Masked Language Model を用いた語群から語の連想の検討

相馬佑哉，堀内靖雄，黒岩眞吾
千葉大学
yuya10101@chiba-u.jp，{hory, kuroiwa}@faculty.chiba-u.jp

概要

本稿では，Masked Language Model を用いて 5 つの刺激語から 1 つの正解連想語を予測させるタスクの検討を行った. 使用したモデルは BERT および gMLP である．対象とした語連想タスクでは，刺激語・連想語ともに名詞のみを用いていることから，本稿で新たに学習したモデルでは MASK するトークンを名詞に限定した．実験では「富士山、浜名湖、…、うなぎから連想する都道府県は MASK です。」等の文を与え，MASK に連想語が出力されるようにした. この際，MASK の前後に鍵括弧「」を付与することも検討した．実験の結果，MASK の前後に「」を付与することで最も高い正答率 49%（上位 5 語以内に正解が含まれる率 72%）が得られた.

1 はじめに

本稿では日本語 Wikipedia で学習した BERT[1]と gMLP[2]を用いて語群から語の連想を行った．本稿における語連想とは，単語を 5 つ与えて（刺激語），連想する単語（連想語）を 1 つ回答させるタスクである.

複数の刺激語から 1 つの連想語を出力する研究として, 米谷ら[3], 白川ら[4], 芋野ら[5]の研究がある．米谷らは感覚判断システム，白川らはコンテキストを考慮した関連エンティティ，芋野らは語概念連想を用いて，各々人間の連想を模擬できるかの検証を行った．これらに対し，我々は BERT を用いて 1 つの刺激語から複数の連想語を出力する研究[6][7]を行ってきた．一方で，近年，BERT の学習タスクである Masked Language Model（以下，MLM と表記）において, タスクに対応する単語を MASK することでそのタスクの精度を向上させる手法が提案されている[8][9]. そこで，本稿では複数の刺激語から 1 つの連想語を出力する課題に BERT を代表とする MLM の適用を試みる. 具体的には,「＜刺激語 1＞、…、<刺激語 5>から連想する都道府県は MASK です。」等の文を与え, MASK に連想語が出力されるようにした．また，本稿の語連想では名詞のみを扱うため，通常の BERT に加え，名詞のみを MASK して学習させた BERT と gMLP を用いて，複数の刺激語から 1 つの連想語を出力する実験を行った.

2 語群から語の連想課題

本稿では, 実験に用いる語連想課題として『CD 版そのまま使える失語症教材 2』[10] 中の教材の『名詞の想起』より，5 つの語から特定の語を想起させる『まとめる語想起』を利用した. 例を以下に示す.

次の言葉から連想する都道府県はどこですか。
富士山 浜名湖 茶畑 みかん うなぎ → 静岡県

実験では,『まとめる語想起』132 題のうち，20 代男子大学生 3 人の解答が正解の連想語（以下，正解語）と一致し，なおかつ 3.3 のモデルで正解語が 1 つの MASK トークンで出力可能な 87 題を採用した. また，課題の種類ごとに『上位下位』『部分全体』『都道府県』等の 13 カテゴリに分類した（筆者らによる独自の分類）. 課題の詳細を表 1 に示す.

3 Masked Language Model

本稿では MLM として BERT と gMLP を使用した.

3.1 BERT

BERT[1] は Bidirectional Encoder Representations from Transformers であり，Transformer をベースに構成されたモデルである．BERT は 2 つのタスクで学習を行い, その 1 つである Masked Language Modeling は「トマトの色は赤色である」を「トマトの色は MASK である」のように単語を MASK し，その単語を予測するタスクである．12 層の Transformer 層の後で, 1 層の全結合線形層により正解ラベル（単語）のスコアが大きくなるように学習する.

3.2 gMLP

gMLP[2]は Transformer の Attention 機構の特徴であるトークン間の情報を学習できる点を MLP（Multi-Layer Perceptron）で表現したモデルであり，Attention の代替として SGU（Spatial Gating Unit）を採用したモデルである．入出力層を BERT と同一にすることで，MLM タスクを解くことができる.

3.3 実験で使用する MLM

実験では, (1)東北大学 乾・鈴木研究室の Wikipedia で訓練済み日本語 BERT モデル(BERT-base_mecab-ipadic-bpe-32k_whole-word-mask)[11]（東北大 BERT と表記），(2)Wikipedia データセットを用いて MASK する単語を名詞に限定して学習した日本語 BERT モデル（名詞のみ BERT と表記），(3)日本語 gMLP モデル（名詞のみ gMLP と表記）を使用した．モデルの詳細を付録・表 3 に示す. なお, (1)は Next Sentence Prediction（NSP）も学習を行っているが，(2)と(3)は MLM のみで学習を行った.

4 MLM を用いた語連想手法

本稿では，MLM を用いて MASK が連想語となる文（以下, 連想文と表記）を作成して実験を行った．以下に作成した文の例を示す.

<刺激語 1>、*<刺激語 2>*、*<刺激語 3>*、*<刺激語 4>*、*<刺激語 5>から連想する都道府県は MASK です。*

正解語が『静岡県』の場合，刺激語 1～5 には富士山や浜名湖等の単語が入り，MASK は連想語が入るトークンとなる．また，本項では刺激語と連想語として名詞のみを用いるため，MASK として予測された単語のうち名詞（代名詞を除く）のみを出力する.

表 1 に実験で使用した連想文を示す. 表中の『〇』は刺激語を表す．この連想文は文献[10]を基に作成した．また，文献[6]において連想結果が改善されたことに鑑み，本稿では，表 1 に加え MASK に鍵括弧「」を付与した連想文も使用した. 以下に例を示す.

<刺激語 1>、*<刺激語 2>*、*<刺激語 3>*、*<刺激語 4>*、*<刺激語 5>から連想する都道府県は「MASK」です。*

5 語群から語の連想実験

表 1 で示した連想文を用いて,「」の有無の両パターンで語群から語の連想実験を行った.

5.1 実験結果

表 2 に実験結果を示す．表はモデルの連想語上位 1 語，および 5 語以内に正解語が出力された課題の数の割合を正答率として示したものであり，正答率の最大値は 1 である．カテゴリ欄のカッコ内は各カテゴリの課題の総数である．上位 1 語で最も高い正答率は「」有の BERT と名詞のみ gMLP の 0.49 であり，約半数を正解した．また，上位 5 語以内で最も高い正答率は「」有の名詞のみ gMLP の 0.72 である．同様の設定で 1 つの刺激語から複数の連想語を出力させた実験[6]では，人間の連想語上位 4 語と一致した数は最大でも平均 0.6/4.0 語であり，複数の刺激語から 1 つの連想語を出力するタスクの方が精度が高いと言える.

また,「」を付与することで，全てのモデルで正解語が 1 位に出力されやすくなった．東北大 BERT では「」の付与によって正解語が 1 位に出力された課題が 28 題，1 位に出力されなくなった課題が 3 題であり，符号検定（有意水準 5%）により，連想結果が改善されたと言える．また，名詞のみ gMLP では改善 21 題，改悪 3 題であり，符号検定（有意水準 5%）により, 同様に連想結果が改善されたと言える．一方で，モデル間の正答率に有意差はなかった.

表 1　実験で使用する連想文

カテゴリ	連想文
上位下位	〇、〇、〇、〇、〇は MASK の仲間です。
部分全体	〇、〇、〇、〇、〇からできているものは MASK です。
色	〇、〇、〇、〇、〇から連想する色は MASK です。
都道府県	〇、〇、〇、〇、〇から連想する都道府県は MASK です。
国	〇、〇、〇、〇、〇から連想する国は MASK です。
スポーツ	〇、〇、〇、〇、〇から想像できるスポーツは MASK です。
季節	〇、〇、〇、〇、〇から連想する季節は MASK です。
場所	〇、〇、〇、〇、〇から連想する場所は MASK です。
家の中の場所	家の中で〇、〇、〇、〇、〇がある場所は MASK です。
行事	〇、〇、〇、〇、〇から連想される行事は MASK です。
使ってすること	〇、〇、〇、〇、〇を使ってすることは MASK です。
どんなときに持つ	〇、〇、〇、〇、〇は MASK のときに持っていきます。
メニュー	〇、〇、〇、〇、〇から連想されるメニューは MASK です。

表 2　連想語上位 1 語に正解語が出力された課題数の割合（カッコ内は上位 5 語以内）

	鍵括弧「」無			鍵括弧「」有		
課題のカテゴリ（課題数）	東北大 BERT	名詞のみ BERT	名詞のみ gMLP	東北大 BERT	名詞のみ BERT	名詞のみ gMLP
全体（87）	0.21 (0.46)	0.34 (0.52)	0.29 (0.47)	**0.49** (0.69)	0.44 (0.66)	**0.49** (**0.72**)
上位下位（38）	0.32 (0.50)	0.24 (0.34)	0.34 (0.45)	0.39 (0.53)	0.32 (0.45)	**0.42** (**0.55**)
部分全体（6）	0.00 (0.00)	0.33 (0.33)	0.00 (0.17)	0.33 (0.83)	**0.83** (0.83)	0.67 (**1.00**)
色（8）	0.13 (0.88)	0.25 (0.75)	0.13 (0.88)	**0.75** (**1.00**)	0.50 (0.88)	0.25 (0.63)
都道府県（8）	0.13 (0.25)	0.38 (0.63)	0.13 (0.50)	0.63 (0.88)	0.38 (0.88)	**0.75** (**1.00**)
国（4）	0.25 (0.75)	**1.00** (**1.00**)	**1.00** (**1.00**)	**1.00** (**1.00**)	**1.00** (**1.00**)	**1.00** (**1.00**)
スポーツ（4）	0.75 (**1.00**)	**1.00** (**1.00**)	0.75 (0.75)	**1.00** (**1.00**)	**1.00** (**1.00**)	0.75 (**1.00**)
季節（4）	0.00 (0.75)	0.25 (**1.00**)	0.25 (0.75)	**0.50** (0.75)	**0.50** (**1.00**)	**0.50** (**1.00**)
場所（4）	0.00 (0.50)	**0.75** (**1.00**)	0.25 (0.50)	**0.75** (**1.00**)	**0.75** (**1.00**)	0.50 (**1.00**)
家の中の場所（4）	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	**0.25** (**0.50**)	0.00 (**0.50**)	0.00 (**0.50**)
行事（3）	0.00 (0.00)	0.33 (**0.67**)	0.33 (0.33)	0.33 (**0.67**)	0.33 (**0.67**)	**0.67** (**0.67**)
使ってすること（2）	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	**1.00** (**1.00**)
どんなときに持つ（1）	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
メニュー（1）	0.00 (0.00)	**1.00** (**1.00**)	0.00 (0.00)	0.00 (**1.00**)	0.00 (**1.00**)	0.00 (**1.00**)

5.2 結果の分析

全てのモデルで「」の付与により正答率が改善したことから，まずは「」の効果について分析する．

「」の付与で連想結果が特に改善されたモデルの 1 つが東北大 BERT であり，特に**色カテゴリ**で正答率が 0.13→0.75（8 題中 1 題→6 題）に上昇した．正解語が赤の課題では「」の付与により，上位 5 語が『黄色 **赤** 青 緑 ピンク』→『**赤** 青 緑 黄色 ピンク』となった．「」無でも色の単語は出力されているが，8 題中 7 題で 1 位に黄色が出力された．黄色が出力されやすい要因を分析するために，刺激語を含めない文『連想する色は MASK です。』を入力した．その結果，『黄色 水色 青 赤 カラー』の順で出力された．この結果から，『～連想する色は MASK です。』の MASK として黄色が出力されやすいと考えられる．黄色が出力されやすい要因をさらに調査するために，日本語 Wikipedia での出現頻度を調べたところ，最も高い色は赤であり（黄色は 6 位），黄色が出力されやすい要因は頻度ではなかった．そこで，Attention について調査した．図 1 に Attention を可視化した図を示す．「」無では MASK 周辺の単語である『色』『は』『です』への Attention が大きく，「」を付与することで，『は』『です』への Attention が小さくなった（図 1a→b）．これにより，黄色の出力が抑制され正解語の赤が 1 位に出力されるようになったと考えられる．他の色においても，「」の付与により，『は』『です』への Attention は小さくなった．一方で，図 1 のような刺激語への Attention の変化は殆ど無かった．

東北大 BERT では**国カテゴリ**も同様に「」の付与によって正答率が 0.25→1.00（4 題中 1 題→4 題）に上昇した．正解語がアメリカの課題では『世界中 以下 **アメリカ** 世界 日本』→『**アメリカ** 合衆国 ニューヨーク ホワイトハウス 世界』となっていた．「」無の不正解では，国名以外の単語が正解語よりも上位に出力された．色の分析と同様に刺激語を含めない文『連想する国は MASK です。』を入力すると『日本語 実名 以下 日本 不明』が出力された．このことから，『連想する国は』に対して色とは異なり，国名以外の単語も出力されやすいと言える．上記の文に「」を付与した連想文では『日本 アメリカ 中国 ロシア イタリア』となり，国名が出力されやすくなっていた．Attention を調査したところ，「」の付与によって刺激語への Attention が大きくなっていた．また，**色カテゴリ**と同様に「」の付与によって MASK の前後の単語『は』『です』への Attention が小さくなっていた．しかし，『国』への Attention に大きな変化は無く，国名が出力されやすくなった．国名が出力されやすくなった理由については，今後より詳細に分析していく．

a.「」無　　b.「」有

図 1　東北大 BERT での正解語が赤の課題における最終 Transformer 層の Attention．12 色は各々の Attention Head を表し，色の濃さは Attention の大きさを表している．[MASK]から各トークンに向かう線の色は Attention Head の色を混合したものである．[12]

次に，「」の付与で連想結果が改善された名詞のみ gMLP では，特に**使ってすることカテゴリ**で正答率が 0.00→1.00（2 題中 0 題→2 題）に上昇した．正解語が洗濯の課題では「」の付与により『無理 危険 不可 苦手 禁止』→『**洗濯** 安心 仕事 便利 安全』となった．

次に，名詞のみ BERT の結果について分析する．名詞のみ BERT では，「」の有無による正答率の差が東北大 BERT より小さく（0.34→0.44），特に**国カテゴリ**では正答率が変化しなかった（1.00→1.00）．アメリカの課題では「」の付与で『**アメリカ** メキシコ 日本 カナダ フランス』→『**アメリカ** 日本 カナダ メキシコ フランス』となり，「」の有無に関わらず，国名のみが出力されている．これに対し，東北大 BERT では「」の付与で『世界中 以下 アメリカ 世界 日本』→『**アメリカ** 合衆国 ニューヨーク ホワイトハウス 世界』となっており，この変化が名詞のみで学習することによる特徴であると考えられる．

なお，モデルや「」の有無に関わらず**スポーツカテゴリ**（野球，水泳，バレーボール，サッカー）では正答率が高かった．これは，スポーツに関連した名詞が日本語 Wikipedia 内で出現頻度が高いことに要因があると考えられる．

6 おわりに

本稿では，MLM として BERT と gMLP を用いて語群から語の連想実験を行った．連想課題として 5 つの刺激語から 1 つの連想語を回答するタスク（例: 富士山 浜名湖 茶畑 みかん うなぎ→静岡県）を使用し，MASK に連想語が入る連想文を作成した．実験では東北大 BERT に加え，MASK するトークンを名詞に限定して学習した BERT と gMLP を使用し，MASK に鍵括弧「」を付与した場合の検討も行った．実験の結果，「」を付与した東北大 BERT と名詞のみ gMLP で最も高い正答率 49%が得られた．また，上位 5 語以内に正解が含まれる割合は約 7 割であった．

本稿の結果から，MASK の前後に「」を付与する手法は，複数の刺激語から 1 つの連想語を出力する場合でも有効であることがわかった．さらに，BERT を名詞のみで学習させることで，正答率が向上することも確認できた（ただし，「」の付与によりその効果は失われる）．これらのことから，学習の段階で名詞・固有名詞に「」を付与することで，語連想タスクの精度向上が期待できる．また，Transformer の代替モデルである gMLP の精度も高かったことから，gMLP の精度向上の要因を分析するとともに，今後は gMLP を用いた自由連想タスクも検討する．

謝辞

本研究を進めるに当たり，乾・鈴木研究室の訓練済み日本語 BERT モデルをお借りしました．モデルを公開してくださったことに厚く御礼を申し上げ，感謝の意を表します．本研究は JSPS 科研費 JP20K11860, JP21K02052 の助成を受けたものです．

参考文献

1．Jacob Devlin，et al．*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*．s.l. : arXiv preprint arXiv:1810.04805，2018．
2．Hanxiao Liu，et al．Pay Attention to MLPs．s.l. : arXiv preprint arXiv: 2105.08050，2021．
3．米谷彩，渡部広一，河岡司．語の共起情報を考慮した感覚連想メカニズムに関する研究．情報処理学会 研究報告 2005-NL-166，2005．
4．白川真澄，中山浩太郎，原隆浩，西尾章治郎．複数語句から構成されるコンテキストを考慮した連想関係の抽出．DEIM Forum 2011 F3-1，2011．
5．芋野美紗子，吉村枝里子，土屋誠司，渡部広一．語概念連想を用いた複数単語からの連想語生成手法の提案．言語処理学会第 18 回年次大会(NLP2012)，2012．
6．相馬佑哉，堀内靖雄，黒岩眞吾．人間と BERT の語から語の連想の比較．言語処理学会第 27 回年次大会(NLP2021)，2021．
7．相馬佑哉，堀内靖雄，黒岩眞吾．検索画像を介在させた語から語の連想模擬法の検討．第 20 回情報科学技術フォーラム(FIT2021)，2021．
8．Yu Sun，et al．ERNIE: Enhanced Representation through Knowledge Integration．s.l. : arXiv preprint arXiv: 1904.09223，2019．
9．Hao Tian，et al．SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis. ACL2020，2020．
10．鈴木勉 宇野園子監修．CD 版そのまま使える失語症教材 2．エスコアール，2022 出版予定．
11．東北大学乾・鈴木研究室．Pretrained Japanese BERT models．(引用日: 2022 年 01 月 14 日.) https://github.com/cl-tohoku/bert-japanese.
12．Jesse Vig．A Multiscale Visualization of Attention in the Transformer Model．Florence，Italy : Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations"，2019．

A 付録

表 3　実験で使用するモデルの詳細

	東北大 BERT[11]	名詞のみ BERT	名詞のみ gMLP
dataset	日本語 Wikipedia（2020/08/31）	日本語 Wikipedia（2021/08/30）	
pre-training task	MLM + NSP	MLM	
batch size	256	256	128
learning step	100k	100k	200k
max seq length	512	128	128
hidden size	768	768	768
intermediate size	3072	3072	3072
num hidden layers	12	12	24
num attention heads	12	12	
parameter	1.1M	1.1M	1.1M
vocab size	32768	32768	32768
tokenizer	Mecab + WordPiece	Mecab + WordPiece	Mecab + WordPiece
learning time		17days（RTX8000×2）	23days（RTX8000×2）

表 4　連想課題の例

カテゴリ	正解語	連想文
色	赤	ポスト、りんご、消防車、トマト、いちごから連想する色は MASK です。
国	アメリカ	自由の女神、大統領、ワシントン、カルフォルニア、ミシシッピー川から連想する国は MASK です。
使ってすること	洗濯	洗濯機、洗濯ばさみ、洗剤、柔軟剤、ハンガーを使ってすることは MASK です。
スポーツ	バレーボール	アタック、セッター、ボール、ネット、レシーブから想像できるスポーツは MASK です。

表 5　東北大 BERT での色カテゴリにおける連想語上位 5 語

正解語	鍵括弧「」無	鍵括弧「」有
赤	黄色，**赤**，青，緑，ピンク	**赤**，青，緑，黄色，ピンク
青	黄色，**青**，赤，水色，虹	**青**，炎，赤，黒，緑
黄色	**黄色**，ピンク，赤，水色，紫	ピンク，赤，青，**黄色**，オレンジ
白	黄色，**白**，赤，不明，茶色	**白**，赤，青，黄色，黒
緑	黄色，**緑**，赤，紫，青	**緑**，赤，青，ピンク，紫
黒	黄色，赤，ピンク，紫，水色	**黒**，青，赤，白，藍
茶色	赤，黄色，緑，**茶色**，青	赤，緑，黒，青，**茶色**
ピンク	黄色，**ピンク**，赤，水色，紫	**ピンク**，赤，紫，緑，黄色