法政大学学術機関リポジトリ
HOSEI UNIVERSITY REPOSITORY

PDF issue: 2025-01-15

少数資源言語に対する大規模言語モデルを用いた解析支援

菅原, 太樹 / SUGAWARA, Taiki

(出版者 / Publisher)
法政大学大学院理工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学研究科編
(巻 / Volume)
65
(開始ページ / Start Page)
1
(終了ページ / End Page)
4
(発行年 / Year)
2024-03-24
(URL)
https://doi.org/10.15002/00030745

少数資源言語に対する大規模言語モデルを用いた解析支援

ANALYSIS SUPPORT USING A LARGE-SCALE LANGUAGE MODEL FOR A LOW RESOURCE LANGUAGES

菅原太樹
Taiki SUGAWARA
指導教員　藤井章博

法政大学大学院理工学研究科応用情報工学専攻修士課程

Linguistic models have not been trained for languages with small corpora, which are called low resource languages. In this study, we pre-trained a language model using BERT for Classical Japanese, one of the low resource languages. This study verified the effectiveness of pre-training of language models for languages with small corpora.

Key Words *: NLP, Japanese, BERT*

１．序論

近年，自然言語処理の分野において，大規模なコーパスを用いた言語モデルの学習が盛んに行われている．BERT や GPT に代表される言語モデルは，大規模なコーパスを用いて学習されたモデルであり，様々な自然言語処理のタスクにおいて，高い精度を示している．

しかし，これらの言語モデルは，英語や中国語などの大規模なコーパスが存在する言語に対してのみ，行われており，少数資源言語と呼称される，コーパスが少ない言語に対しては，言語モデルの学習が行われていない．

本研究では，少数資源言語の一つである古典日本語に対して，BERT による言語モデルの事前学習を行い，コーパスの少ない言語に対する言語モデルの事前学習の有効性を検証した．

古典日本語は現在使用されておらず，既に発展的消滅をした言語であるが，古典籍に記されている情報については国語学の分野で研究が行われている．

また，本研究をもとにして他の少数資源言語の自然言語処理の研究につながることが期待される．

２．研究背景

（１）コーパスに依存する自然言語処理

世界には 7,168 もの言語が存在する．しかし，昨今急速に発展を遂げる自然言語処理の対象となるのは，そのうちのごく少数の言語のみである．

P. Joshi ら(2020)[1]は，ラベル付けされたリソース数と，ラベル付けされていないリソース数の二つの特徴軸を用い，自然言語処理から見た言語の分布を調査した．

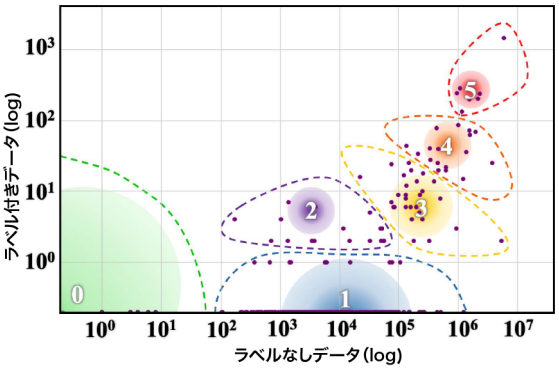

図１　言語資源の分布図
P. Joshi ら(2020)をもとに，著者翻訳

図１では，資源の量による言語の分布を示している．

調査対象の言語は 2485 個である．縦軸が LDC(Linguistic Data Consortium)が扱うアノテーションデータの LDC コーパスと，ELRA(the ELRA Language Resources Association)の提供する ELRA マップに基づいたラベル付きデータ数である．横軸は Wikipedia をもとにしたラベルなしデータ数である．

本研究では，5 番以外の言語を少数資源言語と呼称する．

また，1-4 番の言語を少数資源言語の中でも研究意義の高い言語と位置付ける．

（２）消滅危機言語

国語学の分野では，言語話者の減少を危惧し研究が進

められている．UNESCO(2010)は，消滅の危機に瀕する言語をリスト化している[2].日本語や英語などの「安全な言語」から、「消滅した言語」まで 6 種の分類がされており、そのうち安全な言語を除いた 5 種類がリストに掲載されている。

言語が消滅し，理解できる人がいなくなると，その言語で記された文献や言い伝えを人類は理解できなくなる.これはその言語の文化や歴史を人類が失うことを意味する.特に消滅危機言語で残る文化や歴史は，他の言語では

残されていないものが多い.そのため，消滅危機言語の保存・継承は重要であり，国語学分野ではもとより，前章で述べた自然言語処理をはじめとする分野でも積極的に研究をしていかなければならない.

（３）日本語の歴史

ここで，日本語について考える.現代日本において標準語と言われる現代日本語は，消滅危機言語ではなく，少数資源言語のグループ分けでは 5 番に分類される言語である.

では，現代日本語以外の日本語はどうだろうか.加藤ら(1991) の分類によると、表 1 のような分類になる[3]。

表１　日本語の歴史分類

名称	年代	時代
古代日本語	B.C. 13000 - A.D. 600	縄文・弥生・古墳時代
上代日本語	600 - 784	飛鳥・奈良時代　約200年間
中古日本語	784 - 1184	平安・院政時代　約400年間
中世日本語	1184 - 1603	鎌倉・室町時代　約400年間
近世日本語	1603 - 1867	江戸時代　約300年間
近代日本語	1868 - 1945	明治・大正・昭和前半時代　約80年間
現代日本語	1946 -	昭和後半 -

近世日本語以前の日本語は，わずか 150 年前の文献であるにもかかわらず，多くの日本語話者が理解することができない. 本論文では，上代日本語以降，近世日本語以前の日本語を古典日本語と呼称し，研究対象とする.

古典日本語は，のちに近代日本語と現代日本語に発展した言語であり，2.2 章で述べた消滅危機言語には含まれない.しかし，自然言語処理の分野で解析が進んでいないのは事実であり，2.1 章で述べた，コーパスが少ない言語である.古典日本語に対する自然言語処理の研究を進めることで，発展的に消滅危機言語や他の少数資源言語に対する研究につながることが期待される.

３．関連研究

（１）人力による翻刻

古典籍の文字を，テキスト化することを翻刻という.現時点で完全な OCR 技術は存在しないため，多くは人力によって行われている.また，OCR を行うためにはアノテーションデータが必要であり，翻刻作業は情報科学分野においても重要な課題である.

作業主体としては、京都大学古地震研究会が，2017 年から「みんなで翻刻」というシチズンサイエンスプロジェクトを行っている．また、人間文化研究機構国文学研究資料館をはじめとする国内の学術機関は，文科省の大規模学術フロンティア促進事業の一環として，「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」を進めている.

（２）古典籍の OCR

くずし字の OCR モデルとしては，人文学オープンデータ共同利用センターの KuroNet や，その後継の RURI(瑠璃)がある.KuroNet は，最大で 0.8984 の F1 スコアを出した，End-to-End のくずし字 OCR モデルである [15].また，TOPPAN 株式会社はくずし字資料の解読や公開をサポートするサービスとして，「ふみのは」を提供しており，90% 以上の精度を謳っている[4]

しかし，OCR をされてもテキスト化されたデータが多く公開されていない.現在，人文学オープンデータ共同利用センターが公開している日本古典籍データセットの中で，本文を公開しているのは 29 点しかない.画像のみ公開している古典籍が 3,126 点あるため，これは非常に少ないことがわかる.

（３）現代日本語以外の自然言語処理

表 1 に示したように，本研究で研究対象となる古典日本語と現代日本語の間には，近代日本語と呼ばれる時期が存在する.謝ら (2023) は，近代日本語に対して BERT モデルを作成し，近代日本語の誤り訂正を行なった[5].

謝らは，BERT モデルを使用することによって誤り訂正タスクを行い，OCR の精度向上を図った.近代文 BERT は，2022 年 7 月 30 日時点の青空文庫のデータと，国立国語研究所の近代語コーパスによって学習された.

その後，謝らは近代文の誤り訂正のためのデータセットを作成した.近代文データの一部の文字を類似文字に書き換え，誤りのあるデータセットを作成して検証に用いた.検証に当たっては，東北大 BERT を比較に用いて実験を行っている.結果は以下の通りである.

表２　近代文 BERT の誤り訂正の正解率

近代文BERT	近代文コーパス	5	0.39
近代文BERT	近代文コーパス	10	0.54
近代文BERT	近代文コーパス	15	0.56

表３　近代文 BERT と東北大 BERT の比較実験

モデル	コーパス	Train	Test	Epoch	Acc
近代BERT	近代文コーパス	165087	11899	15	0.56
東北大BERT	近代文コーパス	85673	4143	15	0.2
近代BERT	現代文コーパス	104199	2252	10	0.1
東北大BERT	現代文コーパス	143027	2252	5	0.76

表３に示されるように，近代文 BERT は，近代文の誤り訂正において， 東北大 BERT を上回る性能を示した.注目したいのは，近代文 BERT と近代文コーパスの学習

結果の正解率 0.56 と，東北大 BERT と現代文コーパスの正解率 0.76 である.近代文 BERT は東北大 BERT に比べてデータ量が 0.11% であるのにも関わらず，性能の開きが大きくない.これは，少量のデータセットでも BERT モデルを学習する意義がある可能性を示している.

４．要素技術

（１）BERT

BERT は，あらかじめ大規模なコーパスを用いて事前学習を行い，事前学習済みモデルを用いて特定のタスクに適した追加学習を行う.この追加学習をファインチューニング（Fine-tuning）と呼び，ファインチューニングによって特定のタスクに適したモデルを作成することができる.事前学習には大量のデータセットが必要となり，各言語では Wikipedia などの大規模なデータセットを用いて事前学習が行われている[6].

（２）形態素解析器と辞書

先述した BERT を含め，日本語の自然言語処理においては，形態素解析器が必要となる.日本語は，分かち書きと呼ばれる単語ごとに空白で区切る構造をしておらず，形態素解析器を用いて単語ごとに切り分ける.形態素解析器は，形態素解析エンジンと辞書によってなる.

UniDic は，古文用の辞書が存在する反面，その分割が細かく，使用する辞書を選択することが難しい．本研究で対象とする古典日本語は，上代日本語から近世日本語を含むものであることからも，なおいっそう UniDic のように 細かい分類が行われている辞書を用いることは適切でない.しかし，古典日本語全般に特化した辞書は現時点で他に存在しない.

（３）Sentencepiece

Sentencepiece は，Google が開発したトークン化ライブラリである．Sentencepiece は，高頻度語は従来通り 1 単語として扱う一方，低頻度語はより短く数文字単位で分割する.これにより，未知語がなくなり，あらかじめ決められた語彙サイズに分割することができる[7].

５．研究手法

（１）データセット

本研究に使用するデータは，テキストが存在する表４の 17 点の古典籍である．日本古典籍データセットから与えられたテキストは，プレーンテキストまたは DOCX 形式で配布されている．データには脚注や凡例，古典籍のメタデータなどが無作為に含まれているため，学習にあたってそれらを削除，空行や特殊文字を成形する．

表４に示される古典籍は，1633 年から 1852 年に書かれたものである．よって，これらの古典籍は表 1 に示す近世日本語にあたる.しかし，先述したように古典籍は明確な区分があるわけではなく，語彙や文法において別の時代区分のものが含まれている可能性がある.

表４　本研究で使用する古典籍

書名（統一書名）	（西暦）	冊数
二人比丘尼	1633	1冊
源氏物語	1654	54冊
曾我物語	1671	12冊
日本永代蔵	1688	6冊
武家義理物語	1688	6冊
世間胸算用／大晦日ハ一日千金	1699	5冊
御前菓子秘伝抄	1761	1冊
料理珍味集	1764	4冊
傾城買四十八手	1790	1冊
餅菓子即席／手製集	1805	1冊
飯百珍伝	1833	1冊
虚南留別志	1834	1冊
歌学提要	1850	1冊
鼎左秘録	1852	1冊
椿説弓張月	1807～1811	10冊
浮世風呂	1809～1813	9冊
南総里見八犬伝	1814～1842	101冊

（２）BERT

本研究には，BERT を用いて事前学習モデルを作成することによって，実験を行う.4.1 章で述べたように，BERT は大量のデータセットを用いて事前学習を行い，その後に特定のタスクに適した追加学習を行うことができる．また，双方向での学習を行っている.これらの特性は，古典日本語をはじめとする低資源言語の解析において有効であると考えられる.

BERT は，トークンを入力とすることから形態素解析器または Sentencepiece によってトークン化を行う必要がある．形態素解析器では，古典日本語全般に適用可能な辞書が存在しない.また，古典籍の語彙を全て網羅していない可能性がある.一方，Sentencepiece は，入力文からトークンを生成するため，理論上未知語がなくなる.よって，本研究では，Sentencepiece を用いてトークン化を行う．

実装には，rinna 社がオープンソースで公開している BERT の実装[8]を元に，古典日本語の学習に調整したモデルを作成する.

６．研究手法

いくつかのモデルを作成し，結果は下記の表のようになった.

表５　古典日本語 BERT モデルの誤り訂正の loss

Batch Size	epoch	loss
2	5	9.9192
2	8	9.5994
8	6	10.414

また，loss の推移は以下の図のようになった.

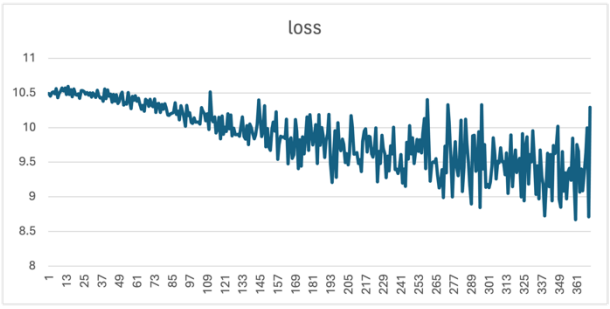

図２　学習時の loss の推移

７．考察

表 10 にある通り，古典日本語 BERT モデルは，loss が若干大きく，モデルとしての性能が低い.また，図 4 に示されるように，loss の推移も安定していない.しかし，loss が下がっていることもわかり，一定の性能が満たされていると考えられる.

この結果は，データ量の少なさによるものと考えられる.日本古典籍データセットは，画像ファイルは多いがテキストデータが少ない.本研究によって，大規模データを使用したモデルにはテキストデータが不足していることが明らかになった.しかし，同様に過小のデータ量で BERT モデルを学習した近代文 BERT は，東北大 BERT を上回る性能を示した.近代文 BERT はモデルが公開されていないことから，検証が不可能であるが，古典日本語 BERT も同様の性能を持つ実装がある可能性があることを示している.

本モデルは，今後データセットが拡充されるにつれて，ファインチューニングを行うことができる.今後の追加学習によって，少ないコストで性能を向上させることができるため，本モデルは有用であると考えられる.

８．考察

本研究では，古典日本語 BERT を作成し，その精度を検証した.古典日本語は現在使われていない言語であり，この結果は少数資源言語や消滅機器言語の自然言語処理においても有用である.本研究のモデルによって，古典日本語の翻刻作業が効率化されるため，結果的に古典日本語のテキストデータが増加することが期待される.これは，他の少数資源言語でも同様であり，現在テキスト化や翻訳が行われていない言語でも，同様の手法を用いることができる.英語や現代日本語のように，完全なモデルを作成して全てを解決することは難しいが，一定の性能が得られるモデルを使用して，人間のアシストをすることでその言語の研究が進むことが期待される.

謝辞：

本研究のみならず，大学院での研究生活全般において，多くの方々にご支援いただきました.

特に大学院より受け入れていただき，研究活動をご指導いただいた法政大学大学院理工学研究科 藤井章博先生に，深く感謝いたします.

参考文献

1) P. Joshi, S. Santy, A. Budhiraja, K. Bali と M. Choudhury, 「The State and Fate of Linguistic Diversity and Inclusion in the NLP World」, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter と J. Tetreault, 編, Online: Association for Computational Linguistics, 7 月 2020, pp. 6282–6293. doi: 10.18653/v1/2020.acl-main.560.

2) UNESCO, *Atlas of the world's languages in danger*, 3rd 版. Memory of peoples, no. 23. France: UNESCO, 2010. [Online]. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000187026

3) 加藤 彰彦, 佐治 圭三, 森田 良行と椎名 和男, 日本語概説, 初版 23 刷. おうふう, 1991. [Online]. Available at: https://cir.nii.ac.jp/crid/1130000795417501568

4) A. Lamb, T. Clanuwat と A. Kitamoto, 「KuroNet: Regularized Residual U-Nets for End-to-End Kuzushiji Character Recognition」, *SN Computer Science*, vol. 1, no. 3, p. 177, 5 月 2020, doi: 10.1007/s42979-020-00186-z.

5) 謝素春,「日本語 BERT モデルによる近代文のテキスト化の精度向上」, 人間情報学研究科年誌, no. 27, pp. 1–5, 6 月 2022.

6) J. Devlin, M.-W. Chang, K. Lee と K. Toutanova, 「BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding」. arXiv, 2019 年 5 月 24 日. doi: 10.48550/arXiv.1810.04805.

7) 「google/sentencepiece」. Google, 2017 年. 参照: 2024 年 2 月 12 日. [Online]. Available at: https://github.com/google/sentencepiece

8) 「japanese-pretrained-models」. rinna Co., Ltd., 2023 年 12 月 5 日. 参照: 2023 年 12 月 7 日. [Online]. Available at: https://github.com/rinnakk/japanese-pretrained-models