情報処理学会論文誌 Vol.63 No.2 269–282 (Feb. 2022)

事前学習モデルを用いた近代文語文のニューラル機械翻訳

喜友名 朝視顕[1,a)] 平澤 寅庄[1,b)] 小町 守[1,c)] 小木曽 智信[2,d)]

受付日 2021年5月18日, 採録日 2021年11月2日

概要：明治・大正期に広く用いられた近代文語文は，現代の日本語話者にとって，専門知識がないと読むことが難しい．近代文語文の特徴として，現代の書き言葉と共通する単語は多いが，共通する n の大きな n グラムはほとんどないことがあげられる．本研究では，『学問のすゝめ』（1872–1876）と『人世三宝説』（1875）の翻訳に焦点を当てる．ニューラル翻訳モデルの学習に使用できる対訳コーパスが少ないという問題に対応するため，事前学習モデルを用いる．実験の結果，対訳コーパスを用いず単言語コーパスのみを用いることで，原文との類似度が高い出力が得られた．加えて，既存の自動評価指標とその変種がどの程度ポストエディットのコストを考慮できているかを調査した．

キーワード：近代文語文，機械翻訳，事前学習モデル

Neural Machine Translation of Classical Japanese Texts in the Late 19th Century Using Pretrained Language Models

Tomoshige Kiyuna[1,a)] Tosho Hirasawa[1,b)] Mamoru Komachi[1,c)] Toshinobu Ogiso[2,d)]

Received: May 18, 2021, Accepted: November 2, 2021

Abstract: Classical Japanese texts in the late 19th century are difficult for contemporary Japanese to read without specialized knowledge. Specifically, this paper focuses on the translation of “An Encouragement of Learning” (1872–1876) and “Three Treasures in Human Life,” (1875) which have many identical unigrams between the source and reference sentences but no significant overlapping larger n-grams. We approach this task by using pretrained language models to address the associated data acquisition bottleneck. The results show that the use of an unsupervised method without fine-tuning on parallel data provides translation outputs with a high degree of similarity to the source text. In addition, we investigate the extent to which existing automatic evaluation metrics and their variants are able to account for post-editing cost.

Keywords: classical Japanese in the late 19th century, machine translation, pretrained language models

1. はじめに

明治・大正期に広く用いられた近代文語文は，現代の日本語話者にとって，専門知識がないと読むことが難しい．語彙の多くは今日まで連続して用いられているものの，文法面で大きく異なっているためである．しかし，新聞や雑誌などのマスコミが発達したこの時期の資料は大量に残されていることから，近代文語文で書かれた情報にアクセスするための機械翻訳技術は，必ずしも歴史的資料を専門とはしない人々から，人文・社会科学などの多くの分野で求められている．本研究では，近代の文語文として，福澤諭吉の『学問のすゝめ』（1872–1876）と西周の『人世三宝説』（1875）の現代語訳に焦点を当てる．近代文語文とその現代語訳には共通する単語は多いが，接続詞や助動詞が異なっているため共通する n の大きな n グラムはほとんどない．**表 1** に『学問のすゝめ』および『人世三宝説』の 2 文とその現代語訳を示す．内容的には，前者が大衆向けに比較的平易に書かれた啓蒙書であるのに対し，後者が知識

[1] 東京都立大学
Tokyo Metropolitan University, Hino, Tokyo 191–0065, Japan
[2] 国立国語研究所
National Institute for Japanese Language and Linguistics, Tachikawa, Tokyo 190–8561, Japan
[a)] kiyuna-tomoshige@ed.tmu.ac.jp
[b)] hirasawa-tosho@ed.tmu.ac.jp
[c)] komachi@tmu.ac.jp
[d)] togiso@ninjal.ac.jp

表 1 『学問のすゝめ』および『人世三宝説』の 1 文とその現代語訳
Table 1 Example sentences from “An Encouragement of Learning” and “Three Treasures in Human Life” and their contemporary translations.

『学問のすゝめ』	
原文	されば賢人と愚人との別は学ぶと学ばざるとによりてできるものなり。
現代語訳	つまり、賢い人と愚かな人との違いは、学ぶか学ばないかによってできるものなのだ。
原文	されば今より後は日本国中の人民に、生まれながらその身につきたる位などと申すはまずなき姿にて、ただその人の才徳とその居処とによりて位もあるものなり。
現代語訳	これからは、日本中ひとりひとりに生まれつきの身分などといったものはない。ただその人の才徳や人間性や社会的役割によって、その位というものが決まるのだ。
『人世三宝説』	
原文	而て苟も此大本に反すれば天殃の降る踵を回らすに暇あらず。
現代語訳	逆に言えば、もしもこの大本に反するならば、即座に禍が降りかかってくることになる。
原文	白刃を蹈み砒霜を飲むが如き人亦孰か甘んじて之を侵さむ。
現代語訳	白刃を踏めば足が切れ、ヒ素を飲めば死ぬという法則は誰でも知っているので、誰もそんな馬鹿なことを試みようとはしない。

層を対象とした論説文であるという違いがある．

古文の翻訳は，翻訳モデルの学習に使用できる対訳コーパスの量が限られているため，高品質な機械翻訳モデルを訓練することは難しい課題の 1 つである．この問題を低減する方法の 1 つに，事前学習モデルの利用がある．Conneau ら [1] は，言語横断的な言語モデル cross-lingual language model（XLM）を用いて，Transformer [12] ベースのニューラル機械翻訳（neural machine translation: NMT）モデルの事前学習を行い，教師あり NMT と教師なし NMT で実験を行い，両者ともに翻訳性能が向上することを示した．XLM の学習方法には，各単言語コーパスを用いた教師なし手法（masked language modeling: MLM）がある．

本研究でも，単言語コーパスを用いて事前学習した XLM を用いる．言語横断的な学習が，2 言語間に共通する単語の学習に効果的に働くことが期待できる．具体的には，近代文語文と現代文の各単言語コーパスを用いて，MLM により XLM を学習する．そして，事前学習済みの言語モデルを用いて初期化をした翻訳モデルに対し，教師なし手法の追加学習と教師あり手法の fine-tuning の両方，またはいずれか一方を行う．

加えて，既存の自動評価指標とその変種がどの程度ポストエディットのコストを考慮できているかを調査する．一般に，NMT モデルの出力は流暢であるが，誤りが多く，ポストエディットが必要である．専門知識を持つポストエディタを利用できる場合，NMT モデルは原文の単語をそのまま出力することが許される．このような NMT モデルの出力をポストエディットの観点で評価する際には，参照訳だけでなく原文も考慮すべきかもしれない．その場合，機械翻訳の分野で一般に用いられている自動評価指標である，bilingual evaluation understudy（BLEU）[7] は普通，参照訳のみを考慮するため，原文も考慮するには，原文に対する BLEU および参照訳に対する BLEU の平均といった変種を考える必要があるだろう．本研究では，BLEU および編集距離とその変種がどの程度ポストエディットのコストを考慮できているかを調査するために，翻訳システムの出力に対しポストエディットの観点で人手評価を行い，各自動評価指標の評価値との相関を調べた．

本研究における貢献は，次の 2 点である．

(1) 近代文語文から現代文への機械翻訳タスクに事前学習モデルを用いて，原文との類似度が高い出力を得る．この出力は，人手評価により，ポストエディットの観点で原文よりも良いと判断された．
(2) 原言語と目的言語に共通する単語は多いが，共通する n の大きな n グラムはほとんどない場合の，ポストエディットのコストを考慮する自動評価指標として，既存の自動評価指標とその変種がどの程度適しているかを調査した．

2. 関連研究

近代より昔の日本語を現代語に翻訳する研究がいくつかある．星野ら [18] は，古代から近世までの古文とその現代語訳からなる段落単位の対訳コーパスから文単位の対訳コーパスを抽出する手法を提案し，抽出した対訳コーパスを用いて統計的機械翻訳を行った．Takaku ら [11] は統計的機械翻訳の代わりに，通時適応した事前学習済み単語分散表現を用いて，教師あり NMT を行った．

最近では，近代文語文の翻訳に焦点を当てた研究もいくつかある．稲見ら [13] は，自ら逐語訳したデータセットを使って 3 つの NMT モデルの学習を行い，畳み込みニューラルネットワーク [3] が再帰ニューラルネットワーク [10] と Transformer [12] よりも翻訳性能が良いことを報告した．竹内ら [14] は，Transformer ベースのモデルに対し，対訳

表 2 訓練・開発データに含まれる n グラムの数と重複率
Table 2 Number of all n-grams and ratio of overlapping n-grams in the training and development data.

	1 グラム		2 グラム		3 グラム		4 グラム	
	トークン	タイプ	トークン	タイプ	トークン	タイプ	トークン	タイプ
『学問のすゝめ』の訓練・開発データ								
原文	55,687	5,355	54,174	21,373	52,662	37,111	51,160	44,803
現代語訳	60,925	4,803	59,412	20,589	57,899	37,991	56,397	48,191
合計	116,612	7,787	113,586	36,202	110,561	69,822	107,557	89,646
重複率	**85.1**%	30.4%	46.0%	15.9%	19.7%	7.6%	8.0%	3.7%
『人世三宝説』の訓練・開発データ								
原文	5,522	1,202	5,278	3,202	5,034	4,139	4,794	4,409
現代語訳	8,250	1,283	8,026	3,875	7,803	5,790	7,583	6,687
合計	13,772	2,146	13,304	6,745	12,837	9,812	12,377	11,056
重複率	**62.2**%	15.8%	15.1%	4.9%	3.7%	1.2%	0.9%	0.4%

コーパスおよび単言語コーパスを用いて事前学習を行った後，人手で説明的な参照訳を取り除いた対訳コーパスで fine-tuning を行い，事前学習の有効性を示した．これらの研究はいずれも教師あり NMT を行っている．本研究では，説明的な参照訳を含む対訳コーパスを用いて，教師あり手法と教師なし手法の両方で実験を行い，ポストエディットの観点から評価を行った．

3. 近代文語文から現代文への機械翻訳

3.1 タスクの概要

近代文語文を，コンピュータを用いて人手を介さずに自動的に現代語の書き言葉に翻訳するタスクである．本タスクの特徴として，原言語と目的言語は同じ日本語であり，その差は通時的に生じたものであることがあげられる．そのため，基本語順や格標示の方法は同じであり，共通の単語が多い．しかし，接続詞や助動詞などの機能語は大きく異なるため，n グラムの重複は n が大きくなるほど少なくなる．

3.2 データセット

本タスクで使用するデータセットは次の 2 つである．

(1) 福澤諭吉の『学問のすゝめ』(1872–1876) とその現代語訳[*1]から抽出した段落単位の対訳コーパス．
(2) 西周の『人世三宝説』(1875) とその現代語訳[*2]から抽出し，文ごとに対応づけた対訳コーパス．

各対訳コーパスに対し 6.1 節の方法で単語分割をしたときの，言語別の単語トークン数[*3]と単語タイプ数[*4]を**表 2** に示す．単語タイプ数は，単語の種類の数のことであり，表層形の同じ単語が複数回出現していても 1 回の出現と数える．『学問のすゝめ』の場合，原文の単語タイプのうち 44.3% は現代語訳にも出現しており，現代語訳の単語タイプのうち 49.4% が原文にも出現している．『人世三宝説』の場合，原文の単語タイプのうち 28.2% は現代語訳にも出現しており，現代語訳の単語タイプのうち 26.4% が原文にも出現している．

ここで，表 1 の 2 つ目および 4 つ目の例のように，2 つのデータセットに含まれる現代語訳が説明的であることに注意する．これらの現代語訳は，近代文語文が持つ歴史的・文化的な違いを現代人にも理解しやすいように原文を噛み砕き，説明が加えられている．**表 3** に，原文に出現する単語のうち，現代語訳と重複する単語について，品詞ごとに集計した結果を示す．『学問のすゝめ』と『人世三宝説』に共通する特徴として，動詞や形容詞は重複が少なく，多くの単語が現代の単語に置き換えられていることが分かる．また，『人世三宝説』の場合，副詞と接続詞の重複が『学問のすゝめ』に比べて極端に少ない．原文に出現するが現代語訳で出現しない副詞には「亦」や「苟も」などがあり，接続詞には「而て」や「即ち」などがある．しかし，説明的な訳を目指さないのであれば，機能語を中心に翻訳することで，原文を現代語として文法的に正しい文に変換することが期待される．

また，重複率が高くても，対応する原文と現代語訳で同じ単語が必ず使われているとは限らない．助動詞はその典型的な例である．近代語の「ざるべからず」という文末表現は現代語では，たとえば「ずにはいられない」と翻訳される．このとき，近代語の「ざる」と「ず」はそれぞれ，「ず」と「ない」に翻訳される．原文と現代語訳にはどちらも「ず」が存在しているが，これらは対応していない．

*1 齋藤 孝（訳）『現代語訳 学問のすすめ』2009 年，ちくま文庫．
*2 菅原 光・相原耕作・島田英明（訳）『西周現代語訳セレクション』2019 年，慶應義塾大学出版会．
*3 述べ語数ともいう．
*4 異なり語数ともいう．

表 3 原文に出現する単語のうち，現代語訳と重複する単語の詳細
Table 3 Details of words in the original texts that overlap with those in contemporary Japanese.

品詞	トークン		タイプ		例		
	数	割合	数	割合	上位 3 単語		
『学問のすゝめ』の訓練・開発データ							
助詞	15,253/15,639	97.5%	35/62	56.5%	の (3,331)	を (3,053)	に (2,232)
名詞	11,310/15,045	75.2%	1,518/3,230	47.0%	こと (459)	人 (342)	もの (342)
補助記号	4,914/6,913	71.1%	8/15	53.3%	、(3,105)	。(1,468)	」 (111)
動詞	4,752/7,215	65.9%	508/1,459	**34.8**%	し (636)	する (293)	あり (225)
助動詞	3,755/4,919	76.3%	28/66	42.4%	ず (738)	に (643)	ざる (321)
副詞	880/1,307	67.3%	86/200	43.0%	ただ (116)	必ず (63)	もとより (57)
接続詞	421/439	95.9%	9/14	64.3%	あるいは (171)	また (151)	すなわち (71)
形状詞	253/419	60.4%	60/142	42.3%	明らか (30)	よう (24)	わずか (23)
形容詞	190/789	**24.1**%	28/193	**14.5**%	なく (69)	多し (28)	多く (10)
その他	2,425/3,002	80.8%	103/257	40.1%	その (661)	これ (470)	この (312)
合計	45,734/55,687	82.1%	2,371/5,355	44.3%			
『人世三宝説』の訓練・開発データ							
助詞	1,551/1,597	97.1%	17/26	65.4%	を (354)	の (344)	に (263)
名詞	783/1,598	49.0%	210/647	32.5%	社交 (60)	人 (56)	者 (56)
動詞	345/871	39.6%	56/338	**16.6**%	し (91)	する (57)	得 (29)
補助記号	255/255	100.0%	7/7	100.0%	。(238)	〔 (5)	〕 (5)
助動詞	229/427	53.6%	10/33	30.3%	ざる (62)	に (58)	ず (48)
形状詞	39/57	68.4%	5/19	26.3%	貴重 (23)	同一 (12)	盛ん (2)
副詞	10/162	**6.2**%	2/52	**3.8**%	必ず (7)	極めて (3)	
形容詞	5/82	**6.1**%	3/35	**8.6**%	なく (3)	なけれ (1)	深く (1)
接続詞	0/74	**0.0**%	0/11	**0.0**%			
その他	84/399	21.1%	17/56	30.4%	相 (18)	我が (13)	上 (9)
合計	3,445/5,522	62.4%	339/1,202	28.2%			

以下に，対訳コーパスの分け方を説明する．

3.2.1 『学問のすゝめ』

訓練・開発データには『学問のすゝめ』の 2 編から 17 編を用いる．段落単位の対訳コーパスから，文ごとに対応づけた対訳コーパスを作成する（1,513 対）．

評価データには『学問のすすめ』の初編を用いる．16 段落からなる段落単位の対訳コーパスであり，近代文語文は 75 文，その現代語訳が 121 文である．

3.2.2 『人世三宝説』

訓練・開発データには『人世三宝説』の 2 編から 4 編を用いる（192 対）．評価データには『人世三宝説』の初編を用いる（78 対）．

4. 機械翻訳の自動評価手法

一般に，NMT モデルの出力には，文意と無関係な単語が出力されるなど，多くの誤りを含むため，ポストエディットが必要である．専門知識を持つポストエディタを利用できる場合，原文の単語を多く含むような，原文との類似度が高い出力は，ポストエディットのコストが小さいと考えられる．なぜなら，原文の単語の修正は，修正箇所の特定や修正に必要なコストが，文意と無関係な単語のときと比べて小さいからである．

しかし，機械翻訳の自動評価指標として一般に用いられている BLEU [7] では，原文との類似度が高い出力を正しく評価できない可能性がある．BLEU は通常，1 つ以上の参照訳を考慮して翻訳システムの出力を評価する．具体的には，出力の n グラムがいずれかの参照訳に含まれるかをもとに計算される修正 n グラム適合率の n が 1 から 4 までの幾何平均と，出力と参照訳の長さの比をもとに計算される brevity penalty との積が評価値であり，値が大きいほど良いシステムであることを意味する．修正 n グラム適合率は，出力が参照訳より短いときに高くなりやすい性質を持つため，短い出力に対して評価値が小さくなるように brevity penalty を掛けている．このように BLEU は参照訳との表層的な一致で評価しているため，仮に出力の単語が翻訳として妥当であったとしても，もしくは，原文の単語と一致していても，参照訳に含まれていなければそのような単語は過小評価されてしまうという問題がある．ま

た，本タスクで参照訳として用いる現代語訳は説明的であるため，参照訳と表層的に一致させることは困難である．このような問題に対する単純な解決策として，原文に対する BLEU の使用が考えられる．

一方，ポストエディットのコストは，原文や参照訳に対する編集距離で見積もることも考えられる．そこで本研究では，BLEU および編集距離がどの程度ポストエディットのコストを考慮できているのかを調査する．

また，BLEU と編集距離はどちらも，参照訳または原文のどちらか一方に対する表層的な一致により評価することが通例であるが，本研究では，参照訳と原文の両方を考慮する評価指標として，各評価指標に対し以下の 4 種類の方法を適用した変種もあわせて調査する．

マルチリファレンス（BLEU のみ） BLEU は複数の参照訳を考慮することができるため，原文も参照訳に加えて評価する．

算術平均 原文に対する評価値と参照訳に対する評価値の算術平均を評価値とする．a と b の加重算術平均は次式で求まる：

$$w \times a + (1-w) \times b.$$

幾何平均 原文に対する評価値と参照訳に対する評価値の幾何平均を評価値とする．a と b の加重幾何平均は次式で求まる：

$$a^w + b^{1-w}.$$

調和平均 原文に対する評価値と参照訳に対する評価値の調和平均を評価値とする．a と b の加重調和平均は次式で求まる：

$$\frac{1}{w/a + (1-w)/b}.$$

ただし，重み w は実数であり，本稿では $w = 0.5$ とした．

5. 事前学習モデルを用いた NMT

翻訳モデルの学習に使用できる対訳コーパスが低リソースであるという問題に対処するために，事前学習モデルを用いる．しかし，近代文語文で事前学習された公開されている言語モデルは存在しない．そこで，まず，近代文語文と現代文の各単言語コーパスを用いて言語モデルを事前学習する．そして，事前学習済み言語モデルで初期化した翻訳モデルに対し，追加学習と fine-tuning の両方，またはいずれか一方を行う．以上の学習の流れを**図 1** に示す．各手法は，Conneau and Lample [1] の翻訳手法に従う．追加学習と fine-tuning はどちらも，単一のモデルで，近代文語文か

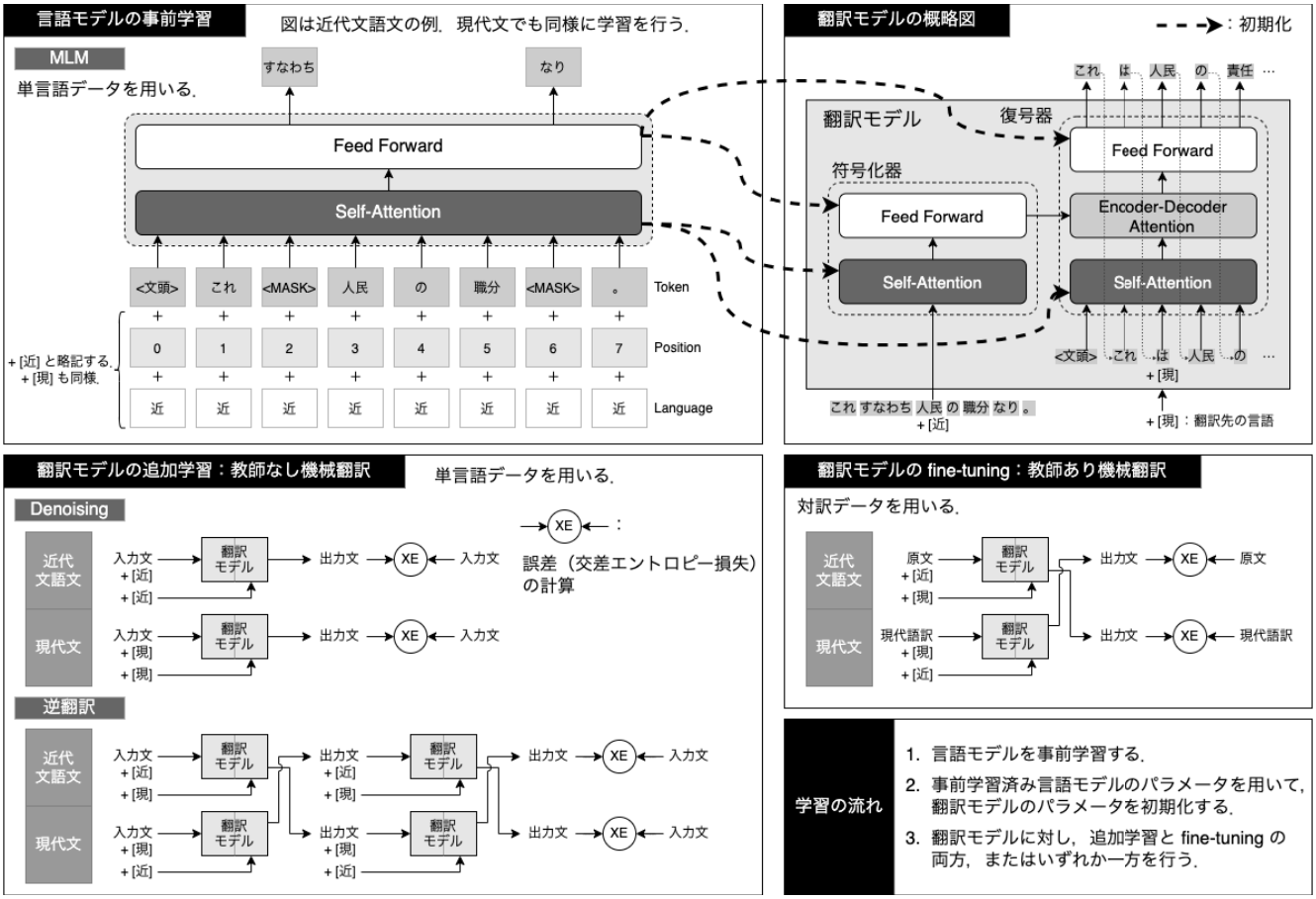


図 1 学習の全体像

Fig. 1 Overview of training process.

ら現代文への翻訳と，現代文から近代文語文への翻訳の両方を行うため，原文の単語を多く出力することが期待できる．

なお，本研究で使用する翻訳モデルは，系列変換モデルである．系列変換モデルは，符号化器と復号器からなる．符号化器は，入力文中の各単語をベクトル表現に変換し，それらをもとに符号を生成する．復号器は，符号化器の生成した符号を用いて出力文を生成する．Conneau and Lample の翻訳手法の場合，言語の情報を符号化器と復号器にそれぞれ入力することで翻訳の方向を制御する．

5.1 翻訳モデルの事前学習

原言語と目的言語に共通する語彙が多いため，言語横断的な言語モデルを用いることで，高い翻訳性能が期待できる．そこで，言語モデルには言語横断的な言語モデルの 1 つである XLM [1] を用いる．近代文語文と現代文の各単言語コーパスを用いて，MLM により共通の言語モデルを作成する．そして，事前学習済み言語モデルのパラメータを用いて，翻訳モデルのパラメータの初期化を行う．

5.2 翻訳モデルの追加学習

本稿では，翻訳モデルに対し単言語コーパスを用いて教師なし手法で学習することを，対訳コーパスを用いる教師あり手法によるものと区別して追加学習と呼ぶ．教師なし手法は，Lample らの手法 [6] と同様である[*5]．すなわち，各単言語コーパスを用いて，denoising と逆翻訳により学習を行う．

5.3 翻訳モデルの fine-tuning

教師あり手法により翻訳モデルの fine-tuning を行う．教師あり手法は，Ramachandran ら [8] の手法を多言語 NMT [4] に拡張した手法であり，対訳コーパスを用いて系列変換の目的関数に従い学習を行う．

6. 実験

6.1 データセット

2 つの対訳データセットに加えて，言語モデルの事前学習に必要な単言語データセットを用意した．対訳データセットは翻訳モデルの fine-tuning にのみ用いるが，単言語データセットは言語モデルの学習と翻訳モデルの追加学習の両方に用いた．

MeCab [5] を用いて形態素解析を行った．形態素解析の辞書には，近代文語文は近代文語 UniDic [17] [*6]，現代文は現代書き言葉 UniDic [15] [*7] を用いた．また，語種が漢語または固有名の場合は，表層形の代わりに語彙素を用いた．

各文は，byte pair encoding（BPE）[9] によりサブワードに分割した．BPE のマージ過程は，後述の単言語データセットの近代文語文と現代文を連結したデータから学習し，マージ回数は 60,000 回に設定した．

6.1.1 対訳データセット

3.2.1 項の訓練・開発データのうち，9 割を訓練データ（1,362 対），1 割を開発データ（151 対）として使用した（『学問のすゝめ』対訳データセット）．

3.2.2 項の訓練・開発データのうち，9 割を訓練データ（173 対），1 割を開発データ（19 対）として使用した（『人世三宝説』対訳データセット）．

6.1.2 単言語データセット

訓練データは近代文語文と現代文の各単言語コーパスを用いた．近代文語文は 226,793 文で固定し，現代文の文数が異なるデータセットを 3 つ用意した：近代文語文と同じ文数，2 倍の文数，5 倍の文数．訓練データに使用した近代文語文と現代文の各単言語コーパスは以下のとおりである．

近代文語文の単言語コーパス 『日本語歴史コーパス』[16] に収録されている文語記事と，国語教科書の文語部分を用いた．

現代文の単言語コーパス 『現代日本語書き言葉均衡コーパス』[*8] の「出版・書籍」を用いた．

単言語データセットの開発データには福澤諭吉の『文明論之概略』（1875）とその現代語訳[*9] から抽出し，文ごとに対応づけた対訳コーパスを使用した（3,306 対）．

6.2 翻訳手法

実験には Conneau and Lample [1] の実装[*10] を利用し，8 つの手法で翻訳モデルの学習を行った．特に言及がない限り，ハイパーパラメータは既定値を使用した．

6.2.1 翻訳モデルの事前学習

近代文語の事前学習済み言語モデルは提供されていない．そのため，独自の言語モデルを学習する必要がある．以下の 2 つの方法で，言語モデルを学習し翻訳モデルを初期化した．

言語横断的な初期化 単言語データセットの 2 つの単言語コーパスを用いて学習した XLM で，翻訳モデルの符号化器と復号器[*11] を初期化する（**XLM**）．

言語横断的でない初期化 近代文語文と現代文の各単言語コーパスを用いて近代文語文と現代文の言語モデル BERT [2] [*12] をそれぞれ学習し，翻訳モデルの符号化

[*5] XLM のルックアップテーブルだけでなく，XLM 全体を初期化する点が異なる．
[*6] https://ccd.ninjal.ac.jp/unidic/download_all#unidic_kindai（`UniDic-kindai_1603`）
[*7] https://ccd.ninjal.ac.jp/unidic/download（`unidic-cwj-2.3.0`）
[*8] https://ccd.ninjal.ac.jp/bccwj/
[*9] 齋藤 孝（訳）『現代語訳 文明論之概略』2013 年，ちくま文庫．
[*10] https://github.com/facebookresearch/XLM
[*11] 符号化器の出力を key と value として受け取る multi-head attention sub-layer を除く．
[*12] 言語を区別しない XLM は，next sentence prediction タスクを行わない BERT と見なせる．

器を近代文語の BERT で，復号器を現代語の BERT で初期化する（**BERT**）．

6.2.2 翻訳モデルの追加学習

言語横断的な初期化を行った翻訳モデルに対し，単言語データセットを用いて，追加学習を行った（**手法 3**）．また，単言語のデータセットの現代文の量を増やしたときの性能を調べるため，現代文の文数のみを 2 倍，5 倍にしたデータセットを用いて，追加学習を行った（**手法 7**，**手法 8**）．さらに，言語横断的な事前学習による効果を調べるため，言語横断的でない初期化を行った翻訳モデルに対しても同様に，追加学習を行った（**手法 6**）．

6.2.3 翻訳モデルの fine-tuning

言語横断的な初期化を行った翻訳モデルに対し，『学問のすゝめ』または『人世三宝説』の対訳データセットを用いて教師あり手法により fine-tuning を行った（**手法 1**，**手法 2**）．また，手法 3 により追加学習済みの翻訳モデルに対し，同様に fine-tuning を行った（**手法 4**，**手法 5**）．

6.3 自動評価

各翻訳手法の傾向を調べるために，次の 4 つの評価指標を用いる：(1) 原文に対する BLEU [7]（**s-BLEU**），(2) 参照訳に対する BLEU（**r-BLEU**），(3) 原文に対する編集距離（**s-編集距離**），(4) 参照訳に対する編集距離（**r-編集距離**）．BLEU は，値が大きいほど評価が高くなり，最大値は 1 である．編集距離は，値が小さいほど類似していることを示す．

各評価データを文ごとにモデルに入力し，翻訳結果を得た後，それらを連結してコーパス単位で評価を行う．

6.4 人手評価

機械翻訳の出力結果を公開できる品質の現代文に修正する際，まず対応する部分を見つけ，それから翻訳を修正すると考える．そこで，評価の観点は以下の 3 つとした．

I 文・節・句・語などの様々な単位で，原文と翻訳システムの出力との対応を調べる時間・人手コスト．
II 対応のとれた部分の翻訳が妥当か判断し，妥当でない場合に修正する時間・人手コスト．
III I と II を総合した，公開できる品質の現代文に修正するための時間・人手コスト．

各評価値は，1 から 10 の離散値で，原文を 5 とした．評価値が時間・人手コストの大きさであるため，値が小さいほど良いシステムと判断できる．

本研究で用いた参照訳は説明的であるため，比較的長く観点 I のコストが増えたり，説明的な翻訳では訳が一意に決まるとは限らないため観点 II のコストが増えたりと，参照訳であっても人手評価で最善でない可能性もある．

『学問のすゝめ』と『人世三宝説』の評価データについて，8 つの翻訳手法の出力例と，原文および参照訳を無作為に並べ替え，合計 10 個のシステムとして評価者に提示する．評価者の負担を考慮し，『学問のすゝめ』は無作為に選んだ 2 段落（原文 2,613 語，参照訳 3,151 語，システム出力の平均 2,747 語），『人世三宝説』は無作為に選んだ 2 文とその周辺を含む 7 文（原文 1,505 語，参照訳 2,014 語，システム出力の平均 1,658 語）のみを提示する．

4 人の評価者（A，B，C，D）によるシステム単位の人手評価を行った．彼らは国立国語研究所の「通時コーパスの構築と日本語史研究の新展開」プロジェクトの非常勤研究員である．評価者ごとの各観点の評価値は付録 A.1 に示す．

評価者間の信頼性を調べるために，ケンドールの一致係数を計算した．『学問のすゝめ』の場合，観点 I は 0.83，観点 II は 0.91，観点 III は 0.88 であった．『人世三宝説』の場合，観点 I は 0.81，観点 II は 0.83，観点 III は 0.83 であった．

原文を 5 とするよう指示したが，4 人のうち 2 人の評価者は原文の評価値を 5 以外の値にすることがあった．これは，10 個のシステムの中に原文が含まれていると明示しなかったことが要因と考えられる．

6.5 実験結果

自動評価および人手評価による各翻訳手法の評価値を**表 4** に示す．人手評価で良い出力であると評価されているのは，言語横断的な初期化をした後，追加学習を行い，fine-tuning をしない手法 3，手法 7，手法 8 である．また，言語横断的でない初期化をする手法 6 に比べ，言語横断的な初期化をする手法 3 の方が人手評価の結果が良い．これは，言語横断的な初期化による事前学習の効果といえるだろう．一方，fine-tuning をする手法は，追加学習の有無によらず人手評価の結果が悪い．

追加学習をせず，fine-tuning をする手法 1 と手法 2 は，s-BLEU と r-BLEU のいずれも低い．単言語データセットで追加学習をする手法 3，手法 6，手法 7，手法 8 は，s-BLEU が高い．追加学習（手法 3）をした後に fine-tuning をする手法 4 と手法 5 は，s-BLEU が比較的低い一方で，評価データと同じドメインで fine-tuning をしたときに，r-BLEU が最も高くなっている．

s-編集距離の小ささは，s-BLEU の大きさとおおむね似た傾向を示しており，追加学習のみを行う手法は原文との類似度が高いことが分かる．また，r-編集距離は，追加学習をしない，もしくは，『人世三宝説』対訳データセットで fine-tuning をしたときに大きくなっている．

人手評価は，追加学習のみを行う手法で高く，fine-tuning をする手法で低かったが，同様の傾向を示している評価指標は s-BLEU と s-編集距離であることが分かった．7 章では，変種も含めた評価指標のメタ評価を行う．ただし，メタ評価に用いる人手評価は観点 III の評価値を用いる．こ

表 4 各翻訳手法の自動評価および人手評価による評価値．↑は値が大きいほど良い，↓は値が小さいほど良い評価指標であることを表す．「学問」「人世」「単言語」はそれぞれ『学問のすゝめ』対訳データセット，『人世三宝説』対訳データセット，単言語データセットを表す

Table 4 Evaluation scores of each system by each automatic metric and human evaluation. The ↑ indicates that the higher the better, and the ↓ indicates that the lower the better.

	事前学習	追加学習	fine-tuning	BLEU		編集距離		人手評価（平均）		
	手法	データ	データ	s ↑	r ↑	s ↓	r ↓	I ↓	II ↓	III ↓
評価データ：『学問のすゝめ』										
		原文		100.00	15.03	0	2,509	3.25	5.75	5.25
		参照訳		15.29	100.00	2,509	0	2.75	1.50	2.00
1	XLM	—	**学問**	6.98	15.97	2,644	3,068	9.00	10.00	9.50
2	XLM	—	人世	2.69	8.93	2,824	3,212	10.00	10.00	10.00
3	XLM	単言語	—	57.77	20.65	1,040	2,692	2.50	4.25	4.00
4	XLM	単言語	**学問**	21.20	**24.42**	2,224	2,789	7.00	8.50	8.25
5	XLM	単言語	人世	2.99	11.63	2,978	3,258	9.25	10.00	9.75
6	BERT	単言語	—	56.68	19.70	945	2,647	4.50	6.75	6.75
7	XLM	単言語 2×	—	**61.17**	20.20	877	2,561	**1.75**	4.25	3.75
8	XLM	単言語 5×	—	59.37	20.26	**863**	**2,538**	**1.75**	**3.50**	**3.25**
評価データ：『人世三宝説』										
		原文		100.00	3.63	0	2,776	4.00	6.25	6.25
		参照訳		3.82	100.00	2,776	0	3.25	2.75	2.75
1	XLM	—	学問	0.00	13.01	2,748	2,898	9.75	10.00	9.75
2	XLM	—	**人世**	0.00	15.99	3,536	3,425	9.75	9.75	9.75
3	XLM	単言語	—	34.79	9.70	1,296	2,878	4.75	7.25	6.75
4	XLM	単言語	学問	4.02	15.01	2,464	2,816	8.50	9.00	9.50
5	XLM	単言語	**人世**	1.01	**17.25**	3,468	3,349	8.75	9.50	9.25
6	BERT	単言語	—	26.34	10.47	1,535	2,889	6.50	8.25	7.75
7	XLM	単言語 2×	—	**37.79**	10.05	**1,188**	2,841	**2.25**	**5.50**	**5.50**
8	XLM	単言語 5×	—	31.92	11.07	1,334	**2,828**	4.25	5.75	6.00

れは，観点 I と観点 III，および観点 II と観点 III の，評価者別のピアソンの積率相関係数の平均値はそれぞれ，『学問のすゝめ』で 0.92 と 0.98，『人世三宝説』で 0.84 と 0.97 と高いためである，

6.6 出力例

各翻訳手法で学習したシステムの出力例を**表 5** に示す．1 つ目の例は『学問のすゝめ』の出力例であり，各システムの傾向を反映している．s-BLEU の低い手法 1，手法 2，手法 5 では，原文とまったく異なる意味の文を出力している．s-BLEU は次に低いが，r-BLEU が最も高い手法 4 では，原文の単語を多く出力しているものの，原文と無関係な単語も出力している．s-BLEU の高い手法 3，手法 6，手法 7，手法 8 では，原文の単語のほとんどをそのまま出力し，助詞や活用語尾の多くを現代の単語に翻訳している．

2 つ目の例は，対訳データセットを用いた fine-tuning をすることで得られる恩恵を示している．手法 4 のみ「肝要」を原文と同じではなく，「大事」と変えて出力している．これは，対訳データセットを用いた fine-tuning を行うことで，機能語だけでなく内容語も翻訳できるようになることを示している．

3 つ目の例は『人世三宝説』の出力例である．単言語データセットを用いた追加学習をした手法 3，手法 4，手法 6，手法 7，手法 8 は，文末の表現を正しく翻訳できている．また，s-BLEU と r-BLEU がともに高い手法である手法 8 のみ，「たる」を「である」と翻訳している．

6.7 分析：単言語コーパスの量的効果

以下，単言語データセットは近代文語文と現代文の文数が 226,793 文のものを基準とする．

6.7.1 単言語コーパスの量

単言語データセットとして十分な文数を調査するため，学習データの量を変えて翻訳モデルの学習を行い，各自動評価指標で評価した．翻訳モデルの学習は，手法 3 と同様

表 5 各翻訳手法で学習したシステムの出力例．下線は，評価データの同じドメインで fine-tuning を行う手法を表す

Table 5 Examples of the translation output. The underlined numbers correspond to methods that perform fine-tuning on the same domain of the evaluation data.

翻訳手法	出力
評価データ：『学問のすゝめ』	
原文	教え ざる 子孫 なれ ば その 愚 なる も また 怪しむ に 足ら ず 。
参照訳	きちんと 教育 され なかっ た 子ども たち が 、また 愚か に なっ て いく こと も 不思議 で は ない 。
1	そ の よう な こと を 認識 し て 、それ を 持つ もの と すれ ば 、遠慮 なく これ を 怒り 、用人 の 身 を 恐れ て い ない 。
2	然 ども 室 家 の 樂 み より 子孫 の 計 を なす が 如き 皆 、どう し て 比較 する こと など が できる だろう か 。
3	教え て い た 子孫 だ から その 愚 さ も また 怪しむ に 足ら ない 。
4	教え ない 子孫 が 作っ て いく に あたっ て は 、その 愚か さ も また 軽々しく 出 て は いけ ない 。
5	子孫 の こと こそ を 大事 に し 、そ の うえ で 人間 が 生き て いく うえ で の こと に よっ て その 人 の 富有 を 増進 さ せる こと が できる 場合 に は 、そう し て も よい 。
6	教え られ ない 子孫 だ から その 愚 さ も また 怪しむ に 足ら ない 。
7	教え ない 子孫 で 、その 愚 さ も また 怪しむ に 足ら ない 。
8	教え ない 子孫 だ から その 愚 さ も また 怪しむ に 足ら ない 。
原文	学問 を する に は 分限 を 知る こと 肝要 なり 。
参照訳	学問 を する に は 、なす べき こと を 知る こと が 大事 で ある 。
1	学問 を する に は 、人間 の 見識 品格 は 、この よう な 学問 を する 人 で ある 。
2	何 と なれ ば 、墨子 の 説 に 加担 し た よう に 誤解 され る 可能 性 が ある の で 、「 公益 」 と いう 字 に つい て 一 つ の 「 命題 」 を 設け て 説明 し て みよう 。
3	学問 を する に は 分限 を 知る こと が 肝要 です 。
4	学問 を する に は 、分限 を 知る こと が 大切 で ある 。
5	個々 人 の 道徳 上 の こと に つい て は すでに 論じ た の で 、これ 以上 は 述べ ない 。
6	学問 を する に は 、分限 を 知る こと が 肝要 だ 。
7	学問 を する に は 分限 を 知る こと が 肝要 で ある 。
8	学問 を する に は 分限 を 知る こと が 肝要 で ある 。
評価データ：『人世三宝説』	
原文	此 人 畢生 人生 の 快樂 何物 たる を 知る 能は ず 。
参照訳	この よう な 人 は 、生涯 、人 が 生きる うえ で の 本当 の 快楽 が 何 な の か 知る こと が でき ない 。
1	その 人 は 、物事 の ようす を 見る と 、その 人間 の 親 が 生まれ た 原因 の 一 つ で は ない 。
2	早朝 から 夜 まで 仕事 を 体験 し て み て 、そ の うち の どれ ほど が 、完全 に 自分 の ため の もの だ と 言える だろう か 。
3	この 人 畢生 人生 の 快 樂 何物 たる を 知る こと は でき ない 。
4	人間 の 自然 な 働き を 、西洋 人 に 頼る こと を 知る こと は でき ない 。
5	彼 は 、一 生涯 を 尽くし て 増やし 溜め込ん だ 富有 を 他人 に 分け与える など と いう こと は でき ない が 、彼 の 死後 に は 、残され た 富有 を 他人 の ため に 生き て いる と 言えよう 。
6	この 人 が 畢生 の 人生 の 快 樂 何物 たる を 知る こと が でき ない 。
7	この 人 畢生 人生 の 快 樂 何物 たる を 知る こと は でき ない 。
8	この 人 が 畢生 人生 の 快 樂 何物 で ある か 知る こと は でき ない 。

にして行った．

単言語データセットの学習データの量を 1%，5%，25%，100%に変更したときの各自動評価指標の評価値を**表 6** に示す．25%のときでも，性能がほとんど変わらないことが分かる．さらに学習データの量を減らすと，性能が急激に低下している．

6.7.2 単言語コーパスの現代文の量

近代文語文の文数を増やすことに比べ，現代文の文数を増やすことは比較的容易である．そこで，現代文の文数のみを増やしたときに性能がどの程度向上するのか調べるために，手法 3，手法 7，手法 8 を比較する．

各手法について，3 回の実験を行ったときの各自動評価指標の評価値の平均値を**表 7** に示す．現代文を 5 倍程度増やしただけでは，性能はほとんど変わらないことが分かる．表 6 の 25%と 100%の変化が緩やかなことも合わせると，これらの結果は，単言語コーパスの文数を線形に変えても，性能の著しい向上は見込めないことを示唆している．また，本研究で使用したよりも文数の多い現代文を用いて

表 6 単言語データセットの学習データの文数を変えたときの結果．各評価値は 1 回の実験を行ったときの値

Table 6 Results for different numbers of sentences in the training data of the monolingual dataset. Each evaluation score is for single run.

	1%	5%	25%	100%
近代文語文	2,267	11,339	56,698	226,793
現代文	2,267	11,339	56,698	226,793
評価データ：『学問のすゝめ』				
s-BLEU ↑	1.55	23.22	56.94	**57.77**
r-BLEU ↑	6.61	11.63	18.97	**20.65**
s-編集距離 ↓	2,740	1,995	1,041	**1,040**
r-編集距離 ↓	3,194	2,884	**2,650**	2,692
評価データ：『人世三宝説』				
s-BLEU ↑	0.00	17.11	27.75	**34.79**
r-BLEU ↑	5.78	8.46	**11.16**	9.70
s-編集距離 ↓	2,362	1,832	1,583	**1,296**
r-編集距離 ↓	2,851	**2,796**	2,942	2,878

表 7 単言語データセットの現代文の学習データの文数を変えたときの結果．各評価値は 3 回の実験を行ったときの平均値

Table 7 Results for different numbers of contemporary Japanese sentences in the monolingual dataset. Each evaluation score is the average for three runs.

	1×	2×	5×
近代文語文	226,793	226,793	226,793
現代文	226,793	453,586	1,133,965
評価データ：『学問のすゝめ』			
s-BLEU ↑	60.70	**62.83**	60.42
r-BLEU ↑	19.59	19.57	**19.73**
s-編集距離 ↓	907.00	**804.67**	890.67
r-編集距離 ↓	2,604.00	**2,527.00**	2,571.33
評価データ：『人世三宝説』			
s-BLEU ↑	34.86	**36.91**	31.57
r-BLEU ↑	9.57	9.48	**10.51**
s-編集距離 ↓	1,268.00	**1,206.00**	1,311.33
r-編集距離 ↓	2,843.33	2,837.00	**2,823.33**

学習された，公開されている事前学習モデルの活用も検討したい．

7. メタ評価

BLEU および編集距離とその変種がどの程度ポストエディットのコストを考慮できているか調べる．そこで，観点 III について，各自動評価指標の評価値（原文および参照訳の評価値を含む）と人手評価の平均値との相関係数を計算し，各自動評価指標を評価する．相関係数の計算には，ピアソンの積率相関係数およびスピアマンの順位相関係数を用いる．ピアソンの積率相関係数で 2 変数間の線形関係を，スピアマンの順位相関係数で 2 変数間の単調関係を評価する．

自動評価指標は，6 章で用いた 4 つだけでなく，4 章に示した，原文と参照訳の両方を考慮する 4 種類の手法を BLEU と編集距離にそれぞれ適用した，計 11 個を調べる（マルチリファレンスの手法は BLEU のみであることに注意）．

7.1 評価結果

図 2 にメタ評価の結果を示す．観点 III の人手評価の平均値との相関は，BLEU の場合は幾何平均，編集距離の場合は算術平均が最も高い傾向にある．

7.2 議論

原文に対する評価値は，BLEU も編集距離も，原文（三角）の評価値をピークとする山型になっている．これは，良い編集をすればするほど良い出力，悪い編集をすればするほど悪い出力に近づくことを表している．

参照訳に対する評価値は，BLEU も編集距離も，参照訳（四角）が飛び抜けており，原文（三角）とシステム出力（丸）は同程度の評価値になっていることが分かる．これは，本タスクで使用した参照訳が説明的であることを表している．

原文と参照訳をマルチリファレンスとして用いる BLEU は，原文（三角）と参照訳（四角）の評価値は等しく，最大値となるが，これは人手評価の結果に反する．一方で，原文と参照訳を含まない，システム出力と人手評価の関係は最も線形に近く，ピアソンの積率相関係数は，『学問のすゝめ』で 0.96，『人世三宝説』で 0.99 であった．

BLEU の算術平均，幾何平均，調和平均は，人手評価の結果どおり原文（三角）よりも参照訳（四角）の方が良いという関係を保ちながら，原文と参照訳の評価値の両方を考慮することにより単調増加に近づいていることが分かる．特に幾何平均は，スピアマンの順位相関係数が最も高い．これは，BLEU 自身が修正 n グラム適合率の幾何平均であり，2 つの幾何平均の幾何平均は，全体の幾何平均となることから，原文も考慮するための BLEU の自然な拡張といえる．

編集距離の 3 種類の平均の場合，原文（三角）と参照訳（四角）の評価値は，マルチリファレンス BLEU と同様に等しくなる．このうち，スピアマンの順位相関係数が最も高いのは，算術平均である．これは，編集距離が置換，挿入，削除という 3 つの操作の回数の和であることをふまえると，妥当な平均のとり方のように思える．

また，相関が高い評価指標であっても，人手評価では原文より高いにも関わらず低い評価値を与えるか，参照訳よりも高い評価値を与えていることが分かる．特に前者の場

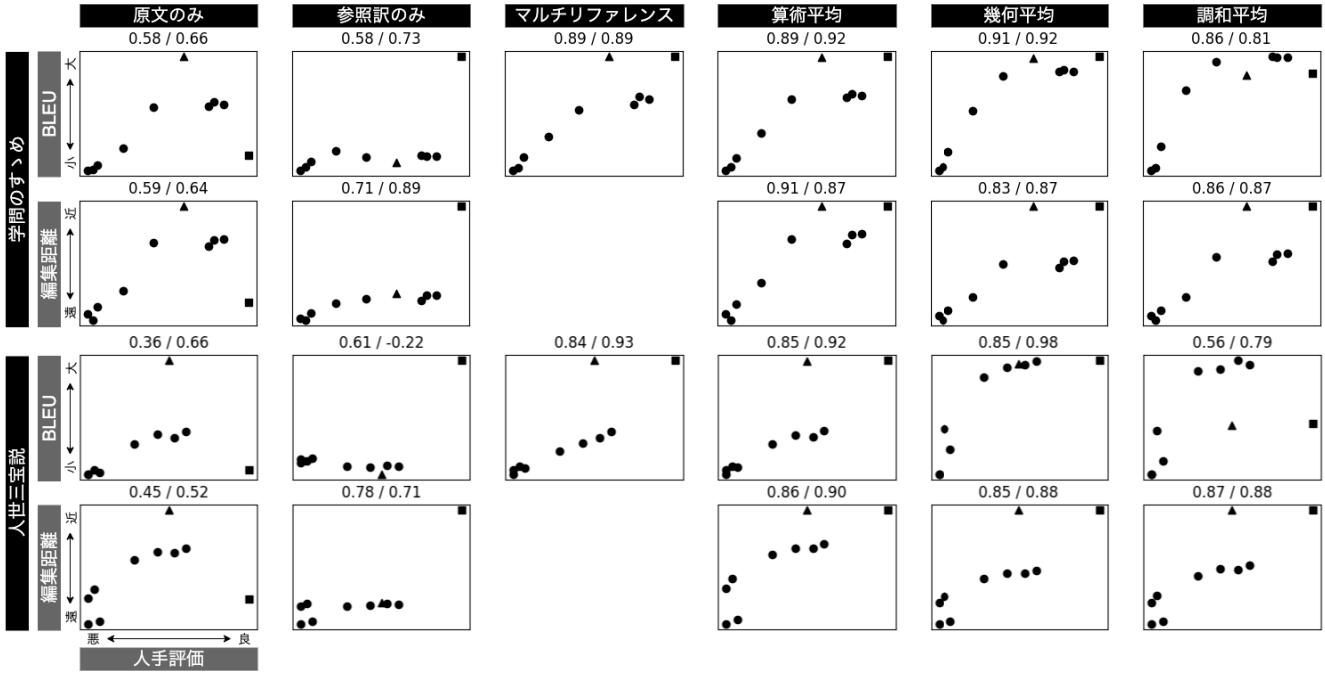


図 2 人手評価の平均値（観点 Ⅲ）と各自動評価指標の評価値の関係．値は，左から順にピアソンの積率相関係数とスピアマンの順位相関係数である．記号は，丸，三角，四角がそれぞれ，システム出力，原文，参照訳に対応している

Fig. 2 Scatter plots of the mean values of human evaluation (viewpoint Ⅲ) and the evaluation scores of each automatic evaluation metric. The values are Pearson product-moment correlation coefficient and Spearman's rank correlation coefficient, from left to right. Circles, triangles, and squares correspond to the system output, source text, and reference translation, respectively.

合，このような評価指標を用いてモデルを選択すると，原文に比べて改善している出力を過小評価する可能性がある．

8. おわりに

本研究では，2 言語間に共通する単語は多いが，共通する n の大きな n グラムはほとんどないという特徴を持つ，近代文語文から現代文へのニューラル機械翻訳を行った．特に，『学問のすゝめ』と『人世三宝説』の翻訳に焦点を当てた．このタスクの，対訳コーパスの量が限られている問題に対応するため，事前学習を用いた．実験の結果，単言語コーパスのみを用いて，原文との類似度が高い出力を得た．ポストエディットの観点で人手評価をすると，これらのシステムの出力は原文よりも良いと判断された．このような出力傾向を持つ NMT モデルを使うことで，ポストエディットがしやすい現代語訳を迅速に提供できる．

一方で，本研究で得られた出力は，機能語や活用語尾の一部を流暢に翻訳するだけにとどまっている．NMT モデルに見られる，一見流暢でも原文の意味とは異なる翻訳を減らすことが今後の課題である．

また，BLEU および編集距離とその変種がどの程度ポストエディットのコストを考慮できているかについて調査した．相関係数の高い評価指標をいくつか見つけることができたが，依然として課題は残っている．

本タスクは，入力文と出力文の重複率が大きいタスクと一般化できる．そのようなタスクには，テキストスタイル変換やテキスト平易化がある．今後は，このような関連タスクの翻訳手法や評価指標をふまえた手法の開発を検討していきたい．

謝辞 本研究は国立国語研究所の共同研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」および所長裁量経費の助成を受けたものです．

参考文献

[1] Conneau, A. and Lample, G.: Cross-lingual Language Model Pretraining, *Advances in Neural Information Processing Systems*, Vol.32, pp.7059–7069 (2019).

[2] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (*Long and Short Papers*), pp.4171–4186 (2019).

[3] Gehring, J., Auli, M., Grangier, D., Yarats, D. and Dauphin, Y.N.: Convolutional Sequence to Sequence Learning, *Proc. 34th International Conference on Machine Learning*, Vol.70, pp.1243–1252 (2017).

[4] Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M. and Dean, J.: Google's Multi-

lingual Neural Machine Translation System: Enabling Zero-Shot Translation, *Transactions of the Association for Computational Linguistics*, Vol.5, pp.339–351 (2017).
[5] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proc. 2004 Conference on Empirical Methods in Natural Language Processing*, pp.230–237 (2004).
[6] Lample, G., Ott, M., Conneau, A., Denoyer, L. and Ranzato, M.: Phrase-Based & Neural Unsupervised Machine Translation, *Proc. 2018 Conference on Empirical Methods in Natural Language Processing*, pp.5039–5049 (2018).
[7] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation, *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, pp.311–318 (2002).
[8] Ramachandran, P., Liu, P. and Le, Q.: Unsupervised Pretraining for Sequence to Sequence Learning, *Proc. 2017 Conference on Empirical Methods in Natural Language Processing*, pp.383–391 (2017).
[9] Sennrich, R., Haddow, B. and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, *Proc. 54th Annual Meeting of the Association for Computational Linguistics*, pp.1715–1725 (2016).
[10] Sutskever, I., Vinyals, O. and Le, Q.V.: Sequence to Sequence Learning with Neural Networks, *Advances in Neural Information Processing Systems*, Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. and Weinberger, K.Q. (Eds.), Vol.27, Curran Associates, Inc. (2014).
[11] Takaku, M., Hirasawa, T., Komachi, M. and Komiya, K.: Neural Machine Translation from Historical Japanese to Contemporary Japanese Using Diachronically Domain-Adapted Word Embeddings, *the 34th Pacific Asia Conference on Language, Information and Computation* (2020).
[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I.: Attention is All you Need, *Advances in Neural Information Processing Systems*, Vol.30, pp.6000–6010 (2017).
[13] 稲見郁乃，竹本有紀，石川由羽，高田雅美，上田　薫，城和貴：邦字新聞における近代文語体と現代口語体の自動翻訳の検討，情報処理学会研究報告，Vol.2020–MPS–131, No.12, pp.1–6 (2020).
[14] 竹内康仁，小林裕之：近代文語体から現代語口語体へのニューラル機械翻訳における事前学習の有効性，電気学会次世代産業システム研究会，pp.15–18 (2021).
[15] 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).
[16] 国立国語研究所：『日本語歴史コーパス』バージョン 2020.3 (2020), 入手先 ⟨https://ccd.ninjal.ac.jp/chj⟩.
[17] 小木曽智信，小町　守，松本裕治：歴史的日本語資料を対象とした形態素解析，自然言語処理，Vol.20, No.5, pp.727–748 (2013).
[18] 星野　翔，宮尾祐介，大橋駿介，相澤彰子，横野　光：対照コーパスを用いた古文の現代語機械翻訳，言語処理学会第 20 回年次大会，pp.816–819 (2014).

付　　録

A.1　人手評価の詳細

評価者ごとの各観点の評価値を**表 A·1** に示す．

A.2　マルチリファレンス BLEU の評価値

マルチリファレンス BLEU による各翻訳手法の評価値を**表 A·2** に示す．4 章で示した，参照訳と原文の両方を考慮する 4 種類の方法のうち，マルチリファレンスを除く，算術平均，幾何平均，調和平均の 3 つを適用したときの評価値は，本文で提示した値から計算できる．

表 A·1 各翻訳手法の人手評価の詳細
Table A·1 Details of human judgment of each translation method.

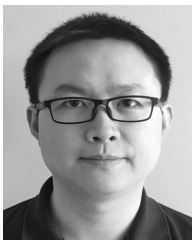
		原文	参照訳	手法 1	手法 2	手法 3	手法 4	手法 5	手法 6	手法 7	手法 8
評価データ：『学問のすゝめ』											
観点 I	A	5	6	7	10	4	7	7	6	1	1
	B	1	2	10	10	1	7	10	3	2	2
	C	2	2	10	10	3	7	10	2	2	3
	D	5	1	9	10	2	7	10	7	2	1
	平均	3.25	2.75	9.00	10.00	2.50	7.00	9.25	4.50	1.75	1.75
観点 II	A	5	2	10	10	7	9	10	7	3	2
	B	5	1	10	10	2	9	10	8	6	4
	C	8	2	10	10	6	8	10	4	6	6
	D	5	1	10	10	2	8	10	8	2	2
	平均	5.75	1.50	10.00	10.00	4.25	8.50	10.00	6.75	4.25	3.50
観点 III	A	5	4	8	10	6	8	9	7	2	1
	B	4	1	10	10	2	9	10	8	6	4
	C	7	2	10	10	6	8	10	4	5	6
	D	5	1	10	10	2	8	10	8	2	2
	平均	5.25	2.00	9.50	10.00	4.00	8.25	9.75	6.75	3.75	3.25
評価データ：『人世三宝説』											
観点 I	A	5	1	9	10	8	10	7	7	1	6
	B	1	2	10	10	1	9	10	8	2	1
	C	5	6	10	10	5	7	10	5	2	5
	D	5	4	10	9	5	8	8	6	4	5
	平均	4.00	3.25	9.75	9.75	4.75	8.50	8.75	6.50	2.25	4.25
観点 II	A	5	1	10	10	7	10	9	8	4	4
	B	5	1	10	10	7	9	10	9	6	5
	C	10	6	10	10	9	8	10	9	8	8
	D	5	3	10	9	6	9	9	7	4	6
	平均	6.25	2.75	10.00	9.75	7.25	9.00	9.50	8.25	5.50	5.75
観点 III	A	5	1	9	10	8	10	8	7	3	5
	B	5	1	10	10	5	10	10	8	7	5
	C	10	6	10	10	8	9	10	9	8	8
	D	5	3	10	9	6	9	9	7	4	6
	平均	6.25	2.75	9.75	9.75	6.75	9.50	9.25	7.75	5.50	6.00

表 A·2 各翻訳手法のマルチリファレンス BLEU による評価値
Table A·2 Evaluation scores of each system by multi-reference BLEU.

評価データ	原文	参照訳	手法 1	手法 2	手法 3	手法 4	手法 5	手法 6	手法 7	手法 8
『学問のすゝめ』	100.00	100.00	21.51	11.58	62.63	38.05	13.54	58.40	68.88	67.22
『人世三宝説』	100.00	100.00	13.32	16.42	36.94	19.37	17.69	30.97	45.77	40.85

喜友名 朝視顕

2021 年東京都立大学システムデザイン学部情報通信システムコース卒業．同年同大学大学院システムデザイン研究科情報科学域博士前期課程に進学．



平澤 寅庄

2021 年東京都立大学大学院システムデザイン研究科情報科学域博士前期課程修了．同年同大学院システムデザイン研究科情報科学域博士後期課程に進学．機械翻訳，特にマルチモーダル機械翻訳，に関心がある．

小町 守 （正会員）

2005 年東京大学教養学部基礎科学科科学史・科学哲学分科卒業．2008 年より日本学術振興会特別研究員（DC2）を経て，2010 年奈良先端科学技術大学院大学情報科学研究科修了．博士（工学）．同年より同研究科助教を経て，2013 年より首都大学東京（現，東京都立大学）システムデザイン学部准教授．大規模なコーパスを用いた意味解析および統計的自然言語処理に関心がある．言語処理学会 20 周年記念論文賞，言語処理学会第 14 回年次大会最優秀発表賞，情報処理学会平成 22 年度山下記念研究賞，2010 年度人工知能学会論文賞等を受賞．人工知能学会，言語処理学会，ACL 各会員．

小木曽 智信 （正会員）

1995 年東京大学文学部日本語日本文学（国語学）専修課程卒業．1997 年東京大学大学院人文社会系研究科日本文化研究専攻修士課程修了．2001 年同博士課程中途退学．2014 年奈良先端科学技術大学院大学情報科学研究科博士課程修了．博士（工学）．2001 年明海大学講師，2006 年独立行政法人国立国語研究所研究員を経て，2009 年人間文化研究機構国立国語研究所准教授，2016 年より教授．専門は日本語学，自然言語処理．言語処理学会，日本語学会各会員．