Transformerによる中国古典詩歌の解釈生成ネットワークの構築

王　藝臻　　荻野　正樹
関西大学大学院　知識情報研究科
{k435374,ogino}@kansai-u.ac.jp

概要

中国古代詩は素晴らしい作品が数多く残されているにも関わらず，言語的な難しさから現代の中国人でも解釈が困難な場合がある．古典的な詩を現代語の解釈として翻訳することができれば，現代の詩人の創作活動の支援としても役立つことが期待できる．このため，本研究では，古代中国の詩からその解釈を生成する言語処理モデルを提案する．提案モデルは Transformer を使った生成モデルとなっており，詩歌を入力とし，その解釈を現代語として出力するように学習する．提案モデルは，Attention is all you need のモデルを使用し，生成された解釈文は BLEU スコアで評価した．トークンとして，文字（漢字）を使用する手法と，単語を使用する手法について比較を行い，後者の方が BLEU スコアの評価が高い結果が得られた．また，長文の解釈文について，要約を用いて文字数を少なくして学習させたが，BLEU スコアとしては要約しない場合と同程度の結果が得られた．

1 はじめに

近年， Generative AI と呼ばれる分野の技術の革新が進み，AI を使った画像生成，文章生成，音楽生成等の研究が盛んに行われており，画像からの文章生成，キーワードからの画像生成などの応用研究やサービスの提供も行われている［1］.

AI を使ったアート生成の試みとして，中国古典詩の生成についても多くの研究が行われている．Ryu らは，事前学習済みモデルに基づく中国古典詩の生成の研究を提案した［2］．彼らは主に BART などの事前学習モデルを利用した FS2TEXT と RR2TEXT を提案したが，これらのモデルは特定のスタイルの詩文を生成することができる．このモデルの性能を検証するため，詩人や詩作研究者のグループに，中国の古典詩と BART で生成した詩についての判別をしてもらい，AI 詩についてのチューリングテストを行った． 600 名以上の参加者についての大規模な調査では，ハイレベルな詩のファンたちでも，AI の作品と人間の作品を区別することができないことが明らかとなった．

詩歌の創作活動は現代の中国でも人気があり，現代中国の詩人の数は 500 万人に達している．ところが，中国には古くからの優れた詩歌の作品が残されているにも関わらず，古典的な素養に関する学習の機会の不足から，古典的な詩歌の作品の解釈が難しく感じる人も多く，古典的作品が創作に有効に活用されているとは言えない．古典的詩歌には現代の人からは理解しにくい表現や文法があり，それらの解釈を支援するサービスがあれば，古典的な作品から創作のインスピレーションを得る機会も増えることが期待される．そこで，本研究では，中国の古典的詩歌について，その意味的な解釈を生成するシステムを提案する．提案するモデルでは Transformer［3］を使い，encoder 側に詩歌を入力し，decoder 側で詩歌を説明する文章を出力するように学習させる．学習後は古典詩歌を入力することで，現代人にわかりやすい解釈文が生成される．

2　提案モデル

ディープラーニングの出現により機械翻訳においてもニューラルネットワークの有用性が着目されてきた．Seq2Seq モデルは 2014 年に Google によって提案された手法で，エンコーダ，デコーダに RNN を用いることにより，エンコーダでは入力された時系列情報から文脈情報を抽出し，デコーダでは文脈情報から別の時系列情報を生成する．このモデルでは学習におけるコストや精度の高い依存関係モデルを構築が難しいという問題があった．2017 年に発表された Transformer は，RNN[4]を使わずに Attention 層だけを使ってニューラルネットワークを構築することで，これらの問題を解決しており，

BERT，GTP をはじめ，現在非常に多くの自然言語処理モデルの基盤技術となっている．本研究では，Transformer を翻訳機として使い，古典的詩歌から現代語の解釈文へ翻訳するように学習するモデルを提案する．

通常，自然言語処理モデルの処理では入力されるトークンの長さの制限のため，長いテキストデータは学習データとして使用することができない．しかし，特にデータ量の少ないトレーニングセットでは，削除せずにデータを活用したほうがより深く学ぶことができると考えられる．

そこで本研究では，入力文の長さが最大入力トークン数を超えた場合は，要約によって短文に変換して入力する方法を提案する．（図 1）まず，長文テキストを抽出する． 次に，BERT ベースの要約生成器を用いて，古文の要約と古文解釈の要約を生成する．最後に，生成された対になる古文書要約と古文書解釈要約を元データと混合し，シャッフルする． その結果，より大きく，より使いやすいデータセットが出来上がる．

本研究は Miller の提案する文章要約モデル[5]をベースに中国語要約プログラムを作成した. Pre-trained Language モデルとして Yang らが公開している Whole Word Masking モデルを使用している [6].

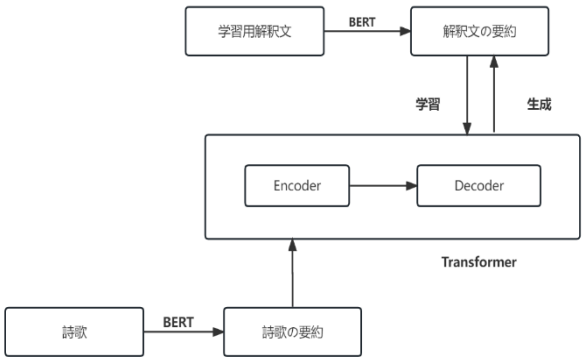

図 1 提案モデル

3 実験

本研究では，実験結果に対して，要約モデルを使用した場合と使用しない場合のデータを比較した．そして，実験テキストがすべて中国語であるのを考量し，NLTK (Natural Language Toolkit）を使って漢字単位で分割使用した場合と，中国語専用の形態素解析ツール Jieba を使用して単語を分離した場合の実験結果に対する影響を比較した．

3.1 実験条件

本研究で用いた計算機環境は，ハードウェア環境として CPU は AMD R7 3800X， GPU は RTX2080 を使用し，ソフトウェア環境は OS として Ubuntu 20.04 ，深層強化学習ライブラリとして Pytorch 1.10.1 を使用している．

3.2 データ説明

古詩文網[7] には中国の古典詩についてタイトル，詩文，解釈が公開されており，本研究ではこのページから学習データとして古詩と古文を計 7207 件抽出し，繁体字と簡体字に変換した． 最後に，それぞれの行をスペースで区切り詩とその解釈をペアにした．表 1 に学習データの一例を示す．

表 1 データの例

詩文	解釈文
寥落古行宫，宫花寂寞红．白头宫女在，闲坐说玄宗	曾经富丽堂皇的古行宫已是一片荒凉冷落，宫中艳丽的花儿在寂寞寥落中开放．幸存的几个满头白发的宫女，闲坐无事只能谈论着玄宗轶事．

3.3 モデルのパラメータ

本研究では，［Attention is all your need］[3]の論文から基本モデル Base と Small のパラメータを参考にした．モデルのパラメータを表 2 に示す．なお，GPU のメモリ量の制限により，big モデルのパラメータは使用していない．

表 2 パラメータ

	N	dm	d_{ff}	h	p	ls	ts
Base	6	512	2048	8	0.1	0.1	2000
Small	2	256	1024	8	0.1	0.1	2000

3.4 形態素解析

本研究では，形態素解析の分割の粒度による学習への影響を調べた．分割には Natural Language Toolkit(NLTK) [8] と Jieba[9] を使った．Natural Language Toolkit(NLTK) [8]は，NLP の分野で最もよく使われている Python ライブラリの 1 つである．このライブラリのユーティリティ word_tokenize を使って，テキストの切り出し処理を行った．Jieba は中国語単語分割によく使われている中国語テキスト分割モジュールである．Jieba には分割の粒度について，いくつかモードが設けられている．表 3 に 3 つのモードでの監視の分割の違いを示す．本研究ではこれらのモードの中で cut_all=True を使用した．

表 3　使用した後の漢詩の分け方

モード	床前明月光，疑是地上霜.
lcu 文章分析に適した，最も精密に文章を切り出そうとする「精密」モード	’床前’，’明月光’，’，’，’疑是’，’地上’，’霜’，’．’
cut_all=True 文中の単語になりうるものをすべてスキャンするため，非常に高速だが，曖昧さの解消にはならない	’床’，’前’，’明月’，’明月光’，’月光’，’，’，’疑’，’是’，’地上’，’霜’，’．’
lcut_for_search 正確なパターンに基づいて，長い単語を再びスライスして，リコールを向上させ，検索エンジンの単語分割に適している.	’床前’，’明月’，’月光’，’明月光’，’，’，’疑是’，’地上’，’霜’，’．’

4 結果と考察

4.1 モデルパラメータによる比較

2 つの異なるパラメータを用いて得られた BLEU スコアを表 4 に示す.

表 4　モデルによる BLEU スコア比較

	BLEU-1	BLEU- 2	BLEU- 3
Base	32.7	16.1	9.0
Small	29.8	15.8	10.9

2 つの結果を比較すると，Base を使用した方が BLEU スコアが高いことがわかる.

4.2 分割単位による比較

入力する文の分割に Jieba を使って単語で分割した場合の BLEU スコアと，NLTK を使って個別の漢字で分割した場合の BLEU スコアを表 5 に示す.

表 5　分割単位による BLEU スコア

分割単位	BLEU-1	BLEU- 2	BLEU- 3
漢字	32.7	16.1	9.0
単語	36.5	27.1	24.0

表 4 に示すように，Jieba を使用することによって，BLEU スコアが大幅に向上することが確認できた.

4.3 提案モデルの使用

解釈文として要約文を使用した場合と使用していない場合の学習曲線を図 2 に示す．両者のモデルとも 2000 エポックで loss が収束した.

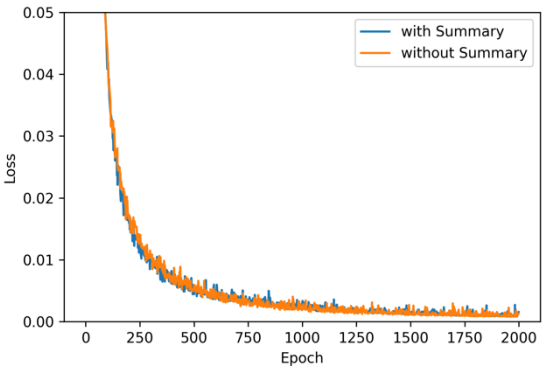

図 2 解釈文出力モデルの学習曲線

学習済みモデルを用いて詩の解釈を生成した例を表 6 に示す．学習データに要約文を使用した場合と使用していない場合もこの解釈が生成されている.

要約文不使用と使用の場合の BLEU スコアを表 7 に示す．表の BLEU スコアから，要約文を使用したデータセットの BLEU スコアは，使用しなかった場合のスコアから大きく変化していないことがわかる.

これは，データ量を同じにした結果である． したがって，長文を要約処理したデータを本来のデータに入れても悪影響を与えないと結論づけられる．しかし同時に，この解釈で生成された結果を評価するために BLEU スコアを使用した．

表 6 詩の解釈を生成した例

詩	造物无言却有情 ，每于寒尽觉春生。千红万紫安排著，只待新雷第一声。
人による解釈	大自然虽然默默无言，但却有情，寒尽而带来春天，悄悄地安排好万紫千红的百花含苞待放大自然早已安排好了万紫千红，只等春雷一响，百花就将竞相开放。
要約文使わないモデルの解釈	大自然虽然默默无言但却有情，每当寒冬将尽便促实在欣赏西山的大自然。
提案モデルによる解釈	大自然虽然默默无言但却有情，每当寒冬将尽便促使春意萌生。大自然早已安排好了万紫千红，只等春雷一响，百花就将竞相开放 。

表 7 BLEU スコア

	BLEU-1	BLEU- 2	BLEU- 3
不使用	38.6	26.5	22.2
使用	38.3	26.8	22.6

通常，翻訳を評価する指標として BLEU が使われるが，このように同じ言語間の解釈生成の時もこの指標を使うことに問題がないかどうかは，今後検討解決されるべき問題である．

まだ，複数の解釈文を読むと，ある言葉の意味は明確に説明されているものの，文章全体がスムーズに読めないことがわかる． これは，通常の日常的な使用に悪影響を及ぼすと考えられる．

さらに，現在では，BERT で事前に学習したモデルを用いて小規模なデータセットで再学習することが一般的になっており，少ない計算資源でより優れた詩の解釈生成モデルを迅速に学習できる可能性がある．今後，BERT の学習済みモデルで再学習を行うことが今後の課題になる．

6 おわりに

本研究は詩歌解釈生成サービスを構築した．古代漢詩や文言文からその解釈を生成することに成功した．

そして，要約生成器を用いて作成した長文要約をトレーニングセットとテストセットに入れ，その結果を比較した．BLEU スコアに大きな変化はないことがわかった． その結果，長文の要約を使用しても学習結果に影響を与えないという結論に達した．次に，形態素解析に Jieba を使用した場合の BLEU の結果への影響を比較したところ，Jieba を使用することによって，BLEU スコアが大幅に上がることが確認できた．

しかし，解釈の評価する指標，文章流暢さの問題など課題が残っている．これらの解決は今後の課題になる．

参考文献

1. Liu, Yusen and Liu, Dayiheng and Lv, Jiancheng and Sang, Yongsheng. Generating Chinese Poetry from Images via Concrete and Abstract Information arXiv preprint arXiv: 2003.10773, 2020.
2. Wang Z, Guan L, Liu G. Generation of Chinese classical poetry based on pre-trained model[J]. arXiv preprint arXiv:2211.02541, 2022.
3. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Kaiser, and I. Polosukhin. Advances in Neural Information Processing Systems , page 5998--6008. (2017)
4. Yi, Xiaoyuan and Li, Ruoyu and Sun, Maosong Generating Chinese Classical Poems with RNN Encoder-Decoder arXiv preprint arXiv: 1604.01537, 2017.
5. Miller D. Leveraging BERT for extractive text summarization on lectures[J]. arXiv preprint arXiv:1906.04165, 2019.
6. Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for chinese bert[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.
7. 古诗文网 https://www.gushiwen.cn
8. Bird, Steven, Edward Loper and Ewan Klein (2009). Natural Language Processing with Python. O'Reilly Media Inc.
9. jieba: https://github.com/fxsjy/jieba