卒 業 論 文 概 要 書

Summary of Bachelor's Thesis

Date of submission: 02/06/2019

学科名 Department	情報通信	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	1w152113-1 CD		
研究題目 Title	機械翻訳におけるドメイン変換に関する研究 Research on Domain Transfer 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. 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.

2. 文章生成モデル

文章生成モデル Sequence to Sequence (Seq2seq) [1] は, Long short-term memory (LSTM) などの Recurrent Neural Network (RNN) を用いて文章の意味を理解する手法であり, 図 1 のように Encoder と Decoder で構成されたネットワークからなる.

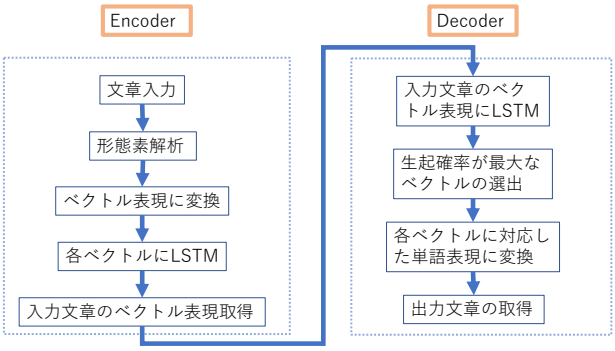

図 1 Seq2seq の概要図

3. 提案手法

Seq2seq 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」[2]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.

ここで, 題材データの量の少なさや精度を補うため, ファインチューニングを適用する. ファインチューニングは, 学習済みモデルの重みを初期値として利用し, 新しいモデルに受け渡す手法である. 学習済みモデルの結果を, 求めたいモデルに反映させることができるため, 少ないデータ量でも質の良いモデルを作成できる.

ファインチューニングを行うため, 学習済みモデルの作成を行う. 精度のいいモデルを作成するために多くの同義文章のペアを学習させる必要があるので, 同義文章のペアとして, やさしい日本語コーパス[3]を用いた. これは, 50000 文をやさしい日本語（平易な日本語表現）に書き換えた対訳コーパスである.

4. 実験結果

やさしい日本語コーパスを用いて作成した学習済みモデルのドメイン変換結果を表 1 に示す

表 1 学習済みモデルによるドメイン変換

Encoder	Decoder
おはよう	みんなにおはようおはよう物に話した.
こんにちは	こんにちは品がなくなる思いが？
さようなら	さようなら保証さようなら.
今日はいい天気ですね	今日はいい天気ですね.
明日はいい天気ですね	明日はいい天気ですね.
今日は祝日でした	今日はへでした
明日は祝日でした	明日はへでした.
明日は何してますか？	明日は何していいですか？
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？	年も年はどのですか？
好きな色を教えてください	好きな色を教えてください.
それで全部です？	それでそれで大丈夫です？
ボンク!	ゲームのような態度を聞いている.
素敵な？	素敵な？

学習済みモデルの重みを利用し，題材データを再学習したファインチューイングモデルのドメイン変換結果を表 2 に示す.

表 2 ファインチューニングモデルによるドメイン変換

Encoder	Decoder
おはよう	だささドアさだよ
こんにちは	学校これハーイをが入っ..
さようなら	痛いいく
今日はいい天気ですね	今日は.もっ
明日はいい天気ですね	もっともひとつさ
今日は祝日でした	早く，チャック_
明日は祝日でした	もっやってるも
明日は何してますか？	もしで！ママ分かるドルはしてる？
明日の天気を教えてください	もっとでいいのがもらわしかのがある⁉
好きな食べ物はなんですか？	彼，いつもたらなんて？
年はいくつですか？	“ら!だったかい？
好きな色を教えてください	彼が練習に行ったな気に入ったいことがぼく
それで全部です？	それだけ，50 のは？
ボンク!	もっが番犬!!
素敵な？	これ犬小屋みんな

ファインチューニングモデルでは，ドメイン変換を取得できず, 文章として不成立であった. 原因として, データの下処理の不十分さや, 学習済みモデルの重みが効果的に反映されてないことが挙げられる. 学習済みモデルの重みを初期値として用いるが，学習率を高く設定していたために題材データの傾向を多く捉えてしまったと考えられる. そのため, 記号の出現や文の関連性を見出し難い結果となり，学習を進めることでモデルの精度が悪くなったと考えられる.

5. まとめ

本研究では，Seq2seq とファインチューニングを用いて，機械翻訳におけるドメイン変換を行った. Seq2seq のモデルでは, 文章を入力することで同義文章を返答する学習済みモデル，ファインチューニングモデルを作成した. ファインチューニングモデルは, 学習済みモデルの重みを利用し, 題材データを再学習したモデルである. 学習済みモデルでは, 高精度で同義文章を取得できたが, 事前モデル, ファインチューニングモデル共に同義文章を取得できなかった.

ファインチューニングを行う際の学習率の最適化と, データの下処理を充分に行うこと, 学習難易度が低い文章の選出を行うことで，モデルの精度向上が期待できる.

参考文献

[1] I. Sutskever, O. Vinyals, and Q. V. Le: “Sequence to Sequence Learning with Neural Networks,” Neural Information Processing Systems (NIPS), pp1-9, Dec. 2014

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

[3] SNOW T15: やさしい日本語コーパス http://www.jnlp.org/SNOW/T15, (平成 31 年(2019 年)1 月現在)

2018 年度　卒業論文

機械翻訳におけるドメイン変換に関する研究

Research on Domain Transfer in Machine Translation

指導教員　渡辺　裕　教授

早稲田大学　基幹理工学部

情報通信学科

1w152113-1

京極　健悟

目次

第 1 章　序論

1.1 研究の背景

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 研究の目的

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 関連研究

文章を自動生成する関連研究として，緒方ら[1]は元となる文章の切り貼りを行う事で，140 字程度の文章を作成する手法を提案している．この手法では，もととなる文章の従属節を別の従属節で置換し，他の小説の会話文と元の会話文を組み替える事で文

章を自動生成している．しかし，この手法は文法として正しい文章を生成しても，意味としては不当な文章を生成する可能性がある．また，文章の意味を理解せず切り貼りで生成しているため，実用性に欠けるという問題点がある．

1.4 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，関連研究について述べる．

第 2 章では， 文章生成モデル Seq2seq の概要および本研究における Seq2seq の利点と課題について述べる．

第 3 章では，ドメイン変換された同義な文章を得る手法について述べる．

第 4 章では，手法の実験結果及び考察について述べる．

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　文章生成モデル

2.1 まえがき

本章では，本論文で用いる文章生成モデルである Sequence to Sequence (Seq2seq) [2]について述べる. Seq2seq は Sutskever らによって考案された文章の意味を理解するために用いられる手法であり，機械翻訳の性能向上やチャットボット，文章要約などで利用されている．

2.2 文章生成モデル Seq2seq の概要

文章生成モデル Seq2seq は，Long short-term memory (LSTM) などの Recurrent Neural Network (RNN) を用いて文章の意味を理解する手法であり，図 2.1 のように Encoder と Decoder で構成されたネットワークからなる．Seq2seq の Encoder では，中国語や日本語など分かち書きをしない文章に対して形態素解析を行い，言葉が意味を持つ最小の単位まで分割する．次に，単語に応じてベクトル変換を行う．さらに，各ベクトル表現に関して LSTM を時系列順に行うことで，Encoder は入力文章のベクトル表現を得る．なおベクトル表現を得る際, 意味が類似している単語のベクトル距離を近くすることで, 単語間の関係性をベクトル演算によって表現できる

Decoder 側では，入力文章のベクトル表現に LSTM を適用し，生起確率が最も高い単語ベクトルを選出する. 次に, ベクトル表現を単語に戻す. この操作を繰り返すことで Decoder 側の出力結果を取得する. 学習時には，予測したベクトル表現と教師データを比較することで誤差を伝播させる.

任意の文章をドメイン変換する場合，Encoder では図 2.2，Decoder では図 2.3 のように動作処理を行う.

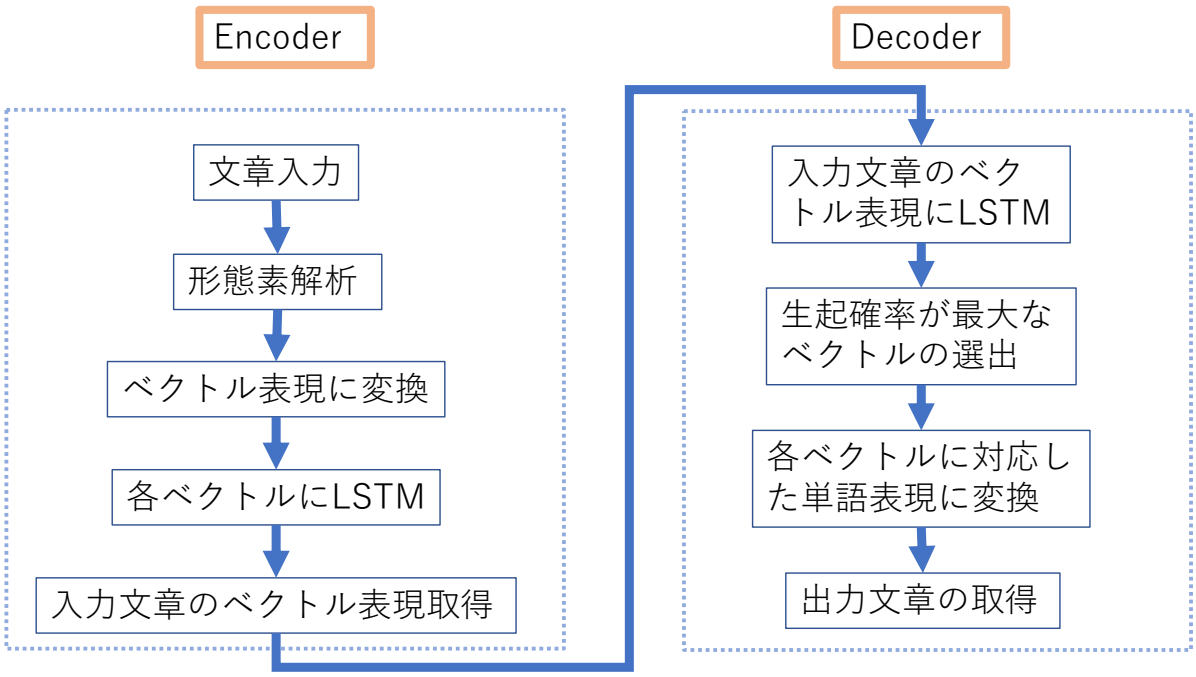

図 2.1 Seq2seq の概要図

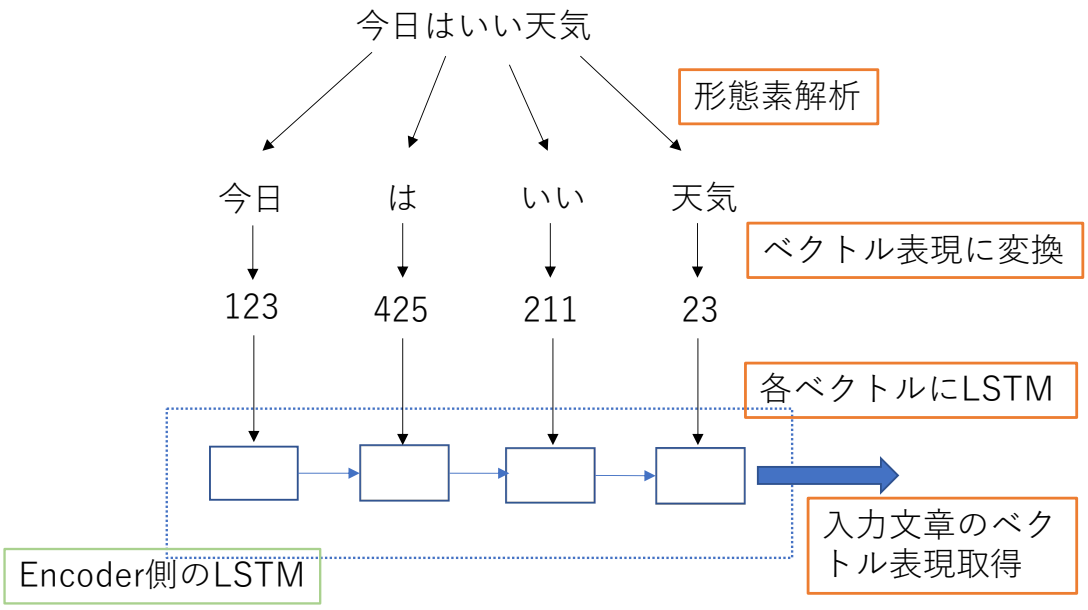

図 2.2 Encoder における動作処理の例

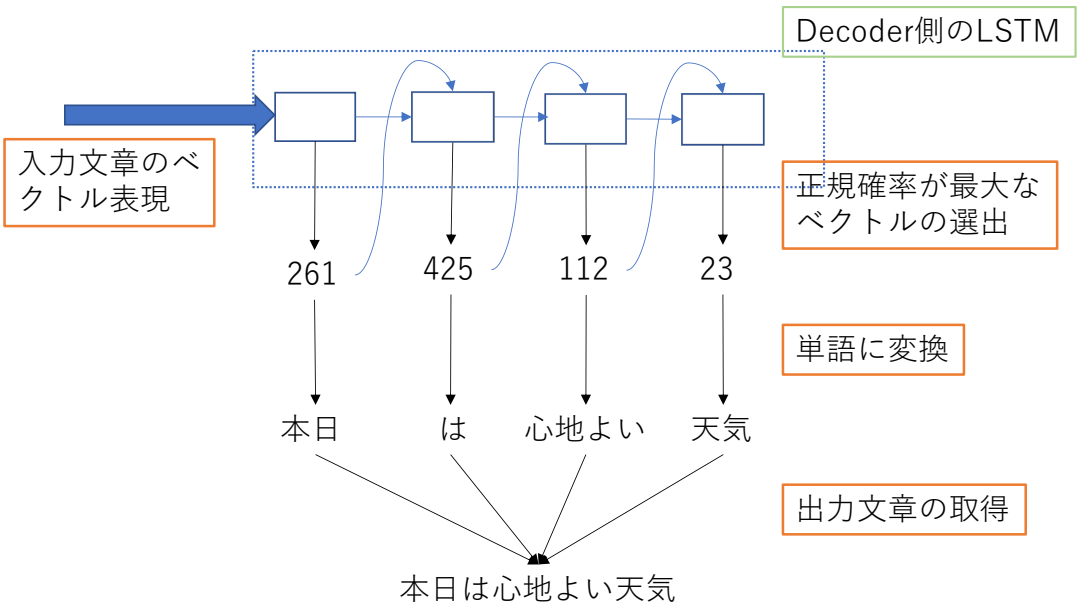

図 2.1 Decoder における動作処理の例

2.3 利点と課題

Seq2seq は，ある文章を異なった文章に変換する機構とみなすことができ，学習させる文章ペアがあれば，その特徴を再現することができる．特徴を再現する幅は広く，同義文章や対義文章だけでなく，言語間翻訳や応対システムなども再現可能である．しかし，その特徴を学習させるためには，学習精度の良い多くのデータが必要となる．データ数が少ないと特徴を掴み取れず，学習精度が悪いと文法として正しい文章生成ができなくなる．データの学習精度を良くするためには様々な方法がある．文章を短くすることや，指示語を削除すること，カタカナ表記と平仮名表記を統一すること，無駄な記号や句読点を削除することなどが挙げられる．

したがって，求めるモデルを作成するためには適切なデータセットを多量に必要とする．

第 3 章　提案手法

3.1 まえがき

本章では，Seq2seq を用いて，入力文章と同義なドメイン変換された文章を得る手法を検討する．

3.2 題材データ

Seq2seq 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」[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．Encoder で入力した機械翻訳が，Decoder で谷川俊太郎風にドメイン変換されることを目標とする．なお，総データ数は，2750 ペアである．題材データの例を表 3.1 に示す．

表 3.1 題材のデータの例

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
彼らは回転式で何をしますか，先生？	みんなロータリーで何するんですか，先輩?
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
座って！	おすわり!
私は学校での会議に遅刻です	学校での集まりにおくれてるんだ

3.3 ファインチューニング

題材データは一文が長く，短い文章を選別する場合ではデータ量が不十分である．また，題材データが人の認識では同義文章となる場合であっても，生成モデルが単語間の関係性を理解し，その特色を再現させるためには大量のデータが必要となる．したがって，題材データの量の少なさや精度を補うため，ファインチューニングを適用する．ファインチューニングは，学習済みモデルの重みを初期値として利用し，新しいモデルに受け渡す手法である．この手法は，学習済みモデルの結果を，求めたいモデルに反映させることができるため，少ないデータ量でも質の良いモデルを作成できる．学習させる際，重みの学習率を任意に変更することで新しいモデルへの反映率を変更できる．学習率が低い場合は，学習済みモデルと類似した結果を取得できるが新しいモデルの傾向を捉えられない．したがって，適切な学習率を設定する必要がある．

3.4 学習済みモデルの作成

ファインチューニングを行うため，学習済みモデルの作成を行う．精度のいいモデルを作成するために多くの同義文章のペアを学習させる必要があるため，同義文章のペアとして，やさしい日本語コーパス[4]を用いた．これは，50000 文をやさしい日本語（平易な日本語表現）に書き換えた対訳コーパスである．この同義文章を seq2seq に学習させることで，同義な文章を返す学習済みモデルを得る．平易な日本語表現に書き換える前の文章を Seq2seq の Encoder に，書き換えた後の文章を Decoder に学習させる．優しい日本語コーパスの例を表 3.2 に示す．

表 3.2 やさしい日本語コーパスの例

変換前	変換後
誰が一番に着くか私には分かりません.	誰が一番に着くか私には分かりません.
私には生きる上での価値がない.	私には生き甲斐がない.
彼女は，母に少し似ている	彼女は，母親に少し似ている.
その仕事を担当しようかと思っている.	その仕事を引き受けようかと思っている.
2時間も待った.	2時間も待たされた.
言っておく物は何もありません.	申告する物は何もありません.
彼は危険をものともしない.	彼は危険をものともしない.
もう夕食は食べ終わりましたか.	もう夕食は済みましたか.
それについてどう思った.	それについてどう思った.
彼らはちゃんと私の顔を見られなかった.	彼らはまともに私の顔を見られなかった.

第 4 章　実験結果と考察

4.1 まえがき

本章では，学習したモデルを用いてドメイン変換した実験結果及び実験の考察を述べる．

4.2 実験結果

ファインチューニングを行わずに題材データを用いて作成した事前モデル，やさしい日本語コーパスを用いた学習済みモデル，ファインチューニングを用いたファインチューニングモデルの三つに対する実験結果を示す．なお，Encoder に入力する文章は任意に選出した文章である．

4.2.1　事前モデル

ファインチューニングを行わずに，題材データを用いて作成した事前モデルのドメイン変換結果を表 4.1 に示す．

表 4.1 事前モデルによるドメイン変換

Encoder	Decoder
おはよう	それにとお兄ちゃんが入ってる_
こんにちは	「はったりって言ってるものは
さようなら	それはページを法お願いある人だよ
今日はいい天気ですね	言葉がいいつもりだよ
明日はいい天気ですね	天気です__私より上手は絶対ますけど
今日は祝日でした	ボクはチャールズと病院に一雨のこともあるわがた
明日は祝日でした	感謝祭は先月をあげるんだね
明日は何してますか？	これはどうやってない？
明日の天気を教えてください	天気チームを見には少年よ
好きな食べ物はなんですか？	よしね，マーシー学用品出来たら？
年はいくつですか？	分かる！は何もやつせてない？
好きな色を教えてください	いい気持ちがどっちで先輩か見ますか
それで全部です？	どこトゥからのその？
ボンク!	ただの意識だ
素敵な？	出来たと思うの？

4.2.2 学習済みモデル

やさしい日本語コーパスを用いて作成した学習済みモデルのドメイン変換結果を表 4.2 に示す

表 4.2 学習済みモデルによるドメイン変換

Encoder	Decoder
おはよう	みんなにおはようおはよう物に話した.
こんにちは	こんにちは品がなくなる思いが？
さようなら	さようなら保証さようなら.
今日はいい天気ですね	今日はいい天気ですね.
明日はいい天気ですね	明日はいい天気ですね.
今日は祝日でした	今日はへでした
明日は祝日でした	明日はへでした.
明日は何してますか？	明日は何していいですか？
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？	年も年はどのですか？
好きな色を教えてください	好きな色を教えてください.
それで全部です？	それでそれで大丈夫です？
ボンク!	ゲームのような態度を聞いている.
素敵な？	素敵な？

4.2.3 ファインチューニングモデル

学習済みモデルの重みを利用し，題材データを再学習したファインチューイングモデルのドメイン変換結果を表 4.3 に示す.

表 4.3 ファインチューニングモデルによるドメイン変換

Encoder	Decoder
おはよう	ださドアさだよ
こんにちは	学校これハーイをが入っ..
さようなら	痛いいく
今日はいい天気ですね	今日は.もっ
明日はいい天気ですね	もっともひとつさ
今日は祝日でした	早く，チャック_
明日は祝日でした	もっやってるも
明日は何してますか？	もしで！ママ分かるドルはしてる？
明日の天気を教えてください	もっとでいいのがもらわしかのがある⁉
好きな食べ物はなんですか？	彼，いつもたらなんて？
年はいくつですか？	“ら!だったかい？
好きな色を教えてください	彼が練習に行ったな気に入ったいことがぼく
それで全部です？	それだけ，50のは？
ボンク!	もっが番犬!!
素敵な？	これ犬小屋みんな

4.3 考察

事前モデルではドメイン変換を取得できず，文章として不成立であった．また，ほとんどの文章において文の関連性を見出し難い結果となった．原因として，題材データの学習精度が悪いことが挙げられる．また，Decoder で記号が散見された原因は，データの下処理の不十分さのためだと考えられる．一方で，「天気」という単語だけに着目すると，Encoder と Decoder の両方で見られるため，局所的にはモデルが作成できていると考えられる．

学習済みモデルでは，事前モデルやファインチューニングモデルと比べ，文章を入力すると同義文章を返すモデルを作成することができた．また，「おはよう」「こんにちは」「さようなら」などの短い単語はそれぞれ 6，14，15 個がやさしい日本語コーパス内に存在し，各文章は短い文章のため学習しやすいと考えられる．しかし，Decoder では入力文章と同じ言葉を繰り返している．原因として，短い単語を一つの文章として認識しており，余剰な言

葉を付け足す傾向にあると考えられる．また，やさしい日本語コーパス内で「祝日」が存在していないため，Decoder では「へ」と変換している．これは，学習済みモデルにおいて前後の文脈から「祝日」と「へ」は同義な単語にドメイン変換されたと考えられる．入力文章では「祝日」が連想できない文脈であったが，「祝日」のような単語辞書にない言葉でも，文章内で「祝日」が連想できる場合，適切な単語にドメイン変換できると考えられる．

ファインチューニングモデルでは，事前モデルと同様にドメイン変換を取得できず，文章として不成立であった．原因として，データの下処理の不十分さや，学習済みモデルの重みが効果的に反映されてないことが挙げられる．学習済みモデルの重みを初期値として用いるが，学習率を高く設定していたために題材データの傾向を多く捉えてしまったと考えられる．そのため，記号の出現や文の関連性を見出し難い結果となり，学習を進めることでモデルの精度が悪くなったと考えられる．

各モデルにおいて，クエスチョンマークは高確率で再現できている．文末で使用され，データ数も多いことで学習が容易であったと考えられる．したがって，単語や文脈において学習難易度の高さに相違があると考えられる．短い文章や記号などの学習難易度が容易なデータから学習させていくことで，モデルの精度が向上することが考えられる．

第 5 章　結論

5.1 結論

本研究では，Seq2seq とファインチューニングという手法を用いて，機械翻訳におけるドメイン変換を行った．Seq2seq のモデルでは，文章を入力することで同義文章を返答する事前モデル，学習済みモデル，ファインチューニングモデルを作成した．ファインチューニングモデルは，学習済みモデルの重みを利用し，題材データを再学習したモデルである．学習済みモデルでは，高精度で同義文章を取得できたが，事前モデル，ファインチューニングモデル共に同義文章を取得できなかった．

5.2 課題

本研究での課題点として，ファインチューニングを行う際の学習率の最適化と，データの下処理を充分に行うこと，学習難易度が低い文章の選出をして，モデルの精度を向上させることが挙げられる．

また，各モデルの数値での評価が不透明なため，正当に評価する指標について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] 緒方, 佐藤, 駒谷: “模倣と置換に基づく超短編小説の自動生成”, 2014 年度人工知能学会全国大会論文集, pp.1CS-OS-14B-2, May 2014

[2] I. Sutskever, O. Vinyals, and Q. V. Le: “Sequence to Sequence Learning with Neural Networks,” Neural Information Processing Systems (NIPS), pp1-9, Dec. 2014

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

[4] SNOW T15:やさしい日本語コーパス
http://www.jnlp.org/SNOW/T15, (平成 31 年(2019 年)1 月現在)

図一覧

表一覧

研究業績

1.京極，渡辺，“機械翻訳におけるドメイン変換に関する検討”，2019 年電子情報通信学会総合大会 2018 年 3 月発表予定