

系列変換タスクにおける ChatGPT の日本語生成能力の評価 Evaluation of Japanese Language Ability of ChatGPT in Text-to-Text Generation Tasks

樽本 宇宙[†]
Sora Tarumoto

島垣 光希[†]
Koki Hatagaki

宮田 莉奈[†]
Rina Miyata

梶原 智之[†]
Tomoyuki Kajiwara

二宮 崇[†]
Takashi Ninomiya

1. はじめに

近年、自然言語処理やその関連分野において、GPT-3 [1]をはじめとする大規模言語モデルの活用が広がっている。これらの大規模言語モデルは、巨大なニューラルネットワークを大量のテキストコーパス上で長時間訓練したものであり、プロンプトと呼ばれる入出力例を含む指示に従って様々なタスクに適応できる。このような大規模言語モデルの中でも特に ChatGPT¹は、多くのタスクと言語において高い性能を発揮するため、2022 年 11 月の公開当初から大きな注目を集めている。

ChatGPT は GPT-3 の大規模言語モデルに基づいているため、流暢な言語生成能力を持つことが知られている。英語においては、機械翻訳 [2]、自動要約 [3]、テキスト平易化 [4] などの系列変換タスクをはじめとして、様々なタスクにおける ChatGPT の性能評価 [5] の結果が報告されている。ChatGPT による言語生成は、英語だけでなく日本語においても非常に流暢であると感じられる。しかし、ChatGPT の日本語生成能力に関する定量的な調査は、現時点ではまだ充分に行われていない。

本研究では、ChatGPT の日本語生成能力を自動評価および人手評価する。具体的には、英語文から日本語文への機械翻訳、日本語文章の自動要約、日本語文のテキスト平易化の 3 種類の系列変換タスクにおいて、ChatGPT の日本語生成能力を評価する。

評価実験の結果、自動評価においては既存の教師ありモデルの方が ChatGPT よりも高い性能を示したが、人手評価においては ChatGPT の方が高く評価される傾向があった。詳細な分析の結果、ChatGPT は全体的には高品質なテキストを生成できるものの、各タスクにおける詳細な要請に答えられていないことが明らかになった。いずれのタスクにおいても ChatGPT は入力文と意味的に等価でかつ流暢なテキストを生成しやすいものの、機械翻訳においては固有名詞などの語彙選択に課題が残っており、自動要約においては指定した長さでの出力ができず、テキスト平易化においては易しすぎる傾向が見られた。そのため、比較的制約の緩い用途においては ChatGPT の活用が有望であり、細かな制御を必要とする用途では従来の教師ありモデルを用いるという、用途に合わせた使い分けが重要であると言える。

2. 関連研究

本節では、ChatGPT の背景技術について概説する。

2.1 Transformer ベースの転移学習

大量のテキストデータを用いた効率的な訓練が可能な Transformer アーキテクチャ [6] の普及とともに、教師なし

学習に基づく事前訓練の後に目的タスクにおいて教師あり学習の枠組みでのファインチューニングを行う転移学習が広く用いられるようになった。Encoder 型 Transformer を採用した BERT [7]、Decoder 型 Transformer を採用した GPT [8]、Encoder-Decoder 型 Transformer を採用した BART [9] などが提案され、目的タスクに応じてこれらのいずれかをファインチューニングするのが、近年の自然言語処理における主流のアプローチとなっている。

事前訓練には、BERT や BART ではマスク言語モデリング、GPT では自己回帰型言語モデリングが用いられている。前者はマスクされたトークンとその周辺トークンを手がかりに推定するタスクであり、後者は所与のトークン列を手がかりとして次に続くトークンを推定するタスクである。これらのタスクは人手によるアノテーションを必要としないため、大量のテキストデータを用いて訓練できる。

その後、Encoder 型では RoBERTa [10]、Decoder 型では GPT-2 [11]、Encoder-Decoder 型では T5 [12] などの改良版が提案され、GLUE [13] などの一部のベンチマークにおいては言語理解能力に関して人間を上回るという報告も見られるようになった。ひとつのモデルで複数の言語を扱える多言語版も開発が進み、Encoder 型では mBERT [7] や XLM [14-15]、Encoder-Decoder 型では mBART [16] や mT5 [17] などが広く用いられている。

2.2 大規模言語モデルの文脈内学習

言語モデルの性能がデータ量とモデルサイズに強く依存するという Scaling Laws [18] に従い、Transformer ベースの事前訓練モデルの大規模化が進められてきた。このような言語モデルの大規模化の流れの中で開発された GPT-3 [1] は 96 層 1,750 億パラメタという巨大な Decoder 型 Transformer であり、3,000 億トークンという大量のテキストデータ上で自己回帰型言語モデリングを訓練している。

大規模言語モデルである GPT-3 は、タスクの説明と少数の入出力例（プロンプト）を指示するだけで、ファインチューニングなしで様々なタスクに適応できる。これは文脈内学習と呼ばれており、GPT-3 の文脈内学習は質問応答などの一部のタスクにおいて従来のファインチューニングと同等以上の性能を達成し、大きな注目を集めた。

2.3 指示チューニングと強化学習

大規模言語モデルが人間の自然な指示に従うよう調整するために、指示チューニング [19] が研究されている。これは、プロンプトの入力に対して適切な出力を行うよう大規模言語モデルをファインチューニングする手法であり、機械翻訳や自動要約など 12 種類の自然言語処理タスクのデータセットに対して設計されたプロンプトが用いられている。

[†] 愛媛大学 Ehime University

¹ <https://chat.openai.com/>

大規模言語モデルが人間の意図に沿った出力をするよう調整するために、人間によるフィードバックを用いた強化学習 (RLHF) [20] が研究されている。自己回帰型言語モデリングにおいてはトークン単位で尤度最大化の訓練を行っているが、それは必ずしも出力系列の全体として最適とは限らないという Loss-Evaluation Mismatch 問題に対処するために、深層強化学習を適用する。RLHF は以下の 3 ステップで構成される。

1. 大規模言語モデルの指示チューニング
2. 報酬モデルの訓練
3. 大規模言語モデルの強化学習

ステップ 2 では、大規模言語モデルからの複数の出力を人間がランキングしたデータセットを用いて、大規模言語モデルに基づく回帰モデルをランキング学習する。ステップ 3 では、ステップ 1 で構築した大規模言語モデルに対してステップ 2 で構築した報酬モデルを用いて、Proximal Policy Optimization アルゴリズム [21] による強化学習を行う。

3. 評価実験

英語文から日本語文への機械翻訳、日本語文章の自動要約、日本語文のテキスト平易化の 3 種類の系列変換タスクにおいて、ChatGPT の日本語生成能力を評価した。

3.1 共通の実験設定

表 1 に示すデータセットを用いて評価実験を行った。タスクごとに BART [9] などの比較手法を設定し、自動評価および人手評価の両方によって性能を評価した。

ChatGPT に入力するプロンプトは、入出力例を提示せずにタスクの説明のみを与える Zero-shot プロンプトおよび入出力例も与える One-shot プロンプトの 2 種類を実験した。プロンプトの例を図 1 から図 6 に示す。なお、入出力例には、各タスクにおける検証用データの先頭の事例を用いた。

人手評価には、機械翻訳の人手評価に用いられる文意解釈性および文意正確性の評価基準 [22] を用いた。ただし、テキスト平易化タスクにおいては、平易性の観点からの人手評価も実施した。これについては 3.4.1 節にて説明する。

文意解釈性 (5 段階評価)

- 5: 文法的に正しく、言葉遣いも自然である
- 4: 文法的に正しいが、不自然な表現がある
- 3: 表記や文法の誤りがあるが、伝達内容の理解は容易
- 2: 表記や文法の誤り、表現の問題でしっかり読まないと言語内容が理解できない
- 1: 何を伝達しようとしているかが理解できない箇所がある (言葉遣いとして内容伝達に失敗している場合を指し、専門用語の意味が分からない等は除く)

文意正確性 (7 段階評価)

- 7: 参照文と文意が同一と考えて差し支えない
- 6: 参照文と文意に僅かな違いがあるが、ほぼ誤解の心配はない
- 5: 参照文と文意に若干の齟齬はあるが、大きな誤解を招くほどではない
- 4: 参照文と矛盾とまではいかないが重要な情報の誤りや過不足があり文意の重大な誤解が起り得る

表 1: データセットの統計

	訓練用	検証用	評価用
ASPEC	2,000,000	1,790	1,812
KFTT	440,288	1,166	1,159
WMT20	-	-	1,000
JParaCrawl	21,887,738	4,000	-
Livedoor NEWS	101,438	664	677
SNOW	82,300	2,000	100

以下の英語を日本語に翻訳してください。
英語: ここに英語が入る
日本語:

図 1: 機械翻訳における Zero-shot プロンプト

以下の英語例と日本語例を参考に、英語を日本語に翻訳してください。
英語例: ここに英語例が入る
日本語例: ここに日本語例が入る
英語: ここに英語が入る
日本語:

図 2: 機械翻訳における One-shot プロンプト

以下の文章を 3 文の日本語で要約してください。
文章: ここに文章が入る
要約:

図 3: 自動要約における Zero-shot プロンプト

以下の文章例と要約例を参考に、文章を 3 文の日本語で要約してください。
文章例: ここに文章例が入る
要約例: ここに要約例が入る
文章: ここに文章が入る
要約:

図 4: 自動要約における One-shot プロンプト

以下の難解文を日本語で平易化してください。
難解文: ここに難解文が入る
平易文:

図 5: テキスト平易化における Zero-shot プロンプト

以下の難解例と平易例を参考に、難解文を日本語で平易化してください。
難解例: ここに難解例が入る
平易例: ここに平易例が入る
難解文: ここに難解文が入る
平易文:

図 6: テキスト平易化における One-shot プロンプト

表 2: ASPEC における機械翻訳の性能評価

	BLEU	chrF++	BERTScore	BLEURT	COMET	解釈性	正確性
Google 翻訳	30.7	30.2	0.873	0.746	0.729	-	-
DeepL 翻訳	29.0	30.1	0.871	0.736	0.707	-	-
Transformer (JParaCrawl)	23.9	26.8	0.852	0.688	0.585	-	-
Transformer (ASPEC)	40.7	35.9	0.890	0.733	0.680	-	-
Transformer (JParaCrawl→ASPEC)	43.0	37.2	0.896	0.746	0.714	-	-
mBART (JParaCrawl)	24.3	27.0	0.851	0.689	0.585	-	-
mBART (ASPEC)	42.6	36.7	0.894	0.741	0.705	-	-
ChatGPT (Zero-shot)	19.3	24.6	0.842	0.703	0.626	-	-
ChatGPT (One-shot)	20.0	24.7	0.846	0.707	0.646	-	-

表 3: KFTT における機械翻訳の性能評価

	BLEU	chrF++	BERTScore	BLEURT	COMET	解釈性	正確性
Google 翻訳	24.6	24.9	0.857	0.689	0.529	4.63	5.51
DeepL 翻訳	23.2	24.5	0.843	0.673	0.464	-	-
Transformer (JParaCrawl)	18.1	20.9	0.814	0.605	0.292	-	-
Transformer (KFTT)	31.4	30.3	0.862	0.679	0.422	-	-
Transformer (JParaCrawl→KFTT)	32.0	30.9	0.868	0.693	0.490	-	-
mBART (JParaCrawl)	17.0	20.4	0.809	0.599	0.272	-	-
mBART (KFTT)	33.4	32.3	0.874	0.706	0.509	4.49	5.29
ChatGPT (Zero-shot)	15.9	10.2	0.816	0.660	0.442	4.55	5.34
ChatGPT (One-shot)	17.2	21.7	0.820	0.665	0.433	4.68	5.42

表 4: WMT20 ニュース翻訳タスクにおける機械翻訳の性能評価

	BLEU	chrF++	BERTScore	BLEURT	COMET	解釈性	正確性
Google 翻訳	27.3	28.0	0.840	0.643	0.542	4.34	5.38
DeepL 翻訳	26.2	26.7	0.840	0.658	0.581	-	-
Transformer (JParaCrawl)	21.8	23.3	0.819	0.585	0.337	-	-
mBART (JParaCrawl)	23.2	24.6	0.823	0.599	0.380	4.00	4.90
ChatGPT (Zero-shot)	21.5	25.1	0.828	0.654	0.580	4.52	5.51
ChatGPT (One-shot)	22.1	24.3	0.831	0.657	0.583	4.47	5.55

- 3: 参照文と矛盾する内容が伝達されている
- 2: 参照文と無関係な内容が伝達されている
- 1: 文意が理解できず、正確性評価に値しない

3.2 機械翻訳タスクにおける評価実験

3.2.1 実験設定

データセット 様々なドメインを扱うために、科学技術論文の抄録から作成された ASPEC² [23]、Wikipedia の京都関連記事から作成された KFTT³、WMT20 ニュース翻訳タスク⁴ [24] の 3 種類の英日対訳コーパスを用いて評価した。また、比較手法の訓練用に大規模な英日対訳コーパスとして、Web から抽出された JParaCrawl v3.0⁵ [25] も用いた。

ベースライン 比較手法の機械翻訳モデルとして、事前訓練なしの Transformer [6]、JParaCrawl を用いて事前訓練された Transformer⁵、マスク言語モデリングによって事前

訓練された mBART⁶ [16] に加えて、オンライン機械翻訳サービスである Google 翻訳⁷ および DeepL 翻訳⁸ を用いた。Transformer および mBART の訓練には、fairseq ツールキット⁹ [26] を用いた。前処理として、事前訓練なしの Transformer には SentencePiece¹⁰ [27] を用いてサブワード分割を行った。語彙サイズは言語ごとに 16,000 に設定した。事前訓練された Transformer および mBART は、それぞれのトークナイザを用いてサブワード分割を行った。また、テキストには NFKC 正規化を施した。

評価指標 自動評価には、SacreBLEU¹¹ [28] を用いて BLEU [29] および chrF++ [30] を計算した。また、分散表現ベースの BERTScore¹² [31]、BLEURT¹³ [32-33]、COMET¹⁴ [34] も用いた。ただし、BERTScore には bert-base-multilingual-cased モデル、BLEURT には BLEURT-20 モデル、COMET には wmt20-comet-da モデルを、それぞれ用いた。

² <https://jipsti.jst.go.jp/aspec/>

³ <https://www.phontron.com/kftt/>

⁴ <https://www.statmt.org/wmt20/translation-task.html>

⁵ <https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/>

⁶ <https://github.com/facebookresearch/fairseq/tree/main/examples/mbart>

⁷ <https://cloud.google.com/translate/>

⁸ <https://www.deepl.com/pro-api/>

⁹ <https://github.com/facebookresearch/fairseq>

¹⁰ <https://github.com/google/sentencepiece>

¹¹ <https://github.com/mjpost/sacrebleu>

¹² https://github.com/Tiiiger/bert_score

¹³ <https://github.com/google-research/bleurt>

¹⁴ <https://github.com/Unbabel/COMET>

表 5: 人手評価における評価者間の一致度 (Quadratic Weighted Kappa)

評価者	KFTT		WMT20		Livedoor NEWS		SNOW		平易性
	解釈性	正確性	解釈性	正確性	解釈性	正確性	解釈性	正確性	
A-B	0.55	0.31	0.36	0.43	0.46	0.45	0.60	0.57	0.50
A-C	0.52	0.49	0.28	0.24	0.61	0.40	0.61	0.26	0.54
B-C	0.45	0.31	0.45	0.30	0.45	0.57	0.62	0.33	0.39

3.2.2 自動評価による実験結果

表 2 および表 3 に、ドメイン内の訓練用データを利用可能な ASPEC および KFTT における機械翻訳の実験結果を示す。実験結果から、ChatGPT はドメイン外の JParaCrawl によって訓練された機械翻訳モデル、つまり、Transformer (JParaCrawl) および mBART (JParaCrawl) と同程度の翻訳品質を持つことがわかる。また、ChatGPT はドメイン内の訓練用データを用いる機械翻訳モデル、つまり、Transformer (ASPEC/KFTT) や Transformer (JParaCrawl) → ASPEC/KFTT) や mBART (ASPEC/KFTT) を下回る翻訳品質を持つことがわかる。さらに、ChatGPT は Google 翻訳 および DeepL 翻訳のオンライン翻訳サービスを下回る翻訳品質を持つことがわかる。

表 4 に、ドメイン内の訓練用データを利用できない WMT20 ニュース翻訳タスクにおける機械翻訳の実験結果を示す。実験結果から、ChatGPT は分散表現ベースの自動評価指標において、Transformer や mBART の機械翻訳モデルよりも高い翻訳品質を示し、オンライン機械翻訳サービスと比べても同等以上の翻訳品質を達成することがわかる。

これらの実験結果をまとめると、ChatGPT は汎用的な機械翻訳モデルと同程度の翻訳品質を持つものの、ドメイン内の訓練用データを利用できる場合には従来のファインチューニングのアプローチを選択すべきであると言える。ドメインについては、ChatGPT はニュースドメインに強く、ニュースドメインにおいてはオンライン機械翻訳サービスとも競合する性能を持つ。また、ChatGPT (Zero-shot) と ChatGPT (One-shot) の比較から、プロンプトに入出力例を含めることで翻訳品質を改善できることを確認できた。

3.2.3 人手評価による実験結果

無作為抽出した 100 文ずつを、日本語母語話者である大学生 3 名が人手評価した結果を表 3 および表 4 に示す。両データセットにおいて、文意解釈性については ChatGPT が最高性能を達成しており、人間にとって自然に感じられる日本語文を生成できていることがわかる。KFTT における文意正確性については、ChatGPT は mBART と Google 翻訳の中間程度の性能を持つことがわかる。また、WMT20 ニュース翻訳タスクにおける文意正確性については、ChatGPT が最高性能を達成しており、やはり ChatGPT はニュースドメインに明るいことがわかる。

人手評価における評価者間の一致度を Quadratic Weighted Kappa [35] を用いて計算した結果を表 5 に示す。WMT20 ニュース翻訳タスクにおいて、評価者間の一致度は充分な一致とは言えないものの、先行研究 [22] における機械翻訳の人手評価においても評価者間の一致度は 0.25~0.39 であるため、許容範囲内であると考えられる。なお、評価者間の一致度については自動要約 (Livedoor NEWS) やテキスト平易化 (SNOW) のタスクにおいても同様の傾向である。

表 6: 自動評価と人手評価の文単位のピアソン相関

	KFTT		WMT20	
	解釈性	正確性	解釈性	正確性
BLEU	0.171	0.239	0.150	0.242
chrF++	0.185	0.241	0.152	0.248
BERTScore	0.178	0.228	0.209	0.288
BLEURT	0.344	0.391	0.377	0.424
COMET	0.384	0.365	0.365	0.382

表 7: “the Council for Kinki Regional Transport” の翻訳

機械翻訳モデル	日本語訳
Google 翻訳	近畿運輸審議会
mBART (KFTT)	近畿地方交通審議会
ChatGPT (Zero-shot)	近畿運輸評議会
ChatGPT (One-shot)	近畿運輸審議会
参照文	近畿地方交通審議会

3.2.4 分析

表 4 において、BLEU や chrF++ の自動評価からは mBART が ChatGPT よりも優れているように見えるが、分散表現ベースの自動評価や人手評価においては ChatGPT の方が高く評価されている。このような評価の振る舞いについて正確に理解するために、自動評価と人手評価の文単位のピアソン相関を求めた結果を表 6 に示す。実験結果から、BLEU および chrF++ の表層マッチングに基づく自動評価は人手評価との相関が低く信頼性に欠けると言える。特に、文意解釈性との相関が低い。また、分散表現ベースの評価指標の中でもトークン単位で処理をする BERTScore は人手評価との相関が低い。これらの結果から、英日機械翻訳の自動評価においては、分散表現ベースの評価指標のうち文単位で処理をする BLEURT や COMET を用いるのが良いと言える。

また、出力事例を観察したところ、ドメイン内の対訳コーパスを用いて訓練したモデルは固有名詞などの当該ドメインに特化した表現を正確に翻訳できている傾向が見られた。その一例を表 7 に示す。この例では、「近畿地方交通審議会」と翻訳すべきところを、ChatGPT は「近畿運輸評議会」や「近畿運輸審議会」などと出力してしまっている。一方で、ドメイン内の対訳コーパスを用いて訓練した mBART は、正しく「近畿地方交通審議会」と出力できている。ChatGPT の出力も意味的には妥当な翻訳であるため、これらは人手評価においては顕著な差としては現れていない。ただし、このような固有名詞や専門用語を正確に翻訳したい用途においては、オンライン機械翻訳サービスや ChatGPT ではなく、ドメイン内の訓練データを用いて従来のファインチューニングのアプローチを選択すると良い。また、表 2 および表 3 において、ドメイン内の訓練データを用いる機械翻訳モデルの BLEU や chrF++ が他のモデルよりも著しく高いのは、このようなドメイン特有の表現を正確に出力できることが要因のひとつであると考えられる。

表 8 : Livedoor NEWS における自動要約の性能評価

	ROUGE-1	ROUGE-2	ROUGE-L	解釈性	正確性	～2 文 (%)	3 文 (%)	4 文～ (%)
LEAD-3	44.81	20.85	41.27	4.48	3.91	-	-	-
BART	51.23	26.19	47.67	4.53	4.24	8.9	89.6	1.5
ChatGPT (Zero-shot)	48.15	20.72	43.62	4.76	4.70	10.6	58.9	30.5
ChatGPT (One-shot)	48.74	21.54	44.48	4.84	4.84	13.9	43.8	42.3

3.3 自動要約タスクにおける評価実験

3.3.1 実験設定

データセット 本実験では、Livedoor NEWS¹⁵から作成された 3 行要約データセット¹⁶ [36] を用いて、自動要約の性能を評価した。これは、ニュース記事の本文とその 3 行要約で構成される日本語の自動要約データセットである。本実験の実施時点で削除されてしまったニュース記事は取得できなかったため、先行研究 [36] よりもデータセットの規模が小さいことに注意されたい。

ベースライン 比較手法の自動要約モデルとして、記事の先頭 3 文を抽出する LEAD-3 およびマスク言語モデリングによって事前訓練された日本語 BART¹⁷ [16] を用いた。LEAD-3 における 3 文の抽出には、ルールベースの文分割である ja_sentence_segmenter¹⁸ を用いた。BART の実験においては、前処理として Juman++ v2¹⁹ [37-38] による単語分割の後にサブワード分割を行った。

評価指標 自動評価には、PythonROUGE²⁰ を用いて ROUGE [39] (ROUGE-1, ROUGE-2, ROUGE-L) の F 値を計算した。なお、テキストには NFKC 正規化を施した。

3.3.2 自動評価による実験結果

ROUGE による自動評価の結果を表 8 に示す。ChatGPT は、抽出型の LEAD-3 よりも高品質であるものの、従来のファインチューニングのアプローチである BART を下回る性能であった。また、機械翻訳と同じく、ChatGPT (Zero-shot) と ChatGPT (One-shot) の比較から、プロンプトに入出力例を含めることで要約品質を改善できることを確認できた。

3.3.3 人手評価による実験結果

無作為抽出した 100 文ずつを、日本語母語話者である大学生 3 名が人手評価した結果を表 8 に示す。文意解釈性と文意正確性の両方において、ChatGPT が最高性能を示した。

3.3.4 分析

出力事例を観察したところ、ChatGPT による要約が 3 文を超える事例が多く見られた。そこで、ja_sentence_segmenter による文分割を用いて、要約の文数を定量的に評価した結果を表 8 の右 3 列に示す。実験結果から、BART は約 9 割が期待通り 3 文の要約を生成できていることがわかる。3 文を超える要約はわずか 1.5% であった。一方で、ChatGPT の要約は 3 割から 4 割が 3 文を超え

る長い要約になっており、2 文以下のものも含めると 4 割から 5 割が要約文数の制約を満たしていない。これらの実験結果から、プロンプトに「3 文」という説明や 3 文に要約した事例を含めるだけでは、要約文数の制約を十分に満たせないことが明らかになった。3.3.3 節で述べたように、ChatGPT による要約は流暢かつ入力との意味的な整合性もとれてはいるため、文数を厳密に指定しない用途での自動要約には適している。しかし、厳密な要約長の制約がある用途では、従来のファインチューニングのアプローチを選択するのが良いと言える。

3.4 テキスト平易化タスクにおける評価実験

3.4.1 実験設定

データセット 本実験では、やさしい日本語コーパス SNOW^{21,22} [40-41] を用いて、テキスト平易化の性能を評価した。これは、教科書などの文を手で平易化して作成された日本語のテキスト平易化のためのデータセットである。

ベースライン 比較手法のテキスト平易化モデルとして、事前訓練なしの Transformer [6] およびマスク言語モデリングによって事前訓練された日本語 BART [16] を用いた。BART の実験設定に揃えるために、Transformer の実験においても Juman++ v2 [37-38] による単語分割の後に、SentencePiece [27] によるサブワード分割を行った。ここで、語彙サイズは 32,000 とした。

評価指標 自動評価には、BERTScore [31] に加えて EASSE²³ [42] を用いて BLEU [29] および SARI [43] を計算した。人手評価には、文意解釈性および文意正確性の観点に加えて平易性の観点から以下の基準で 5 段階評価を行った。

平易性 (5 段階評価)

- 5 : 原文よりも理解しやすい
- 4 : 原文よりも僅かに理解しやすい
- 3 : 原文と理解しやすさに変わりがない
- 2 : 原文よりも理解しにくい
- 1 : 原文と無関係な内容であり、平易性評価に値しない

3.4.2 自動評価による実験結果

自動評価の結果を表 9 に示す。ChatGPT は、3 つの自動評価指標の全てにおいて、Transformer および BART の両方の性能を下回った。また、機械翻訳や自動要約とは異なり、プロンプトに入出力例を含めても性能は改善しなかった。

¹⁵ <https://news.livedoor.com/>

¹⁶ <https://github.com/KodairaTomonori/ThreeLineSummaryDataset>

¹⁷ <https://nlp.ist.i.kyoto-u.ac.jp/?BART> 日本語 Pretrained モデル

¹⁸ https://github.com/wwwcojp/ja_sentence_segmenter

¹⁹ <https://github.com/ku-nlp/jumanpp>

²⁰ <https://github.com/tagucci/pythonrouge>

²¹ <https://www.inlp.org/GengoHouse/snow/t15>

²² <https://www.inlp.org/GengoHouse/snow/t23>

²³ <https://github.com/feralvam/easse>

表 9: SNOW におけるテキスト平易化の性能評価

	BLEU	SARI	BERTScore	解釈性	正確性	平易性
Transformer	69.43	62.70	0.892	-	-	-
BART	85.78	66.40	0.911	4.73	4.91	3.93
ChatGPT (Zero-shot)	36.70	45.50	0.835	4.91	5.80	4.01
ChatGPT (One-shot)	36.57	44.71	0.829	4.87	5.79	4.03
参照文	-	-	-	4.79	6.18	3.72

3.4.3 人手評価による実験結果

評価用データの全体を、日本語母語話者である大学生 3 名が人手評価した結果を表 9 に示す。自動評価とは対照的に、3 つの評価観点の全てにおいて ChatGPT の性能が BART を上回った。

3.4.4 分析

出力事例を観察したところ、ChatGPT が過度に平易化を行っている事例が多く見られた。そこで、SNOW の参照文に対しても人手評価を行い、参照文の平易性と ChatGPT による生成文の平易性を比較することにした。表 9 の最下段に示した実験結果から、SNOW の参照文はあまり顕著な平易化を行っていないことがわかる。それに対して ChatGPT は、参照文を基準にするとは過度に平易化していると言える。言語教育の分野では、学習者の能力よりわずかに高い難易度の教材を用いることで高い学習効果が得られるというインプット仮説 [44] が知られており、言語学習者に対してのテキストの難易度制御 [45] が重要視されている。このような用途で文の難易度を細かく制御することを考えると、過度な平易化を行う ChatGPT は適用しにくい。一方で、入力文よりも平易であれば良いというような設定では、流暢かつ平易な文を生成できる ChatGPT が適していると言える。

4. おわりに

本研究では、機械翻訳・自動要約・テキスト平易化の 3 種類の系列変換タスクにおいて、ChatGPT の日本語生成能力を評価した。いずれのタスクにおいても、ChatGPT が生成した日本語テキストは文意解釈性および文意正確性の両観点で高品質であった。とはいえ、機械翻訳における固有名詞などの語彙選択、自動要約における生成文数の制御、テキスト平易化における生成文の難易度の制御など、各タスクにおける詳細な要請には応えられていないことが明らかになった。そのため、現時点での結論としては、比較的制約の緩い用途においては ChatGPT の活用が有望であるものの、細かな制御を必要とする用途では従来のファインチューニングのアプローチの方が適している。

今後は、他のタスクにおける ChatGPT の日本語能力の評価や、改良版である GPT-4 [46] との比較に取り組みたい。

謝辞

本研究は JSPS 科研費（基盤研究 B、課題番号：JP22H03651）および国立研究開発法人情報通信研究機構の委託研究（課題番号：225）による助成を受けたものです。また、本研究を進めるにあたり、株式会社レトリバの勝又智氏に有益なご助言をいただきました。

参考文献

- [1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, A.a Askell, Jini Aggarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McElish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, pp.1877-1901, 2020.
- [2] Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. arXiv:2301.08745, 2023.
- [3] Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. Exploring the Limits of ChatGPT for Query or Aspect-based Text Summarization. arXiv:2302.08081, 2023.
- [4] Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. Sentence Simplification via Large Language Models. arXiv:2302.11957, 2023.
- [5] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. arXiv:2302.04023, 2023.
- [6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pp.5998-6008, 2017.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.4171-4186, 2019.
- [8] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. *Improving Language Understanding by Generative Pre-Training*. pp.1-12, 2018.
- [9] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp.7871-7880, 2020.
- [10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692, 2019.
- [11] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. pp.1-24, 2019.
- [12] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, Vol.21, No.140, pp.1-67, 2020.
- [13] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp.353-355, 2018.

- [14] Guillaume Lample and Alexis Conneau. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, 2019.
- [15] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp.8440-8451, 2020.
- [16] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, Vol.8, pp.726-742, 2020.
- [17] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.483-498, 2021.
- [18] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. arXiv:2001.08361, 2020.
- [19] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned Language Models Are Zero-Shot Learners. In *Proceedings of the Tenth International Conference on Learning Representations*, 2022.
- [20] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems*, pp.27730-27744, 2022.
- [21] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. arXiv:1707.06347, 2017.
- [22] Katsuhito Sudoh, Kosuke Takahashi, and Satoshi Nakamura. Is This Translation Error Critical?: Classification-Based Human and Automatic Machine Translation Evaluation Focusing on Critical Errors. In *Proceedings of the Workshop on Human Evaluation of NLP Systems*, pp.46-55, 2021.
- [23] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian Scientific Paper Excerpt Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pp.2204-2208, 2016.
- [24] Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pp.1-55, 2020.
- [25] Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp.6704-6710, 2022.
- [26] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp.48-53, 2019.
- [27] Taku Kudo and John Richardson. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp.66-71, 2018.
- [28] Matt Post. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp.186-191, 2018.
- [29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp.311-318, 2002.
- [30] Maja Popović. chrF++: Words Helping Character N-grams. In *Proceedings of the Second Conference on Machine Translation*, pp.612-618, 2017.
- [31] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*, pp.1-43, 2020.
- [32] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp.7881-7892, 2020.
- [33] Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. Learning Compact Metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp.751-762, 2021.
- [34] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp.2685-2702, 2020.
- [35] Jacob Cohen. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, Vol.70, No.4, pp.213-220, 1968.
- [36] 小平知範, 小町守. TL;DR 3行要約に着目したニューラル文書要約. *WebDB Forum*, pp.193-198, 2017.
- [37] Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp.2292-2297, 2015.
- [38] Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. Juman++: A Morphological Analysis Toolkit for Scriptio Continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp.54-59, 2018.
- [39] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp.74-81, 2004.
- [40] Takumi Maruyama and Kazuhide Yamamoto. Simplified Corpus with Core Vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pp.1153-1160, 2018.
- [41] Akihiro Katsuta and Kazuhide Yamamoto. Crowdsourced Corpus of Sentence Simplification with Core Vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pp.461-466, 2018.
- [42] Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. EASSE: Easier Automatic Sentence Simplification Evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing: System Demonstrations*, pp.49-54, 2019.
- [43] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, Vol.4, pp.401-415, 2016.
- [44] Stephen D. Krashen. *The Input Hypothesis: Issues and Implications*. Longman, 1985.
- [45] Daiki Yamamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. Controllable Text Simplification with Deep Reinforcement Learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pp.398-404, 2022.
- [46] OpenAI. GPT-4 Technical Report. arXiv:2303.08774, 2023.

付録

機械翻訳・自動要約・テキスト平易化における ChatGPT の実際の出力例をそれぞれ表 10、表 11、表 12 に示す。

表 10：機械翻訳における ChatGPT の出力例

プロンプト	以下の英語を日本語に翻訳してください。 英語：It has recently been indicated that Naka no Oe no Oji (Naka no Oe no Miko) and Sogano Iruka had a relatively good relationship and similar basic policies. 日本語：
ChatGPT の出力	最近、中大兄皇子（中大兄皇子）と曾我廬穴が比較的良好な関係と似た基本政策を持っていたことが示されています。
BART の出力	中大兄皇子(中大兄皇子)と蘇我入鹿は比較的良好で類似した基本政策を持っていたことが近年指摘されている。
正解文	近年中大兄皇子と蘇我入鹿の関係が比較的良好であり、基本政策も似ていることが指摘されている。

表 11：自動要約における ChatGPT の出力例

プロンプト	以下の文章を 3 文の日本語で要約してください。 文章：日本を訪れた中国人旅行客の多くが、日本の「トイレ」に驚くという。温水洗浄便座の機能だけでなく、トイレという空間そのものが中国とは大きく異なるためだ。中国の公衆トイレはお世辞にも清潔とは言い難く、特に観光地のように多くの人が使用するトイレは有料であることが多い。トイレトペーパーも有料、さらに不潔で酷い有様だ。その点、日本の公衆トイレは清潔であることが多く、トイレトペーパーも無料で使用できる。トイレトペーパーは予備のものが常設されて、誰かが勝手にトイレトペーパーを持ち去ることも少ないだろう。中国メディアの今日頭条はこのほど、「中国人はなぜ日本のトイレを高く評価するのか」と題した記事を掲載し、日本のトイレについて、中国人の間で特に評価の高いポイントを紹介している。まず記事は、前出のとおり、日本のトイレは「無料で利用できる」と紹介し、公衆トイレは無料で利用できるうえに無料のトイレトペーパーが設置されていると称賛。お金を払ってトイレトペーパーを購入し、お金を払って公衆トイレを利用している中国人にとっては「日本のトイレはそれだけで感動的」と伝えた。さらに、女性用トイレに設置されていることの多いトイレ用擬音装置を高く評価。やはり中国人女性も日本人女性と同様にトイレの時の音は気になるようで、トイレ用擬音装置で恥ずかしい思いをしないで済むと高く評価している。また記事は、日本のトイレは使った紙をそのまま流すことができる点も便利であるとしている。中国は排水管が詰まりやすいため、使用済みのトイレトペーパーを流せない場所もある。そういった場合は、トイレ内に設置してあるゴミ箱に捨てるのだが、日本の習慣を知らずに訪日した中国人旅行客が日本のトイレでも使用済みのトイレトペーパーを流さずにゴミ箱に捨てるケースがあるという。日本のトイレは世界的に見ても清潔で、利用者の立場で考えられた設備が数多く存在し、ほかの国のようにトイレは「汚い場所」ではない。中国から日本を訪れて、日本のトイレの先進性や快適さに中国人が驚くのもある意味で当然なのだ。（編集担当：村山健二）（イメージ写真提供：1 2 3 R F） 要約：
ChatGPT の出力	中国人旅行客が日本のトイレに驚くのは、清潔さや無料で利用できる点が大きな理由だ。また、女性用トイレに設置されているトイレ用擬音装置も高く評価されている。中国ではトイレトペーパーを流せない場合もあるため、日本のトイレの利便性にも感心している。日本のトイレは世界的に見ても清潔で、快適な場所とされている。
BART の出力	日本を訪れた中国人旅行客の多くが、日本の「トイレ」に驚くという。温水洗浄便座の機能だけでなく、トイレという空間そのものが中国とは大きく異なるため。日本のトイレは清潔であることが多く、トイレトペーパーも無料で使用できる。
正解文	日本を訪れた中国人旅行客の多くが、日本の「トイレ」に驚くという。無料なことや、トイレトペーパーが設置されていることが評価されている。トイレという空間そのものが中国とは大きく異なるという。

表 12：テキスト平易化における ChatGPT の出力例

プロンプト	以下の英語を日本語に翻訳してください。 英語：数日留守にします。 日本語：
ChatGPT の出力	数日間家を空ける予定です。
BART の出力	何日か留守にします。
正解文	2、3 日の間、留守にします。