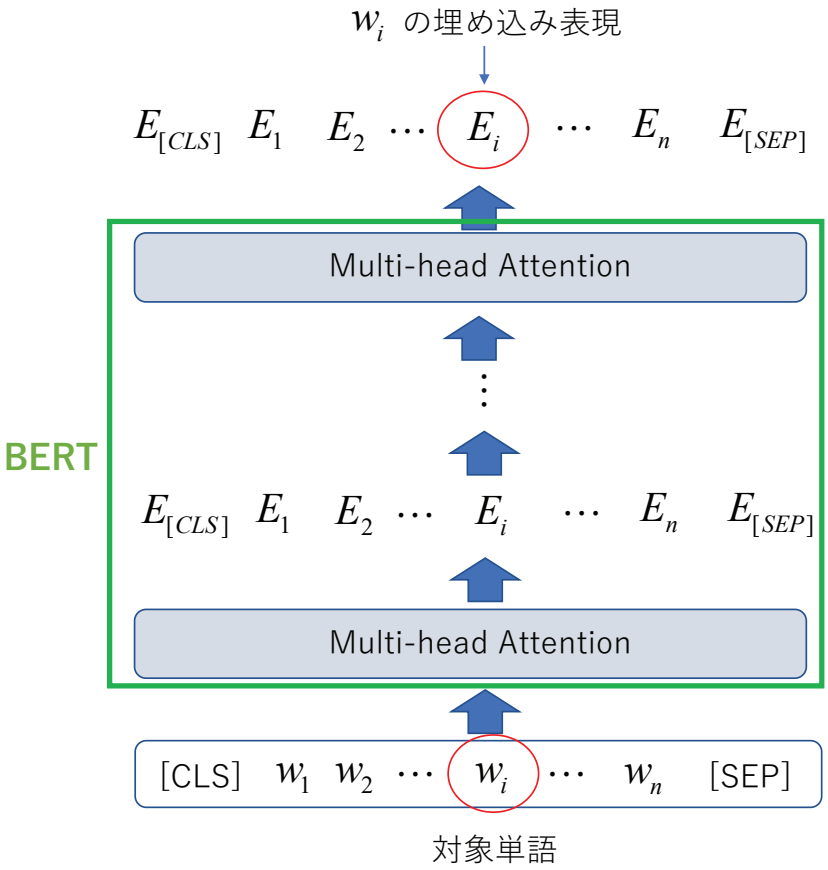

図 1　BERT による用例の特徴ベクトル

4.　実験

4.1　実験データ

対象単語として「意味」を取り上げ，その用例のクラスタリングを行う．単語「意味」は SemEval-2 の日本語 WSD タスクにおいて WSD の対象単語の 1 つになっている．そこで提供された訓練データとテストデータがラベル（語義）付きの用例として利用できる．訓練データとテストデータは 50 用例づつあるが，1 つの用例は新語義となっているので，その用例を除いた計 99 用例をクラスタリングの対象とする．

「意味」は以下の 3 つが岩波国語辞典に記載されている．クラスタリングの正解は，用例がこの 3 つのクラスタに分類されることであり，この観点からクラスタリングの評価を行う．

語義 1　その言葉の表す内容．意義．「辞書を引けば―がわかる」

語義 2　表現や行為の意図・動機．「どういう―でそんなことをしたのか」

語義 3　表現や行為のもつ価値．意義．「そんな事をしても―がない」

4.2 特徴ベクトルの作成

本論文では対象単語の用例の特徴ベクトルとして，BERT から得られる対象単語の埋め込み表現を利用する．比較のために WSD の標準的な特徴ベクトルと分散表現からの特徴ベクトルを使ったクラスタリングも行う．WSD の標準的な特徴ベクトルとしては，SemEval-2 の日本語 WSD タスクの baseline システムで用いられた特徴ベクトルである．また分散表現からの特徴ベクトルは対象単語の前後 2 単語，計 4 単語の分散表現を並べたものである (図 2 参照) (Sugawara et al. (2015))．分散表現としては nwjc2vec (新納浩幸ほか (2017)) を利用する．

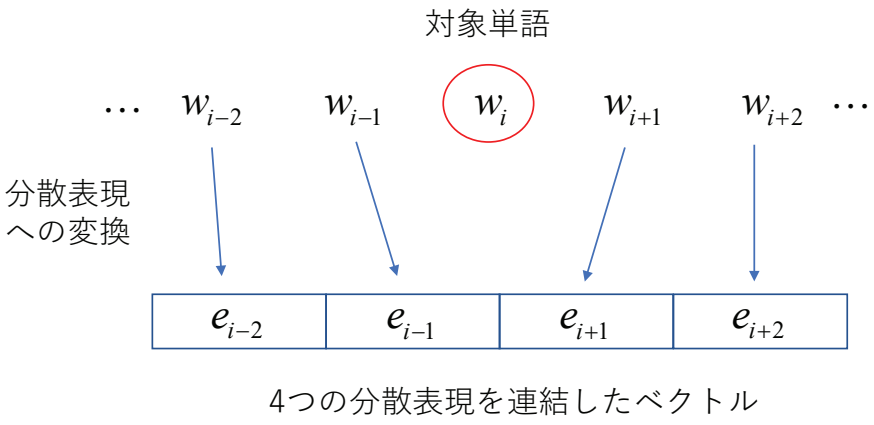

図 2　分散表現を利用した特徴ベクトル

4.3 クラスタリングとその評価

クラスタリング手法として は階層的クラスタリング手法の Ward 法，及び非階層的クラスタリング手法である k-means を用いる．どちらもクラスタ数は 3 に固定する．クラスタリングの評価にはエントロピーを用いる．

4.4 実験結果

実験の結果を表 2 に示す．表 2 の「ランダム」はクラスタの割当をランダムに行い，そのエントロピーを測る実験を 10 回行い，得られた 10 個のエントロピーの平均値である．「WSD」は WSD の標準特徴ベクトルを利用してクラスタリングした結果である．「分散表現」は対象単語の前後 2 単語，計 4 単語の分散表現を連結したベクトルを用例の特徴ベクトルとしてクラスタリングした結果である．「BERT」は BERT による対象単語の埋め込み表現を用例の特徴ベクトルとしてクラスタリングした結果である．

表 1　実験結果 (エントロピー)

	ランダム	WSD	分散表現	BERT
Ward	1.446	1.211	1.169	1.088
k-means	1.446	1.337	1.222	1.101

表 2 より BERT を用いて用例の特徴ベクトルを表した場合が最もエントロピーが低く，BERT から出力される単語の埋め込みが，WSD 用に作られた特徴ベクトルよりもよりよく意味を表現していることがわかる．

5. 考察

5.1 可視化

実験で用いた WSD の標準特徴ベクトル，分散表現を利用した特徴ベクトル，及び BERT からの特徴ベクトルをそれぞれ 2 次元に縮約し，平面上にプロットした図を以下の図 3 に示す．

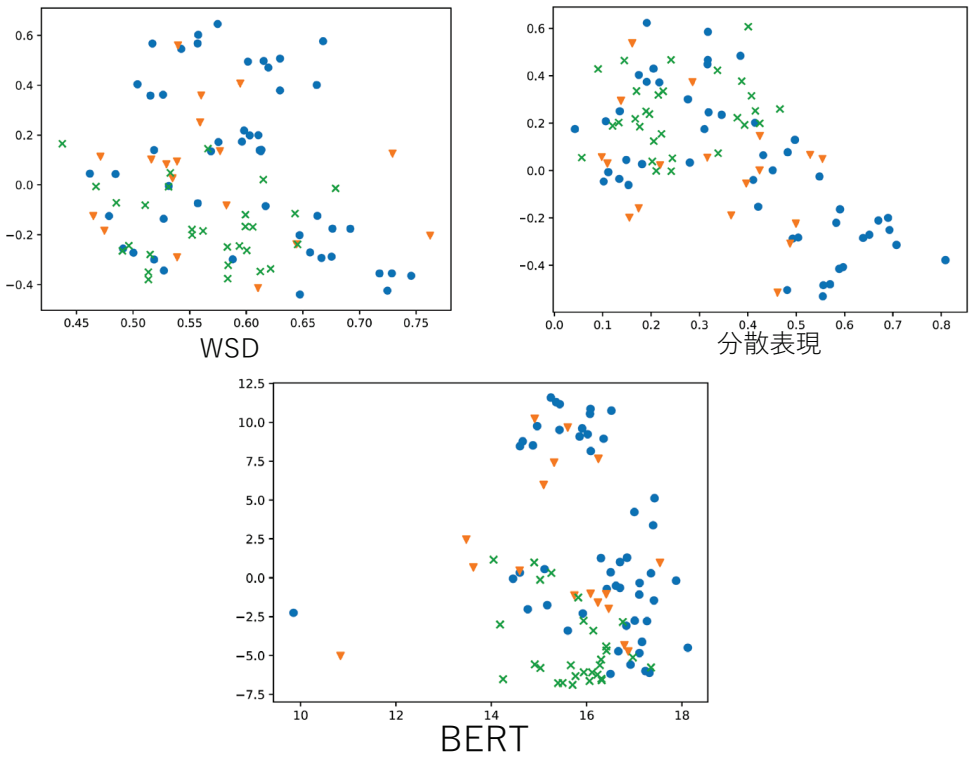

図 3　各特徴ベクトルを用いた場合のデータの分布の可視化

主観的な判断となるが，BERT の図では “×” のデータが比較的良好にクラスターを形成しているように見える．

5.2 別の事前学習モデル

実験では，事前学習モデル BERT として MeCab 版 BERT を利用したが，他にも京都大学の黒橋研究室が公開している BERT が存在する．京都大学版ののモデルでは tokenizer として Juman++ と BPE (Kudo (2018)) を利用している．本論文では以下このモデルを Juman++ 版 BERT と呼ぶことにする．

`http://nlp.ist.i.kyoto-u.ac.jp/index.php?\`
`BERT%E6%97%A5%E6%9C%AC%E8%AA%9EPretrained%E3%83%A2%E3%83%87%E3%83%AB`

ここでは Juman++ 版 BERT を利用して，先の実験と同じ実験を行った．結果を以下に示す．

表 2　事前学習モデルの比較

	MeCab 版 BERT	Juman++ 版 BERT
Ward	1.088	1.122
k-means	1.101	1.158

MeCab 版と比較すると，エントロピーはわずかに高く，本実験では MeCab 版の方がより適切に意味を表現していたことがわかる．

6. おわりに

本論文では BERT から出力される単語の埋め込み表現がその単語の意味を表現していることを確認するために，単語用例のクラスタリングを行った．具体的には単語「意味」の 99 用例を，WSD で標準的に用いられる特徴ベクトル，分散表現から作られる特徴ベクトルおよび BERT から出力される単語の埋め込み表現，それぞれを用いてクラスタリングを行い，そのエントロピーで評価した．結果，BERT から出力される単語の埋め込み表現が最もよい結果であった．このことから BERT から出力される単語の埋め込み表現が，少なくとも従来の特徴ベクトルよりもより適切に単語の意味を表現していると考えられる．今後は BERT から出力される単語の埋め込み表現を WSD や WSI に利用したい．

文　献

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *arXiv preprint arXiv:1810.04805*.

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono (2011). “On SemEval-2010 Japanese WSD Task.” 自然言語処理, 18:3, pp. 293–307.

Suresh Manandhar, Ioannis P Klapaftis, Dmitriy Dligach, and Sameer S Pradhan (2010). “SemEval-2010 task 14: Word sense induction & disambiguation.” *Proceedings of the 5th international workshop on semantic evaluation*, pp. 63–68., Association for Computational Linguistics.

Roberto Navigli, and Daniele Vannella (2013). “Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application.” *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* Vol. 2., pp. 193–201.

Mikael Kågebäck, Fredrik Johansson, Richard Johansson, and Devdatt Dubhashi (2015). “Neural context embeddings for automatic discovery of word senses.” *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pp. 25–32.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). “Improving language understanding by generative pre-training.” *Technical report, OpenAI.*

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

Jeremy Howard, and Sebastian Ruder (2018). “Universal Language Model Fine-tuning for Text Classification.” *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). “Deep Contextualized Word Representations.” *NAACL-2018*, pp. 2227–2237.

Taiichi Hashimoto Toshinori Sato, and Manabu Okumura (2016). “Operation of a word segmentation dictionary generation system called NEologd (in Japanese).” *Information Processing Society of Japan, Special Interest Group on Natural Language Processing (IPSJ-SIGNL)*, pp. NL–229–15.: Information Processing Society of Japan.

Hiromu Sugawara, Hiroya Takamura, Ryohei Sasano, and Manabu Okumura (2015). “Context representation with word embeddings for wsd.” *Conference of the Pacific Association for Computational Linguistics*, pp. 108–119., Springer.

新納浩幸・浅原正幸・古宮嘉那子・佐々木稔 (2017).「nwjc2vec:国語研日本語ウェブコーパスから構築した単語の分散表現データ」 自然言語処理, 24:5, pp. 705–720.

Taku Kudo (2018). “Subword regularization: Improving neural network translation models with multiple subword candidates.” *arXiv preprint arXiv:1804.10959.*