修 士 論 文 概 要 書

Summary of Master's Thesis

Date of submission: 01/25/2021

専攻名(専門分野) Department	情報理工・ 情報通信専攻	氏名 Name	京極 健悟	指 導 教 員 Advisor	渡辺 裕 ㊞
研究指導名 Research guidance	オーディオビジュアル 情報処理研究	学籍番号 Student ID number	5119F027-6 CD		
研究題目 Title	機械翻訳における同義文章生成モデルの検討 A Study on Synonymous Sentence Generation Model in Machine Translation				

1. まえがき

近年, 人工知能(AI)技術を利用した電子機器の開発が活発になっている. 特に自然言語処理の分野では, 人々の生活を支えている製品やサービスが増えている.

本研究では, 翻訳家による特色を機械翻訳により再現する. 翻訳データを学習データとして用い, 翻訳の特色を学習した文章生成モデルを作成することで. 任意の文章に対応したドメイン変換を取得できる. これにより, より自然な文章表現の生成を目指す.

2. 関連技術

2.1. T5

様々な自然言語処理タスクを同じモデルで解くための技術として, Text-to-Text Transfer Transformer (T5) [1]と呼ばれる汎用言語表現モデルが提案されている. T5 は, 自然言語処理タスクに対して汎用的に使用可能な分散表現を算出できる. つまり, T5 を用いることで入力文章を, 文脈を含んだ分散表現に変換できる. 以下の図 1 に T5 の概要図を示す.

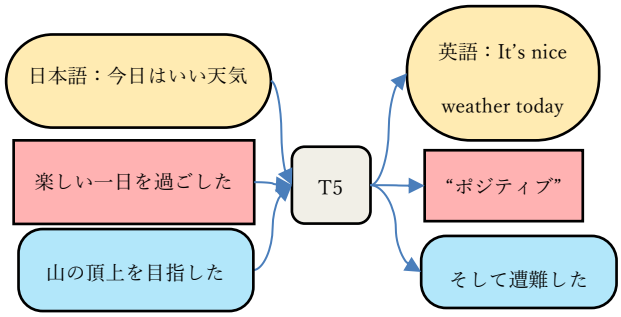

図 1 T5 の概要図

2.2. SentencePiece

文章をサブワードに分割するための技術として SentencePiece[2]と呼ばれるモデルが提案されている. サブワードとは, 学習したテキストに含まれる単語を文字や部分文字列に分解した単語である. SentencePiece の特徴として, 学習したデータセット中に高頻度で現れる単語はそのまま利用し, 低頻度で現れる単語はより細かく分割する. そのため, 使われたデータセットの中に未知語が存在しなくなる.

3. 提案手法

3.1. 提案手法概要

T5 は学習させるデータに基づいて, 作成されるモデルの精度に差が生じると考えられることから, T5 とデータ処理を組み合わせて翻訳家による特色を再現する手法を提案する. そこで, 3 種類の実験と評価を行う. まず, 題材データをそのまま T5 に学習させ, 文章を生成し評価する. これを実験 1 とする. 次に, 題材データの数を増やして T5 に学習させ, 文章を生成し評価する. これを実験 2 とする. 最後に, 題材データをデータの文字数の差に応じた処理を行い T5 に学習させ, 文章を生成し評価する. これを実験 3 とする. 最後に, 各 T5 モデルを BLEU スコアで評価する.

3.2. 評価指標 BLEU

生成された T5 モデルの精度を比較するため, 機械翻訳で用いられる評価手法の一つである Bilingual Evaluation Understudy (BLEU) を採用する. 参照した文章に近いほどスコアが高くなり, 0 から 100 の値でスコアが提示される. BLEU スコアは, 以下の式(1)で計算できる.

$$BLEU = BP \cdot \exp\left(\frac{1}{N}\sum_{n=1}^{N} logP_n\right) \qquad (1)$$

3.3. 実験データ

T5 モデルを作成する題材データとして, 漫画「とっておきスヌーピー」[3]を用いる. これは, 2-4 コマ漫画の総集編でできており, 英語原文と谷川俊太郎による日本語訳で構成されている. 英語原文を機械翻訳で日本語に翻訳したものを Encoder, 谷川俊太郎の日本語訳を Decoder に学習させる.

4. 実験結果

4.1. 実験 1 の結果

題材データ 2750 ペアを用いて T5 に学習を行った. 学習させた T5 モデルの Encoder に任意の文章を入力し, Decoder で文章の生成を行った.

表 1　実験 1 の結果

Encoder	Decoder1.584
おはよう	おはよう
こんにちは	ハイ
さようなら	にかくことだ
ボンク!	ゴチン!
今日はいい天気ですね	今日はいい天気だ
今日は祝日でした	今日は申命でした
明日は祝日でした	明日は申命でした
明日は何してますか？	明日はことなんだい？
明日の天気を教えて	明日の天気をさしてがた
好きな食べ物はなんですか？	歩兵は少なくとも食べ物さえうま?
歳はいくつですか？	歳はがとですかい？
好きな色を教えてください	指を鳴らせて言えばくれよ
それで全部です？	そうよ全部だい？

4.2. 実験 2 の結果

英語原文を機械翻訳で日本語, 英語, 日本語の順に変換を行い, 1 つの英語原文に対して 2 つの日本語訳を取得した．総データ 5500 ペアを用いて T５に学習し，文章の生成を行った.

表 2　実験 2 の結果

Encoder	Decoder
おはよう	おはよう
こんにちは	ハイ
さようなら	やっぱり
ボンク!	ゴチン!
今日はいい天気ですね	今日はいい天気ですね
今日は祝日でした	今日は嬉でした
明日は祝日でした	明日は嬉でした
明日は何してますか？	明日はありません？
明日の天気を教えてください	明日の天気を思うのさ
好きな食べ物はなんですか？	好きな食糧はいったかな?
歳はいくつですか？	歳はてのいかだい？
好きな色を教えてください	持っていた色を思うのさ
それで全部です？	やつせて全部だい？

4.3. 実験 3 の結果と考察

題材データペアの文字数の差が 14 以上あるデータを取り除き T5 に学習を行った. 取り除いたデータ量は 512 ペアであり，元データの約 81%を用いた．総データ 2238 ペアを用いて T５に学習し，文章の生成を行った.

表 3　実験 3 の結果

機械翻訳	谷川俊太郎訳
おはよう	おはよう
こんにちは	ハイ
さようなら	さあ，みんな
ボンク!	ゴチン!
今日はいい天気ですね	今日はいい天気だ
今日は祝日でした	今日は勝利でした
明日は祝日でした	明日は勝利でした
明日は何してますか？	明日は何っていうの？
明日の天気を教えてください	明日の天気を認識すべきよ
好きな食べ物はなんですか？	好きな食糧いいの?
歳はいくつですか？	歳はある？
好きな色を教えてください	好きな色を教えるよ
それで全部です？	それで全部だったか？

4.4. BLEU による評価および考察

各 T5 モデルに対する BLEU のスコアを以下の表 4 に示す.

表 4　各 T5 モデルに対する ＢＬＥＵ スコア

	実験１	実験２	実験３
BLEU	21.70	15.84	25.09

題材データのデータ数を機械翻訳により増やし, T5 に学習を行った実験２が BLEU スコアで一番低い値となった. また, 題材データペアの文字数の差に応じて不適切なデータを取り除き，T5 に学習を行った実験３が BLEU スコアで一番大きい値となった. これにより, 題材データのデータ数に関わらず, 適切なデータを学習に用いることで文章生成モデルの精度が上がることが確認できた.

5. むすび

本研究では，特定著者の作風を再現した同義文章を返答するモデルの作成を提案した．学習させるデータに処理を施し, 3 種類の実験を行った. 実験 1 では, 題材データに処理を施さずそのまま学習を行い, 文章の生成を行った. 実験２では, 題材データのデータ数を機械翻訳により増やし学習を行い，文章の生成を行った. 実験３では, 題材データペアの文字数の差に応じてデータに処理を施して学習を行い，文章の生成を行った．データ数を増やすために不適切なデータの数を増やした場合，生成される文章の質が下がることが確認できた．適切なデータを学習に用いることで文章生成モデルの精度が上がり，同義文章を返答するモデルを作成できることが確認できた.

参考文献

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", Oct 2019

[2] Taku Kudo, John Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing", Aug 2018v

[3] Charles.M.Schulz“とっておきスヌーピー”, 産経新聞社, vol. 1-7，Sept．2000

2020 年度
早稲田大学大学院基幹理工学研究科情報理工・情報通信専攻　修士論文

機械翻訳における同義文章生成モデルの検討

A Study on Synonymous Sentence Generation Model in Machine Translation

京極　健悟

(5119F027-6)

提出日：2021.1.25

指導教員：渡辺裕 印

研究指導名：オーディオビジュアル情報処理研究

目次

第1章　序論

1.1 研究背景

近年，人工知能(AI)技術を利用した電子機器の開発が活発になっている．特に自然言語処理の分野では，人々の生活を支えている製品やサービスが増えている．iOS や macOS などに内蔵されている Siri では，自然言語処理を用いて日常会話の受け答えやメッセージ送信などが可能であり，容易にアシスト機能を楽しめる．その他，テキストの自動翻訳や自動要約，情報検索や情報抽出など自然言語処理は人々の生活を支えている．

また，自然言語処理の中でも特に機械翻訳は精度が大幅に向上している．現在，Google 翻訳ではニューラル機械翻訳を使用している．大量の蓄積データをもとにニューラルネットワークを利用し，機械翻訳を実行している．しかし，翻訳家は言葉の言い回し，時代背景，読み手を意識した翻訳ができる．現状では，機械翻訳は翻訳家による翻訳結果を再現することができない．

そこで，本研究では，翻訳家による特色を機械翻訳により再現する．翻訳データを学習データとして用い，翻訳の特色を学習した文章生成モデルを作成することで．任意の文章に対応したドメイン変換を取得できる．これにより，より自然な文章表現の生成を目指す．

1.2 研究目的

入力文章の意味を理解した上で，文章の自動生成をすることができれば，高度な自動応答システムや，チャットボット，記事の自動作成などに利用することができる．また，入力文章の特色を再現できれば，任意の文章でもその特色を付与することが可能になる．そこで，本研究では，文章の意味を理解し，特定著者の作風を再現した同義文章を返答するモデルの作成を目的とする．

1.3 本論文の構成

本論文の構成を以下に示す．

第 1 章は本章であり，本論文の研究の背景および目的について述べる．

第 2 章では，文章生成モデル T5 の概要および本研究で用いる関連技術を説明する．

第 3 章では，翻訳家による特色を機械翻訳により再現し評価する手法を提案する．

第 4 章では，提案手法における実験および結果について述べる．

第 5 章では，本研究のまとめと今後の課題について述べる．

第2章　関連技術

2.1 まえがき

本章では，本研究で利用する関連技術である T5 と Transformer と SentencePiece について述べる．T5 は自然言語処理タスクを解く汎用言語表現モデルであり，Transformer は文章を分散表現に変換するニューラル機械翻訳モデルであり，SentencePiece は，日本語の文章をワブワードに分割するモデルである．

2.2 T5

Text-to-Text Transfer Transformer（T5）[1]は，Colin らによって開発された，汎用言語表現モデルである．転移学習を利用し，多くの自然言語処理ベンチマークで高いスコアを残している．T5 の特徴として，入力と出力にテキスト形式で学習を行うことで，様々な自然言語処理タスクを同じモデルで解くことができる．T5 は，自然言語処理タスクに対して汎用的に使用可能な分散表現を算出できる．つまり，T5 を用いることで入力文章を，文脈を含んだ分散表現に変換できる．以下の図 2.1 に T5 の概要図を示す．

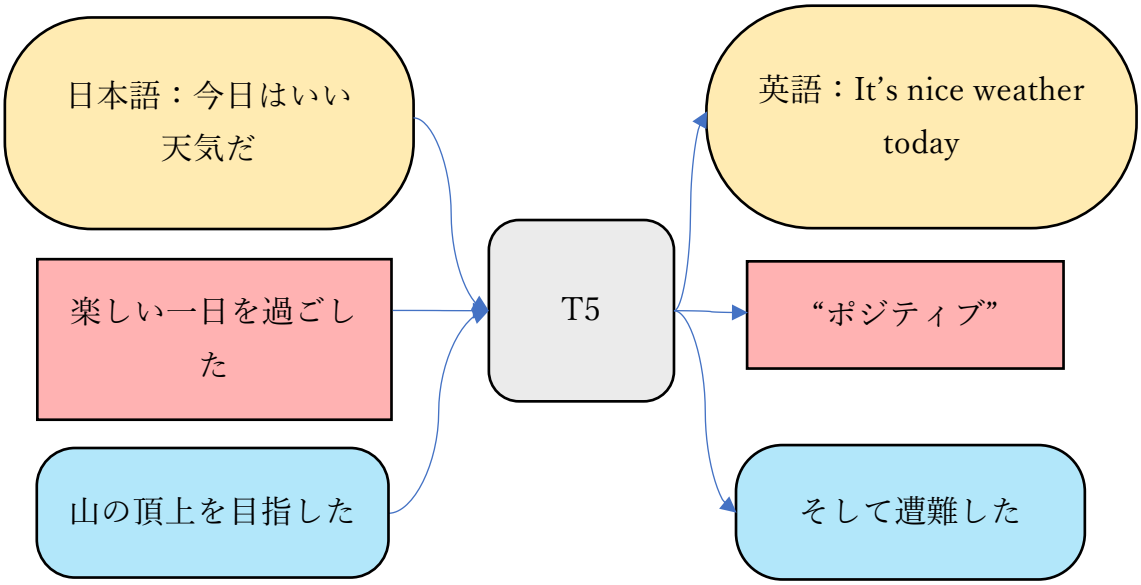

図 2.1　T 5 の概要図

T５は，Encoder-Decoder 型の Transformer から構成され，事前学習として様々な工夫がされている．

事前学習の目的関数として，ＢＥＲＴ[2]と同様 Masked language modeling が採用されている．これは，学習の際に入力文章の 15% を伏せ字にし，伏せ字の内 90% を特定の文字に置き換え，残り 10% をランダムな単語に置き換える．そして，置き換えた文章を元の文章に復元する事前学習である．Masked language modeling を用いることで，文中のマスクされた単語を推定し学習することができる．

事前学習を高速化する工夫として Random spans が採用されている．これは，Masked language modeling の学習時に，伏せ字の間隔をランダムにするために用いる工夫である．

また，T5 は事前学習のデータセットとして， Colossal Clean Crawled Corpus（Ｃ４）[3]を採用している．これは，世界中のウェブサーバーから集められたデータ Common Crawl のクエスチョンマーク・ビックリマーク・ピリオドで終わる文章のみを使用し，スラングや不適切な単語は取り除く，など様々なデータ処理を施した 745GB のデータセットである．

2.3 Transformer

Transformer[4]は，Attention 機構のみを使用したニューラル機械翻訳モデルである．これは，Recurrent Neural Network（RNN）や Convolutional Neural Network（CNN）を使用しないモデルである．Attention 機構は，文字列における単語の間に存在する文脈的な関係を学習する．そのため，文脈に沿った文章を学習することができる．従来の RNN や CNN を用いた自然言語処理は，データセットの学習に多大な時間がかかっていた．また，解析をする文章内に複数の意味を持つ単語が含まれているとき，その単語がどの意味で使われているのか理解できなかった．しかし，Attention 機構は文章全体を参照することで学習時間の短縮と，単語意味の適切な選択ができる．

また, Transformer は, Positional Encoding, Attention 機構, Feed Forward Neural Network で構成されている．Positional Encoding は，入力した単語に位置情報を付与する役割を持つ．Feed Forward Neural Network は，単一方向へのみ入出力が伝播するニューラルネットワークである. 以下の図 2.2 に Transformer の概要図を示す. なお, 図の縦左半分が Encoder, 右半分が Decoder である．

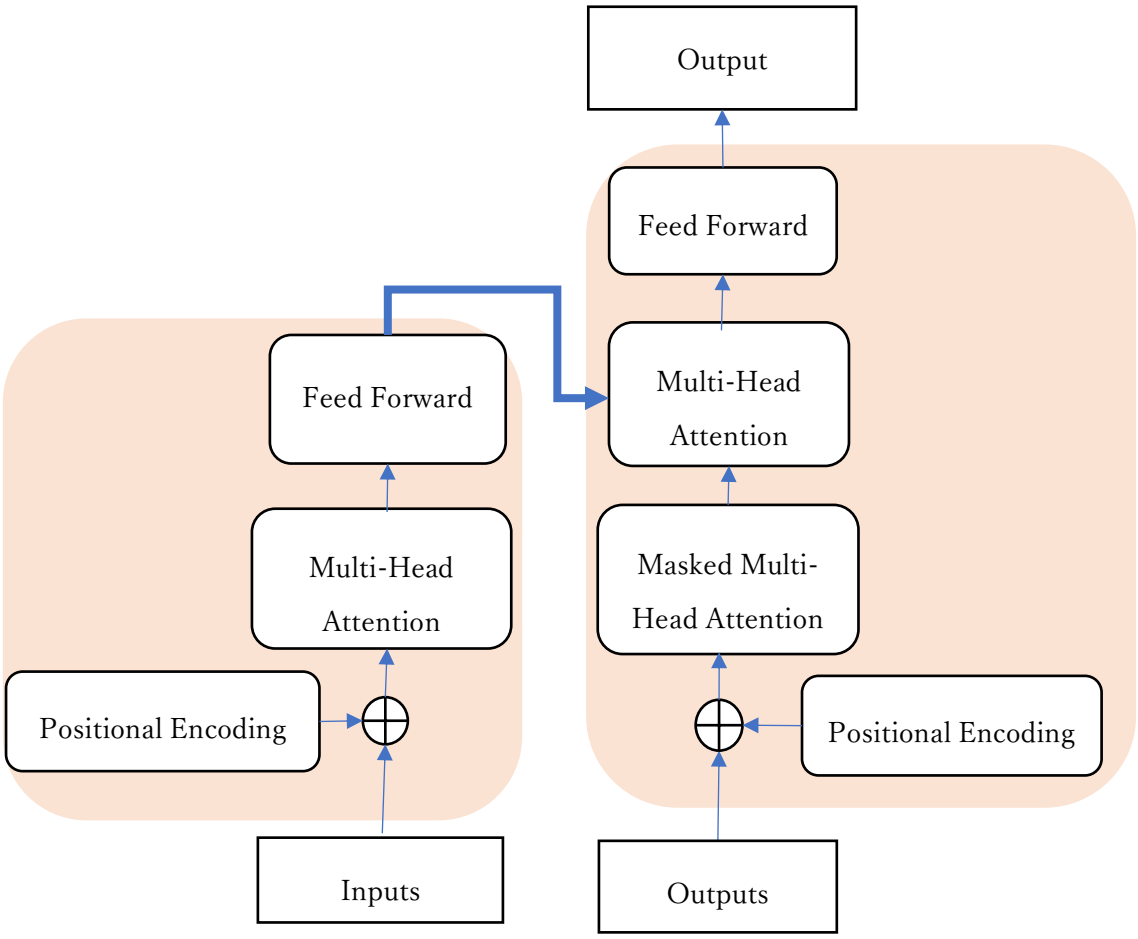

図 2.2　Transformer の概要図

2.4 SentencePiece

SentencePiece[5]は，工藤拓らによって開発された，文章をサブワードに分割するモデルである．サブワードとは，学習したテキストに含まれる単語を文字や部分文字列に分解した単語である．Mecab[6]や JUMAN++[7]など従来の形態素解析器と比べ，サブワードに分割するため，扱う語彙数を少なくすることが出来る．そのため，ニューラルネットワークへの入力の次元数が小さくなり学習時間を短くできる．

SentencePiece の特徴として，学習したデータセット中に高頻度で現れる単語はそのまま利用し，低頻度で現れる単語はより細かく分割する．そのため，使われたデータセットの中に未知語が存在しなくなる．SentencePiece による，文章をサブワードに分割した例を以下の表 2.1 に示す．

表 2.1　SentencePiece による文章の分割例

元の文	分割した文章
このアホ犬!	['この', 'ア', 'ホ', '犬', '!']
とてもいいよ	['とても', 'いい', 'よ']
待たせてごめんね	['待', 'た', 'せて', 'ご', 'めん', 'ね']
ネコぎらいなんです	['ネコ', 'ぎ', 'らい', 'な', 'んです']
みんなロータリーで何するんですか，先輩?	['みんな', 'ロータリー', 'で', '何', 'する', 'んです', 'か', ',', '先輩', '?']
おすわり!	['お', 'す', 'わり', '!']
学校での集まりにおくれてるんだ	['学校で', 'の', '集まり', 'にお', 'くれ', 'てる', 'んだ']
明日は何してますか？	['明日', 'は何', 'して', 'ます', 'か', '?']
明日の天気を教えてください	['明日', 'の', '天気', 'を教え', 'て', 'ください']
好きな食べ物はなんですか？	['好きな', '食べ物', 'は', 'なん', 'ですか', '?']
年はいくつですか？	['年', 'は', 'いく', 'つ', 'ですか', '?']
好きな色を教えてください	['好きな', '色', 'を教え', 'て', 'ください']
それで全部？	['それ', 'で', '全部', '?']
世界的に有名なホッケー選手が大試合にのぞむ	['世界的に', '有名な', 'ホッケー', '選手が', '大', '試合', 'に', 'の', 'ぞ', 'む']
山の頂上を目指した	['山の', '頂上', 'を目指した']

2.5 むすび

本章では，本研究を遂行するにあたり必要となる技術である T5 と Transformer と SentencePiece について述べた.

第3章　提案手法

3.1 まえがき

翻訳家による特色を機械翻訳により再現する目的で，T5 とデータ処理を組み合わせる手法を提案し，その具体的な内容について述べる．

3.2 提案手法概要

本研究の提案手法は，翻訳家による特色を機械翻訳により再現する目的で，T5 とデータ処理を組み合わせる手法である．また，T5 は学習させるデータに基づいて，作成されるモデルの内容に差が生じる．そこで，翻訳家による特色を再現するため，T5 の学習時に入力には英語原文を機械翻訳で日本語に翻訳した文章を用い，出力には谷川俊太郎訳を用いる．また，精度の良いモデルを生成するため，題材データに 2 種類のデータ処理を行う．また，T 5 は多くの自然言語処理タスクで State of The Art(SOTA)を達成している．そのため，学習効率の良い同義文章生成モデルの作成ができる．提案手法の概要図を以下の図 3.1 に示す．

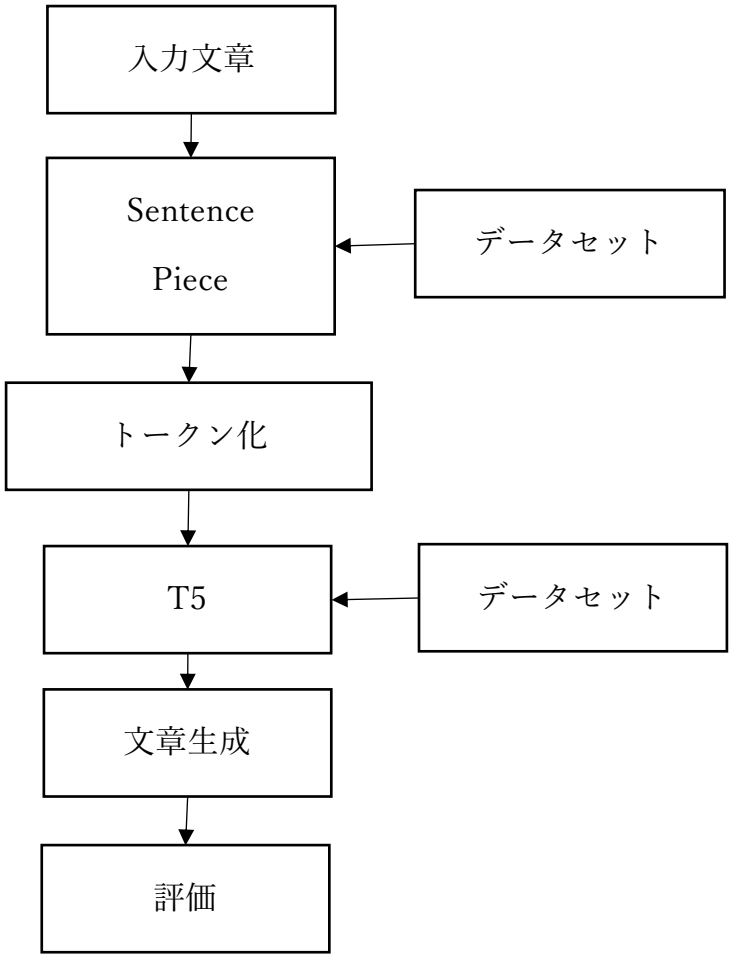

図 3.1 提案手法の概要図

まず，SentencePiece を用いて，入力文章をサブワードに分割し，トークン化する．日本語の文章は分かち書きされていないので，機械にデータを読み込ませるために分割し，トークン化する必要がある．トークン化するため，日本語の文章をあらかじめ SentencePiece に学習させる必要がある．本研究では，学習させる日本語のデータとして，日本語の Wikipedia を用いる．次に，分割し，トークン化された文章を T5 に学習させる．T5 は，事前学習を行うことで日本語の文章に対応した自然言語処理タスクを解くことが出来る．事前学習の日本語データとして，本研究では SentencePiece 同様に，日本語の Wikipedia を用いる．事前学習を行うことで，日本語の自然言語処理タスクに対して汎用的に使用可能な分散表現を取得するモデルが作成できる．次に，学習させた T5 モデルを用いて文章生成を行う．学習済みの T5 モデルの Encoder に任意の文章を入力することで，学習に用いた入力文章の特色を再現した文章を生成できる．最後に，生成された文章を評価指標 BLEU で評価を行う．

3.3 SentencePiece の事前学習データ

入力文章をトークン化するため, 日本語の文章をあらかじめ SentencePiece に学習させる必要がある．学習には，Tensorflow Datasets の公開データセットである日本語版 wikipedia2020[8]を使用する．なお，データサイズは 5.61GiB である．データセットの例を以下の図 3.2 に示す.

本項は，都市対抗野球大会における青森県勢の戦績についてまとめたものである.
概略
青森県は東北地区に属しているが, 宮城, 秋田などの強豪チームの前に東北予選で屈することが多く，本大会に出場したのは第 13 回大会（1939 年）の青森林友 1 チームだけである
その青森林友も本大会 1 回戦で完封負けを喫しており，青森県勢は本大会で 1 勝はおろか,
1 点も挙げていない.
通算成績
（第 89 回大会まで．中止となった第 15 回大会を除く．以下本項において同じ）
延べ出場回数　1 回
優勝回数　なし
準優勝回数　なし
通算勝敗　0 勝 1 敗（勝率 .000）
出場チームの戦績
他県勢との対戦成績
都市別の他都市との対戦成績
（都市名は，最後に対戦した時点での名称を記す.）
青森市
関連項目
都市対抗野球大会

図 3.2　日本語版 wikipedia2020 の例

3.4 T5 の事前学習データ

日本語の文章に対応した自然言語処理タスクを解くため，T5 に日本語の文章を学習させる必要がある．学習には，SentencePiece 同様に Tensorflow Datasets の公開データセットである日本語版 wikipedia2020 を使用する.

3.5 評価指標

本項では，本研究で用いる評価指標について述べる．提案手法によって生成された文章を評価指標 Bilingual Evaluation Understudy（BLEU）で評価する．BLEU は，機械翻訳における評価方法の一つである．参照した文章に近いほどスコアが高くなり，0 から 100 の値でスコアが提示される．BLEU スコアは，以下の式(3.1)で計算できる．

$$BLEU = BP \cdot \exp\left(\frac{1}{N}\sum_{n=1}^{N} logP_n\right) \tag{3.1}$$

なお，条件式は以下の式(3.2)，式(3.3)である．

$$P_n = \frac{\sum_i 翻訳文i と参照訳i で一致したn-gram数}{\sum_i 翻訳文i 中の全n-gram数} \tag{3.2}$$

$$BP = \min\left(1, \exp\left(1 - \frac{参照訳の長さに最も近い翻訳文の長さ}{翻訳文の長さ}\right)\right) \tag{3.3}$$

ただし，n-gram とは，任意の文章における連続した n 個の単語や文字のまとまりを表している．P_nは，翻訳文中における全ての n-gram に対する参照訳で一致した n-gram の割合である．これにより翻訳文と参照訳の連続したフレーズが比較され，類似度を参照できる．また，一般的に自然言語処理のタスクを評価する際は，n=4 が用いられる場合が多い．

BP は，翻訳文の長さが参照訳より短い場合に 1 より小さくなり，翻訳文が参照訳より長い場合に 1 となるペナルティである．翻訳文が参照訳より短い場合は，機械翻訳の精度が悪いとみなされ，BLEU のスコアは低くなる．また，BLEU スコアの大まかな評価指標として，以下の表 3.1 に指標目安を示す．

表 3.1　ＢＬＥＵスコアの指標目安

スコア	指標目安
１０より下	ほとんど役に立たない
１０～１９	主旨を理解するのが困難である
２０～２９	主旨は明白であるが，文法上の重大なエラーがある
３０～３９	理解できる，適度な品質の翻訳
４０～４９	高品質な翻訳
５０～５９	非常に高品質で，適切かつ流暢な翻訳
60 より上	人が翻訳した場合よりも高品質であることが多い

3.6　むすび

本章では，翻訳家による特色を機械翻訳により再現する目的で，T5 とデータ処理を組み合わせる手法を提案し，その具体的な内容について示した．

第4章　評価実験，結果および考察

4.1 まえがき

本章では，T5 モデルに学習させる題材データに処理を施し，3 種類の実験結果及び実験の評価と考察を述べる．

4.2 実験概要

本研究では，T5 モデルに学習させる題材データに処理を施し，3 種類の実験と評価を行う．まず，題材データをそのまま T5 に学習させ，文章を生成し評価する．これを実験 1 とする．次に，題材データの数を増やして T 5 に学習させ，文章を生成し評価する．これを実験 2 とする．最後に，題材データをデータの文字数の差に応じた処理を行い T 5 に学習させ，文章を生成し評価する．これを実験 3 とする．

4.2.1 データセット

本研究では, T5 のモデルを作成する題材データとして, 漫画「とっておきスヌーピー」[9]を用いる．これは，2-4 コマ漫画の総集編でできており，英語原文と谷川俊太郎による日本語訳で構成されている. 英語原文を機械翻訳で日本語に翻訳した文章を T5 の Encoder に学習させる. 谷川俊太郎による日本語訳を Decoder に学習させる. なお, 総データ数は 2750 ペア, 最小の文字数は 1 文字, 最大の文字数は 103 文字, 平均文字数は約 27.5 文字である．また，題材データの文字数は，30 文字以内の文章が 70%を占めている．題材データの例を表 4.1 に示す．また，題材データの文字数の分布図を以下の図 4.1 に示す．

表 4.1　題材データの例

機械翻訳	谷川俊太郎訳
もちろん	もちろんさ
馬鹿犬！	このアホ犬!
非常に素晴らしい	とてもいいよ
世界は太陽の周りを公転しますか？	世界は太陽のまわりを回ってるんですって？
お待たせしました	待たせてごめんね
彼は猫が嫌い	ネコぎらいなんです
傘は役立っているでしょう！	傘一本で助かったんだけど!
私は私のロッカーの組み合わせを思い出すことができません	ロッカーの鍵の組み合わせが思い出せないわ
はい，奥様私の名前はビッグベンであります	はい先生，ぼくビッグベンです
あなたは今まで前に電話で話をしていますか？	これまで電話かけたことあるのかい
そして，ミッキーマウスは，非常に裕福なのですか？	そしてミッキー·マウスは大金持ちだろ?
床屋のポールにサインするために言って！我々はそれを売ることができます！	床屋のサインポールにサインするように言えよ! 売れるぞ!
アンディとオラフはまだ砂漠に住んでいる私たちの兄弟のスパイクを発見していませんでした.	アンディとオラフは荒れ地に住む兄弟のスパイクを依然として見つけていなかった.
泥を通って行き，沼地に左スプラッシュ，沼やゴールラインに，川底を過ぎて，湿地を通過	ぬかるみを突破して，水しぶきあげて左沼から湿地に突っこみ，川底を過ぎ，泥んこからゴールラインへ

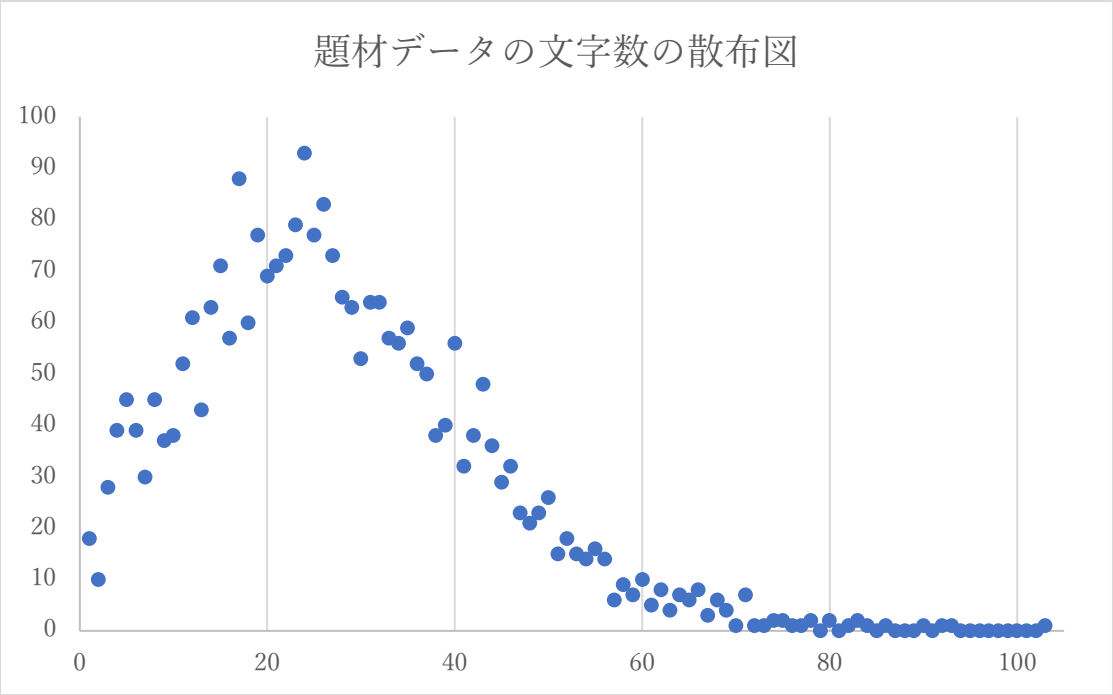

図 4.1　題材データの文字数の分布図

4.3 実験 1

題材データ 2750 ペアを用いて T 5 に学習を行った. 学習させた T 5 モデルの Encoder に任意の文章を入力し，Decoder で文章の生成を行った．その結果を以下の表 4.2 に示す.

表 4.2　実験 1 の結果

Encoder	Decoder
おはよう	おはよう
こんにちは	ハイ
さようなら	やっぱり
ボンク!	ゴチン!
今日はいい天気ですね	今日はいい天気ですね
今日は祝日でした	今日は嬉でした
明日は祝日でした	明日は嬉でした
明日は何してますか？	明日はありません？
明日の天気を教えてください	明日の天気を思うのさ
好きな食べ物はなんですか？	好きな食糧はいったかな?
歳はいくつですか？	歳はてのいかだい？
好きな色を教えてください	持っていた色を思うのさ
それで全部です？	やつせて全部だい？

「おはよう」「今日」「明日」「天気」「何」「歳」などのデータ数が多く，題材データ内で同じ表現がされている単語は，Decoder 側でも同じ単語で再現されている．しかし，類似した挨拶「さようなら」はデータ数が少なく，谷川俊太郎訳では様々な表現として使われているため, 不適切な同義文章の変換が行われた. また, 題材データ内に含まれていない「祝日」は意味の異なる単語 1 文字に置き換わっている．題材データ内で多様に使われている「こんにちは」「ボンク」はそれぞれ「ハイ」「ゴチン」と置き換わっている. 題材データ内では, クエスチョンマークが多く使われているため, 文末に着目すると「の？」「かな？」「だい？」など様々な同義文章の変換が行われた.

T5 の事前学習に大規模な Wikipedia のデータを用いた．そのため，Decoder で生成された文章の構造は適切であることが確認できる. また, 助詞と名詞などをしっかりと区別し文章が生成されたことが確認できる．これは，T5 は事前学習を行うことで，文脈を含んだ文

章の分散表現を取得しているため，文章の構造を適切に把握していると考えられる．しかし，題材データで使われていない未知語や，様々な表現方法がされている単語に対しては，不適切な同義文章の変換が行われた．これは，機械翻訳と谷川俊太郎訳で文章表現の差が大きくなり，T5 モデルが不適切な同義表現を学習したためと考えられる．

4.4 実験 2

題材データのデータ数を機械翻訳により増やし，T5 に学習を行った．題材データはデータ数が限られている．そこで，英語原文を機械翻訳で日本語，英語，日本語の順に変換を行い，1 つの英語原文に対して 2 つの日本語訳を取得した．全ての英語原文に対して行い，総データ数は 5500 となった．以下の表 4.3 に，実験 2 で使用したデータの例を示す．

表 4.3　実験 2 データセットの例

機械翻訳	谷川俊太郎訳
もちろん	もちろんさ
馬鹿犬！	このアホ犬!
優秀な	とてもいいよ
世界は太陽を中心に回転していますか？	世界は太陽のまわりを回ってるんですって？
すみません，待たせました	待たせてごめんね
彼は猫が嫌いです	ネコぎらいなんです
傘は便利です！	傘一本で助かったんだけど!
はい，妻私の名前はビッグベンです	はい先生，ぼくビッグベンです
これまでに電話で話したことがありますか.	これまで電話かけたことあるのかい
そして，ミッキーマウスはとても裕福ですか？	そしてミッキー･マウスは大金持ちだろ?
理髪店の看板に署名すると言ってください！売れます！	床屋のサインポールにサインするように言えよ! 売れるぞ!
アンディとオラフは，砂漠に住む私たちの兄弟であるスパイクをまだ発見していません.	アンディとオラフは荒れ地に住む兄弟のスパイクを依然として見つけていなかった.
泥を通り抜け，沼にスプラッシュを残し，沼またはゴールラインに行き，川床を通り過ぎ，沼を通り過ぎます..	ぬかるみを突破して，水しぶきあげて左沼から湿地に突っこみ，川底を過ぎ，泥んこからゴールラインへ

表 4.3 のデータセットを用い，T5 に学習を行った．学習させた T 5 モデルの Encoder に任意の文章を入力し，Decoder で文章の生成を行った．その結果を以下の表 4.4 に示す．

表 4.4　実験 2 の結果

Encoder	Decoder
おはよう	おはよう
こんにちは	ハイ
さようなら	にかくことだ
ボンク!	ゴチン!
今日はいい天気ですね	今日はいい天気だ
今日は祝日でした	今日は申命でした
明日は祝日でした	明日は申命でした
明日は何してますか？	明日はことなんだい？
明日の天気を教えて	明日の天気をさしてがた
好きな食べ物はなんですか？	歩兵は少なくとも食べ物さえうま?
歳はいくつですか？	歳はがとですかい？
好きな色を教えてください	指を鳴らせて言えばくれよ
それで全部です？	そうよ全部だい？

実験 1 同様,「おはよう」「今日」「明日」「天気」「何」「歳」などのデータ数が多く，題材データ内で同じ表現がされている単語は，Decoder 側でも同じ単語で再現されている．また，題材データ内に含まれていない「祝日」は意味の異なる単語 2 文字に置き換わっている．題材データ内で多様に使われている「こんにちは」「ボンク」もそれぞれ「ハイ」「ゴチン」と置き換わっている. 題材データ内では，クエスチョンマークが多く使われているため，文末に着目すると「かい？」だい？」など様々な同義文章の変換が行われた．しかし，「好きな食べ物はなんですか？」「歳はいくつですか？」「好きな色を教えてください」はそれぞれ「歳」「食べ物」以外の文章に不適切な同義文章の変換が行われた.

実験 1 と比べて，データ数が少なく谷川俊太郎訳で様々な表現方法がされている単語に対して，不適切な同義文章表現の変換が行われたと考えられる．これは，データ処理を施していないデータセットを増やしたことで，機械翻訳と谷川俊太郎訳で文章表現の差がより大きくなり，T5 モデルが不適切な同義表現を学習したためと考えられる．

4.5 実験 3

題材データペアの文字数の差に応じてデータ処理を施し，T5 に学習を行った．Encoder に入力する文章は機械翻訳を行なっているため，谷川俊太郎訳の傾向にそぐわない表現や遠回しな表現など学習に不適切なデータが存在する．そのようなデータをある一定の閾値により取り除き，T 5 に学習を行った．本実験では，文字数の差が 14 以上あるデータを取り除き T5 に学習を行った．取り除いたデータ量は 512 ペアであり，元データの約 81%を実験 3 のデータとして用いる．以下の図 4.2 に題材データの文字数の差の散布図を示す．また，データ処理を施し学習を行った T 5 モデルの Encoder に任意の文章を入力し，Decoder で文章の生成を行った．その結果を以下の表 4.5 に示す．

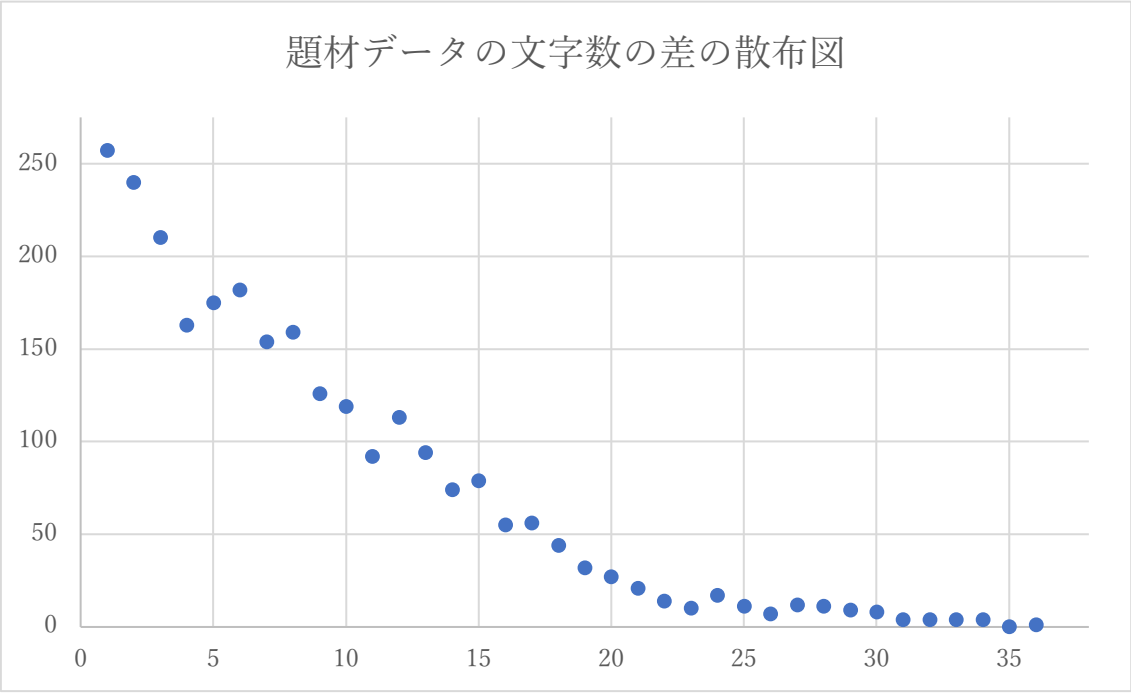

図 4.2 題材データの文字数の差の散布図

表 4.5　実験 3 の結果

機械翻訳	谷川俊太郎訳
おはよう	おはよう
こんにちは	ハイ
さようなら	さあ，みんな
ボンク!	ゴチン!
今日はいい天気ですね	今日はいい天気だ
今日は祝日でした	今日は勝利でした
明日は祝日でした	明日は勝利でした
明日は何してますか？	明日は何っていうの？
明日の天気を教えてください	明日の天気を認識すべきよ
好きな食べ物はなんですか？	好きな食糧いいの?
歳はいくつですか？	歳はある？
好きな色を教えてください	好きな色を教えるよ
それで全部です？	それで全部だったか？

実験 1，実験 2 と同様に，「おはよう」「今日」「明日」「天気」「何」「歳」などのデータ数が多く，題材データ内で同じ表現がされている単語は，Decoder 側でも同じ単語で再現されている．また，題材データ内に含まれていない「祝日」は意味の異なる単語 2 文字に置き換わっている．題材データ内で多様に使われている「こんにちは」「ボンク」もそれぞれ「ハイ」「ゴチン」と置き換わっている．題材データ内では，クエスチョンマークが多く使われているため，文末に着目すると「の？」「ある？」「だったか？」など様々な同義文章の変換が行われた．

実験 1，実験 2 と比べ，不適切な同義文章の変換が行われていない．これは，題材データペアの文字数の差に応じてデータ処理を施したことで，機械翻訳と谷川俊太郎訳で文章表現の差が小さくなり，T5 モデルが適切な同義表現を学習したためと考えられる．

4.6 BLEU による評価および考察

4.3，4.4，4.5 節で学習し作成した各 T5 モデルに対する BLEU のスコアを以下の表 4.6 に示す.

表 4.6　各 T5 モデルに対する ＢＬＥＵ スコア

	実験 1	実験 2	実験 3
BLEU	21.70	15.84	25.09

題材データのデータ数を機械翻訳により増やし，T5 に学習を行った実験 2 が BLEU スコアで一番低い値となった．これにより，題材データのデータ数は一番多いが，不適切なデータの数を増やした場合，文章生成モデルの精度が下がることが確認できた．また，題材データペアの文字数の差に応じて不適切なデータを取り除き，T5 に学習を行った実験 3 が BLEU スコアで一番大きい値となった．これにより，題材データのデータ数は一番少ないが，適切なデータを学習に用いることで文章生成モデルの精度が上がることが確認できた.

4.7 むすび

本章では，T5 モデルに学習させる題材データに処理を施し，3 種類の実験結果及び実験の評価と考察を述べた.

第5章　結論

5.1 結論

本研究では，特定著者の作風を再現した同義文章を返答するモデルの作成を提案した．学習させる題材データに処理を施すことで，3 種類の実験を行った．実験 1 では，題材データに処理を施さず学習を行い，文章の生成を行った．実験 2 では，題材データのデータ数を機械翻訳により増やし T5 に学習を行い，文章の生成を行った．実験 3 では，題材データペアの文字数の差に応じてデータに処理を施し，学習を行い，文章の生成を行った．データ数を増やすために不適切なデータの数を増やした場合，生成される文章の質が下がることが確認できた．適切なデータを学習に用いることで文章生成モデルの精度が上がり，同義文章を返答するモデルを作成できることが確認できた．

5.2 今後の課題

本研究での課題点として，事前学習を行う際の学習率やデータ処理の最適化が挙げられる．提案手法では，題材データペアの文字数の差に応じて元データの 80%を使用するようにデータの削除を行ったが，他の手法についても検討する必要がある．

また，機械翻訳における評価指標の一つである BLEU を用いたが，同義文章を返答するモデルに対する他の評価指標についても検討する必要がある．

謝辞

本研究の実験環境を与えてくださり，適切な指導を賜った渡辺裕教授に深く感謝を申し上げます．また，日頃から相談や問題解決を下さった研究室の皆様に御礼申し上げます．最後に，これまで暖かく見守ってくれた家族に感謝いたします．

参考文献

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", Oct 2019

[2] Jacob Devin, Ming-Wei Chang Kentin Lee, Kristina Toutanova, ”BERT: Pre-training of Deep Bidrectional Transformers for Language Understanding”,NAACL,pp.4171-4186,2019”

[3] TensorFlow Datasets C4<https://www.tensorflow.org/datasets/catalog/c4>参照 2021.1.20

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention Is All You Need", Dec 2017

[5] Taku Kudo, John Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing", Aug 2018v

[6] 京都大学情報学研究所，京都大学情報学研究所ホームページ:”MeCab: Yet Another Part-of-Speech and Morphological Analyzer”, <http://taku910.github.io/mecab/>, 参照 2021.1.20.

[7] 京都大学大学院情報学研究所，黒橋・河原研究室ホームページ:”日本語形態素解析システム JUMAN++”,<http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++> 参照 2021.1.20.

[8] TensorFlow Datasets Wikipedia <https://www.tensorflow.org/datasets/catalog/wikipedia>参照 2021.1.20

[9] Charles.M.Schulz“とっておきスヌーピー”，産経新聞社, vol. 1-7，Sept. 2000

[10] オブジェクトの広場，技術部アドバンステクノロジーセンター，鵜野 和也”はじめての自然言語処理”< https://www.ogis-ri.co.jp/otc/hiroba/technical/similar-document-search/part7.html#13>参照 2021.1.20.

図一覧

表一覧