BERT の Masked Language Model を用いた教師なし語義曖昧性解消

新納浩幸
茨城大学大学院理工学研究科
情報科学領域
hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

馬ブン
茨城大学大学院理工学研究科
社会インフラシステム科学専攻
19nd302h@vc.ibaraki.ac.jp

1 はじめに

本論文では BERT[1] の Masked Language Model (以下 MLM と略記) を利用して，教師なしの語義曖昧性解消を試みる．対象単語をマスクし，マスク位置に推定された単語と分類語彙表から得られる候補語義を持つ類義語とを比較することで語義を推定する．

語義曖昧性解消は自然言語処理の重要な要素技術であり，古くから研究されている．近年は事前学習済みモデルの BERT を用いた語義曖昧性解消が行われている [2][3][4] [5][6][7][8]．BERT を用いることで対象単語の文脈をある空間に埋め込むことができるので，語義のラベルを同じ空間に埋め込めれば教師なし学習も可能である．

本研究では語義のラベルを同じ空間に埋め込むのではなく，BERT の MLM を利用して，対象単語の位置に現れる単語を推定することで語義曖昧性解消を行う．MLM により推定される単語は当然，対象単語である筈だが，MLM では全ての単語に対して，その単語がマスク位置に現れる確率を付与できるため，対象単語の語義を持つ類義語が現れる確率が得られる．これによって対象単語の語義を推定できる．なお本研究では対象単語の語義を持つ類義語は分類語彙表から予め得ておく．

実験では「言葉」「声」「電話」「国際」「市場」の 5 単語に対して，本手法を試み，教師なしの語義曖昧性解消手法として利用できることを確認した．

2 関連研究

語義曖昧性解消は典型的な分類問題であるため，通常の教師あり学習を用いて高精度に解決が可能である．ただし訓練データの構築コストが高いために，全単語を対象にした all-words 型の語義曖昧性解消あるいは教師なし語義曖昧性解消 [9] が研究の中心となっている．

近年は埋め込み表現を語義曖昧性解消に利用する研究が活発である．単語の埋め込み表現自体が語義曖昧性解消の素性として利用できることもあるが，語義曖昧性解消は対象単語の文脈と語義のラベルを同じ空間に埋め込むことで解決できるからである [10]．そのため word2vec のアルゴリズムを利用して語義の分散表現を構築する手法 [11][12] が提案されている．さらに BERT などの事前学習済みモデルでは，入力文の各単語に対して文脈依存の埋め込み表現を精度良く構築できるため，その単語の埋め込み表現を素性として利用するだけでも語義曖昧性解消の性能はかなり向上する [3]．またその埋め込み表現の空間に語義のラベルを埋め込めれば，k-NN 程度の学習手法でもかなりの精度が達成できる [4]．問題は語義のラベルをどのように同じ空間に埋め込むかであるが，一つの方法としては語義の用例を辞書の定義文により求め，その用例を BERT に入力することで語義のラベルの埋め込み表現が獲得できる [2]．あるいは BERT の中に意味ネットワークや意味表現を組み込むことで語義曖昧性解消が可能である [13] [5][6][7]．

本研究では語義のラベルの埋め込み表現を構築するというアプローチではなく，本質的には，文脈の埋め込み表現と類似の埋め込み表現をもつ単語を調べ，その単語の語義から対象単語の語義を推定する．結局，語義毎の用例を構築するか，語義毎の類義語を構築するかの違いとも言える．日本語の場合，前者は困難であるが，後者では分類語彙表が利用できるため，容易に実現可能である．

3 提案手法

3.1 MLM によるマスク単語の推定

BERT の学習には，2 文の接続可否を判定するタスクとマスク単語を推定するタスクが利用される．マスク単語を推定するネットワーク部分も含めて BERT のモデルを取り出したものが MLM である．MLM ではマスク単語を含む文を入力とし，そのマスク単語の位置に単語 w が現れる確率 $p(w)$ を出力する．BERT の登録語彙リストを D とした場合，$\arg\max_{w \in D} p(w)$ によりマスク単語を推定できる．

3.2 MLM を利用した語義曖昧性解消

ここでは MLM を語義曖昧性解消に利用する．まず対象単語 w の語義は $g_1, g_2, \cdots, g_K$ の K 種類とする．文 s 内に対象単語 w が現れたときに，文 s 内の単語 w の語義 g_j を識別するのが語義曖昧性解消である．ここでは文 s の w をマスク単語 [MASK] に変更し，その変更した文を MLM に入力する．出力としては語彙リスト D 内の各単語 x が [MASK] に入る確率 $p(x)$ が得られる．ここで $p(x)$ の値が高い単語 x は語彙が g_j である単語だと考えられる．そこで各語義 g_i を持つ単語のリスト R_i を予め作成しておき，以下により語義 $g_{\hat{j}}$ を推定する．

$$\hat{j} = \arg\max_j \left(\max_{x \in R_j} p(x) \right)$$

3.3 分類語彙表による語義別の類義語の設定

分類語彙表は一種の概念辞書である．各単語には 10 桁からなる分類番号が与えられている．本研究ではこの分類番号の上位 5 桁を語義 id と見なして，語義曖昧性解消を行う．語義 g_i が分類語彙表の上位 5 桁の記号である場合，語義 g_i を持つ単語のリスト R_i は分類語彙表から得ることができる．

例えば「声」には「笑い声」のような「音」としての語義と，「国民の声」のような「意見」としての語義がある．分類語彙表は「声」には '1.3031' と '1.3061' の 2 つの id が付与されている．'1.3031' が「音」の語義に対応し，'1.3061' が「意見」の語義に対応している．'1.3031' および '1.3061' の語義を持つ単語は「声」の他，分類語彙表には表 2 のよう単語が登録されている．

4 実験

4.1 実験の設定

語義曖昧性解消の対象単語は「言葉」「声」「電話」「国際」「市場」の 5 単語とする．

利用するデータは国立国語研究所から公開されている BCCWJ-WLSP[1)] である．これは『現代日本語書き言葉均衡コーパス』に分類語彙表番号が付与されたコーパスである．このコーパスから前記した 5 単語を含む文を取り出し，対象単語の語義をラベルとした評価用データを構築した．表 1 に取り出した文の数を語義毎に記す．またこの数から MFS の値が算出される．

表 1 評価用データの語義毎の文数

単語	語義のデータ数	MFS
言葉	1.3100 (87), 1.3101 (21), 1.3110 (7)	0.757
声	1.3031 (75), 1.3061 (29)	0.721
電話	1.3122 (70), 1.4620 (32)	0.686
国際	1.2530 (37), 1.3500 (50)	0.574
市場	1.2600 (72), 1.2640 (5)	0.935
(平均)	(97.0)	0.735

利用する日本語 BERT のモデルは以下で公開されている 'bert-base-japanese' とする．

https://github.com/cl-tohoku/bert-japanese

本手法では分類語彙表から得られる類義語が BERT モデルの語彙リストに登録されている必要がある．そのため各語義に対する類義語は実際に分類語彙表から得られるものの一部である．その一部を表 2 に示す．

4.2 実験結果

本手法により正解率を測った．この結果を表 3 に示す．

ランダムな解答の正解率と比較すると明らかに正解率は高く，ある程度はこの手法で正解を導けることが分かる．

語義別の正解率も調べた．この結果を表 4 に示す．

1) https://pj.ninjal.ac.jp/corpus_center/goihyo.html

表 2 語義の類義語

単語	語義	類義語
言葉	1.3100	言動, 発言, 弁, 口頭, 一言, 一口, つぶやき, 寝言, 無言, ...
	1.3101	言語, 言, 辞, キーワード, 国語, 外国語, 日本語, 中国語, ...
	1.3110	単語, 詞, 語彙, 用語
声	1.3031	音声, ボイス, 叫び, 阿鼻叫喚, 絶叫, 悲鳴, 鳴き声, とき
	1.3061	思い, 考え方, 哲学, アイディア, 発想, 構想, 感想, 妄想, 考慮, ...
電話	1.3122	通信, 返信, 返事, 便り, 手紙, メッセージ, ふみ, 書, ...
	1.4620	家電, 冷蔵庫, 冷凍庫, 洗濯機, 掃除機, 扇風機, クーラー, ...
国際	1.2530	国家, ステート, 連邦, 我が国, 祖国, 自国, 他国, 外国, 隣国, ...
	1.3500	交際, 付き合い, 交渉, 協調, 国交, 外交, 断交, ...
市場	1.2600	天下, 世の中, 世間, 新天地, 社会, 国際社会, 世界, ...
	1.2640	事務所, オフィス, 営業所, 貴社, 当社, 弊社, 同社, 自社, ...

表 3 実験結果

単語	本手法	ランダム	MFS
言葉	0.626	0.333	**0.757**
声	0.692	0.500	**0.721**
電話	**0.716**	0.500	0.686
国際	0.425	0.500	**0.574**
市場	0.688	0.500	**0.935**
(平均)	0.629	0.467	**0.735**

表 4 語義毎の正解率

単語	語義	正解率
言葉	1.3100	63/87 = 0.724
	1.3101	6/21 = 0.286
	1.3110	3/7 = 0.429
声	1.3031	46/75 = 0.613
	1.3061	26/29 = 0.897
電話	1.3122	53/70 = 0.757
	1.4620	20/32 = 0.625
国際	1.2530	27/37 = 0.730
	1.3500	10/50 = 0.200
市場	1.2600	48/72 = 0.923
	1.2640	5/5 = 1.000
平均		0.653

各語義に対してある程度の正解率を示しており，語義の頻度に影響を受けていないことが確認できる．

5 考察

5.1 類義語の調整

類義語のリストを調整することで正解率を改善できると考えた．「国際」の語義 1.3500 と「言葉」の語義 1.3101 については極端に正解率が悪い．これらの語義の判定で誤りに導いた類義語について調べた．「国際」の語義 1.3500 は 50 個のデータ中 40 個が間違いであった．この間違いを導いた類義語は以下の通りである．

「国家」16 個，「外国」10 個，「各国」5 個，「連邦」 4 個，その他 5 個

そこで類義語のリストから「国家」「外国」を除いて，「国際」について再実験を行ったところ，1.3500 の正解数は 3 つ増えたが，1.2530 の正解数が 2 つ減り，総合的には大きな変化はなかった．類義語のリストを調整することでの正解率の改善は難しいと考えられる．

5.2 他の BERT モデルの利用

本実験では他の BERT モデルを利用することで，正解率を改善できる可能性がある．ただし BERT 毎に Tokenizer が異なるために，単純にモデルだけを交換して実験することはできず，公正な比較は難しい．ただし本実験で用いた ’bert-base-japanese’ と同じサイトで公開されている’bert-base-japanese-whole-word-masking’ は同じ Tokenizer を用いているためにモデルだけを変更して本実験が行える．実験の結果を表 5 に示す．

表 5 他の BERT モデルの利用

単語	本手法 (base model)	whole-word model
言葉	0.626	**0.643**
声	**0.692**	0.644
電話	**0.716**	0.647
国際	**0.425**	**0.425**
市場	0.688	**0.727**
(平均)	**0.629**	0.617

一般に ’bert-base-japanese-whole-word-masking’ は’bert-base-japanese’ よりも少し性能が高いと言われているが，このタスクに対しては性能は低かっ

た．本タスクにはどのような BERT モデルが適しているのかを調べるのは今後の課題である．

5.3 BERT の追加学習

本実験で利用した BERT は素の BERT であり，本タスクに適した形で fine-tuning されたものではない．本タスクに適した形で BERT を fine-tuning しておけば，正解率の改善に繋がると考えられる．ただし本タスクは MLM のタスクそのものであるため，BERT の fine-tuning は BERT の追加学習という形になる．

ここでは実験的に 2008 年度毎日新聞から約 80 万文を利用して BERT の追加学習を行った．作成できたモデルを用いた実験結果を表 6 に示す．

表 6 追加学習した BERT の利用

単語	本手法 (base model)	追加学習 BERT
言葉	**0.626**	0.400
声	**0.692**	0.385
電話	**0.716**	0.412
国際	0.425	**0.494**
市場	**0.688**	0.403
(平均)	**0.629**	0.419

性能は大幅に下がっている．ランダムの結果 (0.467) よりも悪い．追加学習に失敗していると思われるが，原因は不明であり調査中である．BERT の追加学習に関しては何らかの Tips があると思われる．

本タスクに対して BERT の追加学習は有効だと予想しているので，この方向で研究を継続したい．

6 おわりに

本論文では BERT の MLM を利用して，教師なしの語義曖昧性解消を試みた．対象単語をマスクし，マスク位置に推定された単語と分類語彙表から得られる候補語義を持つ類義語とを比較することで語義を推定する．実験では語義タグ付きコーパスである BCCWJ-WLSP から 5 単語を選び，評価用データを作成し，本手法を試みた．性能は低かったが，教師なしの語義曖昧性解消手法として利用できることを確認した．また利用する BERT を追加学習することで，性能は改善できると予想している．その方向で本研究を継続したい．

謝辞

本研究は JSPS 科研費 JP19K12093 および 2020 年度国立情報学研究所公募型共同研究（2020-FC03）の助成を受けています．

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. In *NAACL-2019*, pp. 4171–4186, 2019.

[2]Luyao Huang Chi Sun Xipeng Qiu Xuanjing Huang. GlossBERT: BERT for word sense disambiguation with gloss knowledge. *arXiv preprint arXiv:1908.07245*, 2019.

[3]Hadiwinoto, Christian Ng, Hwee Tou Gan, Wee Chung. Improved word sense disambiguation using pre-trained contextualized word representations. In *EMNLP-IJCNLP-2019*, pp. 5300–5309, 2019.

[4]Gregor Wiedemann Steffen Remus Avi Chawla Chris Biemann. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*, 2019.

[5]Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. *AAAI-2020*, pp. 8758–8765, 2020.

[6]Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. Semantics-aware BERT for language understanding. In *AAAI-2020*, pp. 9628–9635, 2020.

[7]Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. SenseBERT: Driving some sense into BERT. In *ACL-2020*, pp. 4656–4667, 2020.

[8]Tommaso Pasini, Federico Scozzafava, and Bianca Scarlini. CluBERT: A cluster-based approach for learning sense distributions in multiple languages. In *ACL-2020*, pp. 4008–4018, 2020.

[9]鈴木類, 古宮嘉那子, 浅原正幸, 佐々木稔, 新納浩幸. 概念辞書の類義語と分散表現を利用した教師なし all-words wsd. 自然言語処理, Vol. 26, No. 2, pp. 361–379, 2019.

[10]Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. Embeddings for word sense disambiguation: An evaluation study. In *ACL-2016*, pp. 897–907, 2016.

[11]Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A unified model for word sense representation and disambiguation. In *EMNLP-2014*, pp. 1025–1035, 2014.

[12]Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*, 2015.

[13]Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. SenseBERT: Driving some sense into BERT. In *ACL-2020*, pp. 4656–4667, 2020.