静的な単語埋め込みによるカタカナ語を対象としたBERTの語彙拡張

平子 潤
名古屋大学 情報学部
hirako.jun@g.mbox.nagoya-u.ac.jp

笹野 遼平　武田 浩一
名古屋大学 大学院情報学研究科
{sasano,takedasu}@i.nagoya-u.ac.jp

1 はじめに

BERT[1] ではサブワードを利用することで語彙に含まれない未知語を含む文も適切に処理することが可能となっている．しかし，サブワードが意味を持つ適切な単位となっていない場合，その有効性は限定的であると考えられる．こうしたケースはカタカナ語の未知語において特に多く出現すると言える．たとえば本研究で使用した BERT (3.2 節参照) では，「クロワッサン」は「クロ」「##ワ」「##ッサン」の 3 つのサブワードに分割されるが，これらは意味のある単位とはなっていない．

このような未知語を含む文を適切に扱えるようにする 1 つの方法は語彙サイズを大きく設定し，これらの語を語彙に含めてしまうことである．しかし，BERT の事前学習時の語彙サイズを大きくした場合，計算コストが増大してしまう．そこで，本研究では比較的低コストに学習できる静的な単語埋め込みを用いた BERT の語彙拡張手法を提案する．

2 単語埋め込みのマッピング

BERT は，図 1 に示すように語彙に含まれる各単語に対し文脈に依存しない単語埋め込み (BERT Token Embeddings) を持っている．この埋め込み層を拡張することで BERT の語彙の拡張が可能となる．本研究では，より大きな語彙サイズで学習した静的な単語埋め込みを，BERT Token Embeddings と共通する語を手掛かりにマッピングすることで，事前学習済みの BERT の語彙を拡張する．

2.1 既存のマッピング手法

異なる単語埋め込み空間のマッピングは，2 言語間の単語埋め込みの対応付けを目的として行われる場合が多い [2]．教師あり手法の場合，2 言語辞書を用いて対応付けるべき単語ペアを獲得し，各単語ペアの単語ベクトルが近くなるようにソース単語埋め込みをターゲット単語空間にマッピングする．このような手法の先駆けとして，Mikolov ら [3] は，マップ後のソース単語ベクトルとターゲット単語ベクトルの平均二乗誤差が小さくなるような変換行列 W を SGD で求める手法を提案した．また，Xing ら [4] は，ソースとターゲットの全単語ベクトルのノルムを正規化し，変換を直交変換に限定する手法を提案した．さらに，Artetxe ら [5] は，ノルムの正規化や直交変換に加えて，分散の正規化，再重み付け，次元削減を複数ステップで行う枠組みを提案した．

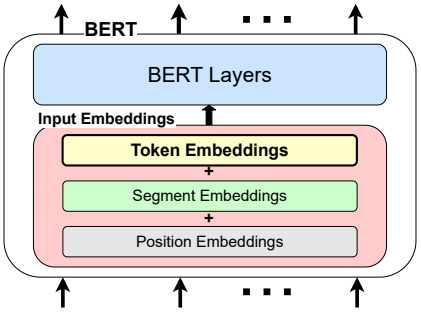

図 1 BERT の構造．BERT Layers は BERT の Input Embeddings 以外の層を表している．

2.2 静的な単語埋め込みからのマッピング

本研究で目的とする，静的な単語埋め込みから BERT Token Embeddings へのマッピングが，既存の言語間の単語埋め込みのマッピングと大きく異なっている点として，言語間マッピングでは基本的に各言語における単語埋め込み生成アルゴリズムは同一のものが使用されるのに対し，本研究におけるマッピングではソース単語埋め込みとターゲット単語埋め込みの生成アルゴリズムが異なっていることが挙げられる．したがって，マッピング対象とする 2 つの埋め込み空間の性質が大きく異なる可能性が考えられることから，本研究で使用した BERT Token Embeddings と，fastText で学習した静的な単語埋め込み (3.2 節参照) の両方に含まれる単語のノルムの関係を調査した．結果を図 2，図 3 に示す．BERT

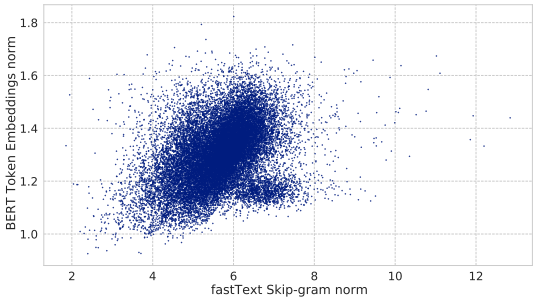

図 2 BERT Token Embeddings と fastText Skip-gram で共通する単語のノルムの関係．相関係数は 0.355.

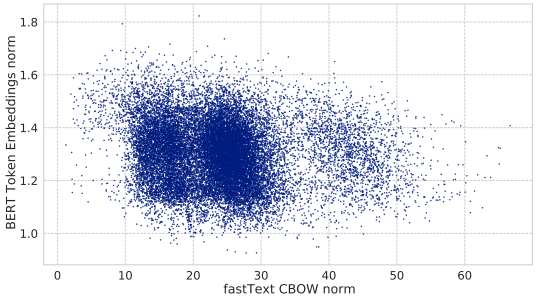

図 3 BERT Token Embeddings と fastText CBOW で共通する単語のノルムの関係．相関係数は-0.142.

Token Embeddings と fastText で学習した静的な単語埋め込みの間で，ノルムに強い相関は確認できず，また，平均値も大きく異なっていることが分かる．[1]

これらの結果から，単純な変換行列に基づくマッピングを行った場合，ノルムの性質の違いを要因とし精度低下が起こる可能性が考えられる．そこで，単純な変換行列に基づくマッピング (補正なし) に加え，ノルムの性質の違いを考慮した 3 つのマッピング手法を含めた下記の 4 つのマッピング手法を考える．これらの手法において直交変換は，変換後のソース単語埋め込みとターゲット単語埋め込みの平均二乗誤差が最小となるように行う．

1. **補正なし**: 静的な単語埋め込み M を直交変換
2. **平均補正**: 静的な単語埋め込み M の平均ノルムが BERT Token Embeddings の平均ノルムと一致するよう M を定数倍してから直交変換
3. **正規化**: 静的な単語埋め込み M と BERT Token Embeddings の全ベクトルを正規化してから直交変換
4. **正規化＋補正**: 静的な単語埋め込み M と BERT Token Embeddings の全ベクトルを正規化してから直交変換した後，BERT Token Embeddings の平均ノルムで変換後の M を定数倍

[1]マッピングの学習に使用した単語に対応するベクトルのノルムの平均値は BERT Token Embeddings は 1.30，fastText Skip-gram は 5.67， fastText CBOW は 23.97 であった．

本研究では，サブワードが適切に機能しないケースが多いと考えられるカタカナ語を対象に，これら 4 つのマッピング手法を用い fastText で学習した静的な単語埋め込みを BERT Token Embeddings にマッピングすることで BERT の語彙拡張を行う．

3 実験

本研究では語彙拡張の有効性をマスク単語予測実験を通して検証する．具体的には，語彙拡張が有効であるならば，拡張した語を含む文において拡張語と依存関係の強い周辺語のマスク単語予測の精度が，語彙拡張前の BERT を使った場合，すなわち拡張語をサブワードの系列として扱った場合と比べ向上するとの仮説に基づき，マスク単語予測データセットを作成し，それらのタスクの精度を通して語彙拡張の有効性を検証する．

3.1 マスク単語予測データセット

語彙拡張の有効性検証のため，上位語予測，兄弟語予測，述語予測，被連体語予測の 4 タイプのマスク単語予測データセットを作成した．以下の文 (1)〜(4) はそれぞれのタイプの例となっている．

(1) <カボチャ>は [野菜] の一種です。
(2) 「<スプーン>」と「[フォーク]」。
(3) <フローラル>の [香り] が気に入った。
(4) 久しぶりに<クロワッサン>を [食べる]。

これらの例において，’<>’ で囲った語が拡張語，’[]’ で囲った語がマスク予測対象語となっており，いずれの場合もマスク予測対象語は拡張前の BERT の語彙に含まれている語である．各タイプのデータセットの作成方法と評価指標を以下にまとめる．

上位語予測 拡張語の上位語が予測対象語となるようにデータセットを作成した．入力文は『<拡張語>は [MASK] の一種です。』とし，対象とする拡張語は，日本語 WordNet[6] に含まれ，拡張語を含む synset の上位 synset に BERT の語彙に含まれる語が存在した語とした．条件を満たした拡張語は 2730 個であった．正解となる上位語は複数存在する可能性があることから評価指標には Mean Average Precision (MAP) を用いた．

兄弟語予測 拡張語の兄弟語が予測対象語となるようにデータセットを作成した．入力文は『「<拡張語>」と「[MASK]」。』とし，対象とする拡張語は，日本語 WordNet に含まれ，拡張語を含む synset の上位 synset の下位 synset から，拡張語が含まれる

表 1　マスク単語予測実験の結果 (太字は拡張なしより高いスコアであることを表す)

予測語タイプ (評価指標)	拡張なし	拡張あり (Skip-gram)				拡張あり (CBOW)			
		補正なし	平均補正	正規化	正規化+補正	補正なし	平均補正	正規化	正規化+補正
上位語 (MAP)	0.109	0.058	**0.135**	**0.138**	**0.137**	0.039	**0.121**	**0.127**	**0.125**
兄弟語 (MAP)	0.092	0.091	**0.121**	**0.118**	**0.121**	0.061	**0.114**	**0.122**	**0.123**
被連体語 (MRR)	0.381	0.317	0.374	0.374	0.374	0.299	0.370	0.371	0.372
述語 (MRR)	0.481	0.346	0.458	0.459	0.458	0.309	0.463	0.465	0.463

synset を除いた synset (兄弟 synset) に BERT の語彙に含まれる語が存在した語とした．条件を満たした拡張語は 3049 個であった．上位語予測と同様に評価指標には Mean Average Precision (MAP) を用いた．

被連体語予測　拡張語が連体修飾する語が予測対象語となるようにデータセットを作成した．具体的には，Common Crawl に含まれるテキスト[2]を，MeCab[3]を用いて形態素解析し，そこから『<拡張語>の<名詞><格助詞>』というパターンを含む文章を抽出し，BERT の語彙に含まれる<名詞>を [MASK] に置き換えた 1 万文を使用した．正解単語は置き換え前の<名詞>とし，正解が 1 つであることから，評価指標には平均逆順位 (MRR) を用いた．

述語予測　拡張語を項とする述語が予測対象語となるようにデータセットを作成した．具体的には，Common Crawl から『<拡張語><格助詞><動詞>。』というパターンを含む文を抽出し，BERT の語彙に含まれている<動詞>を [MASK] に置き換えた 1 万文を使用した．ただし，情報量の少ない「する」や「なる」などの述語を除くため，平仮名のみで構成される<動詞>を含む文は対象から外した．被連体語予測と同様に評価指標には平均逆順位 (MRR) を用いた．

3.2　実験設定

静的な単語埋め込みの作成には，fastText[4]を利用し，Skip-gram と CBOW[7] の 2 つのモデルで学習を行った．コーパスは Wikipedia 日本語版を利用し，MeCab[3]により単語分割を行った．次元数は BERT Token Embeddings と同じ 768 次元とした．また，事前学習済みの BERT として，東北大学で公開されているモデル[5]を利用した．

実験では，語彙拡張を行わない BERT をベースラインとし，4 つのマッピング手法 (補正なし，平均補正，正規化，正規化+補正) で語彙拡張した BERT を比較した．マッピングの学習には，BERT の語彙のうちサブワードトークン (「##」から始まる) と特殊トークン ([MASK] や [CLS] など) を除く，すべての単語を使用した．拡張対象語彙は，静的な単語埋め込みの学習に使用した Wikipedia において， BERT の語彙に含まれない出現頻度上位 32000 個のカタカナ語とした．ただし，カタカナ語とは，MeCab によって 1 語としてまとめられた，カタカナのみで構成される単語である．32000 語を語彙に追加することで，Wikipedia のカタカナ語の頻度ベースのカバー率は 71.3%から 90.7%となった．

3.3　実験結果

表 1 に実験結果を示す．上位語予測と兄弟語予測では，提案手法のうちノルムの違いを考慮した 3 手法はベースライン (拡張なし) を上回った．一方，被連体語予測と述語予測では，いずれの手法を用いてもベースラインを上回ることができなかった．これについては次節で詳細な分析を行う．

ノルムの補正を行わない手法 (補正なし) と行う手法を比較すると，補正を行うことで精度が大幅に向上することが確認できる．一方，ノルムの補正を行う 3 手法間では大きな精度の違いは見られなかった．このことは，図 2，図 3 で示したように， BERT Token Embeddings と静的な単語埋め込みのノルムの性質が大きく異なるため，ノルムの平均の補正や正規化を行わなずに直交変換するだけでは適切にマッピングできないことを示している．

マッピングに使用した静的な単語埋め込みの学習に利用した 2 モデルを比較すると，ノルムの補正を行わなかった場合 (補正なし) は，BERT Token Embeddings とのノルムの相関が比較的大きい Skip-gram を用いた場合の方が，CBOW を用いた場合より高い精度となった．一方，ノルムの補正を行った場合は静的な単語埋め込みの違いによる精度差はほとんど確認できなかった．

4　単語構成と出現頻度による分析

被連体語予測および述語予測において，提案手法がベースラインを上回ることができなかったのは，大きく 2 つの要因があると考えられる．

[2] http://statmt.org/ngrams/で公開されているデータを利用した．
[3] https://taku910.github.io/mecab/ (辞書: mecab-ipadic)
[4] https://github.com/facebookresearch/fastText
[5] https://github.com/cl-tohoku/bert-japanese で公開されているモデルのうち，Wikipedia を MeCab (ipadic) と WordPiece で単語分割し， Whole Word Masking で学習したモデルを利用した．

1 つ目は，単語に構成性があることである．たとえば拡張したカタカナ語に含まれる「ビデオカメラ」は，「ビデオ」と「##カメラ」という 2 つのサブワードに分割されるが，これらはそれぞれ意味を持つ形態素となっている．このようにより細かい形態素に分解可能な語は，カタカナ語であっても意味のあるサブワードに分割することができるため，サブワードが有効に働くと考えられる．

2 つ目は，出現頻度の高いカタカナ語の存在である．高い頻度で出現するカタカナ語の場合，人間が見た場合に意味のあるサブワードに分割することができなかったとしても，それらのカタカナ語の意味を適切に処理できるようにサブワードの埋め込みの学習が行われている可能性がある．たとえば，拡張したカタカナ語に含まれる「ペンギン」は，Wikipedia において 1900 回以上という高い頻度で出現していることから，「ペンギン」を構成する「ペン」および「##ギン」というサブワードは「ペンギン」に近い意味となるように学習が行われている可能性がある．

これら 2 つの要因による影響を検証するため，拡張語を単語構成と出現頻度に基づき分類した 8 つのタイプごとに，被連体語予測と述語予測においてカタカナ語を追加する効果を調査した．具体的には，下記の基準によりカタカナ語を分類し，Common Crawl から各タイプに該当する文を 2500 文ずつ抽出し，正規化+補正手法により語彙を拡張する場合と，しない場合の精度を比較した．

単語構成による分類　Unidic[8] が採用している，短い語の単位である短単位に基づきカタカナ語を形態素解析し[6]，Unidic に含まれる複数の語に分割されたものを複合語，それ以外の語を単純語に分類

出現頻度による分類　32000 語のカタカナ語を，Wikipedia における出現頻度により上位 25%，上位 25~50%，上位 50~75%，下位 25%の 4 つに分類

実験結果を表 2 に示す．表 2 において太字はベースラインより高いスコアであることを示す．単純語かつ出現頻度が高くない単語に対しては，ほとんどの場合において，拡張ありのスコアが拡張なしのスコアを上回っており，語彙拡張が有効であることが確認できる．このことから，拡張するカタカナ語を単純語かつ高頻度でない単語に限定すれば，有用な単語拡張が可能であると考えられる．一方，合成語，および，単純語であっても頻度上位の語については拡張することでスコアが低下しており，被連体語予測および述語予測においてカタカナ語を対象とした語彙拡張によるスコアの向上が確認できなかった要因は，構成的な単語および高頻度語を対象とした語彙拡張が有効でなかったためであると言える．

[6]解析器として MeCab，辞書として https://unidic.ninjal.ac.jp/unidic_archive/cwj/2.1.2/unidic-mecab-2.1.2_bin.zip を用いた．

表 2　単語構成と出現頻度による分析 (MRR)

予測語	単語構成	出現頻度	拡張なし	拡張あり	
				fastText	
				Skip-gram	CBOW
被連体語	単純語	上位 25%	0.410	0.339	0.393
		上位 25~50%	0.348	**0.363**	**0.357**
		上位 50~75%	0.334	**0.348**	**0.343**
		下位 25%	0.354	**0.359**	**0.356**
	合成語	上位 25%	0.435	0.403	0.406
		上位 25~50%	0.343	0.330	0.335
		上位 50~75%	0.387	**0.392**	0.385
		下位 25%	0.357	0.351	0.353
述語	単純語	上位 25%	0.507	0.473	0.478
		上位 25~50%	0.417	**0.430**	**0.432**
		上位 50~75%	0.364	**0.376**	**0.369**
		下位 25%	0.362	**0.364**	0.349
	合成語	上位 25%	0.470	0.438	0.449
		上位 25~50%	0.464	0.435	0.440
		上位 50~75%	0.482	0.473	0.478
		下位 25%	0.391	0.389	0.382

表 3　マスク単語予測において改善した例

入力文	拡張なし	拡張あり
久しぶりにクロワッサンを [MASK]。	用いる	食べる
みなさんと、ヴァチカンに [MASK]。	出会う	向かう
サルスベリの [MASK] を見る。	顔	花
フローラルの [MASK] が気に入った。	音色	香り

表 3 に BERT の語彙を拡張したことによる改善例を示す．語彙拡張を行うことで「クロワッサン」などの語の意味を適切に捉えることができるようになっていることが確認できる．

5　おわりに

本研究では，静的な単語埋め込みから BERT Token Embeddings へマッピングを行うことで， BERT の語彙を拡張する手法を提案した．さらに，マスク単語予測データセットを用いて，カタカナ語を対象とした語彙拡張の有効性の検証を行い，単純語かつ出現頻度が高くないカタカナ語については語彙拡張が有効であることを示した．今後の課題としては，構成的でない漢字語などカタカナ語以外に対する提案手法の有効性の検証や，本研究で用いた直交変換に基づく手法より適したマッピング手法の解明などが考えられる．

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 *Proc. of NAACL'19*, pp. 4171–4186, 2019.

[2] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, Vol. 65, p. 569–631, 2019.

[3] Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv:1309.4168,* 2013.

[4] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proc. of NAACL'15*, pp. 1006–1011, 2015.

[5] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proc. of AAAI'18*, 2018.

[6] Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. Enhancing the Japanese WordNet. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)*, pp. 1–8, 2009.

[7] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proc. of ICLR'13, Workshop Track Proceedings*, 2013.

[8] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源 : 形態素解析用電子化辞書の開発とその応用. 日本語科学, Vol. 22, pp. 101–123, 2007.