卒業論文

Transformer に基づく
やさしい日本語の翻訳モデル生成

指導教官　　村上 陽平 教授

立命館大学　情報理工学部
先端社会デザインコース　4 回生
2600200159-3

黒山晃輝

2023 年度（秋学期）卒業研究 3（CH）
令和 6 年 1 月 31 日

Transformer に基づくやさしい日本語の翻訳モデル生成

黒山　晃輝

内容梗概

近年，グローバル化が進んだことにより，在留外国人の人数が増えている．現在の在留外国人数は，296 万 1，969 人で，この 30 年間で約 3 倍に増えており，国籍の多様化も進んでいる．これらの在留外国人の 8 割の日本語能力は，日常生活に困らない程度の会話力に留まっており，日常生活だけではなく，医療現場や災害時などにおいて，「やさしい日本語」の重要度は高まっている．

「やさしい日本語」とは，普段使われている言葉を，外国人にも分かるように配慮した簡単な日本語であり，出入国在留管理庁と文化庁は，やさしい日本語ガイドラインを出している．このガイドラインに従い，入力された文章や文書の含まれている語彙や文法などからやさしい日本語であるかを診断するサービスや入力された文章のどこがやさしくないのかを指摘してくれるサービスも構築されている．

しかしながら，既存のサービスは，原文に対するルビ振りや用語説明の付記などに限定されており，分かりやすい言い回しへの変換ができていない．

そこで，本研究では，日本語からやさしい日本語を生成するニューラル機械翻訳モデルを作成する．

具体的には，やさしい日本語コーパス（SNOW T15:やさしい日本語コーパス，SNOW T23:やさしい日本語拡張コーパス）と Transformer に基づくニューラル機械翻訳である Open-NMT を利用して，翻訳モデルを構築する．本手法の実現にあたり，取り組むべき課題は以下の２点である．

語彙数の拡大

やさしい日本語コーパスは一般的な日本語コーパスと比べて，サイズが小規模なため，やさしい日本語コーパスのみでニューラル機械翻訳を学習すると，未知語が増え，翻訳精度を低下させる原因となる．したがって，ニューラル機械翻訳が処理可能な語彙数を増やし，未知語を減らす必要がある．

ドメイン適応

日本語からやさしい日本語に翻訳する際に，言語が同じため大部分は入力文のまま出力するものの，一部の難しい用語や言い回しのみをやさしい用語や言い回しに変換する必要がある．したがって，一般的な日本語を出力するモデルをベースとして用い，小規模なコーパスによりベースモデルやさしい日

本語に適応させる必要がある.

前者の課題に対しては，日本語とやさしい日本語の対訳コーパス 8 万件に加えて，第 2 コーパスとして日本語大規模コーパスを併用する．やさしい日本語も日本語のため日本語大規模コーパスの各文を複製して原文と訳文のペアを作成し，翻訳元言語，翻訳先言語をともに日本語とした対訳コーパスを構築する．次に，日本語大規模コーパスとやさしい日本語コーパスを統合して語彙データを作成する．ただ，やさしい日本語コーパスのサイズが日本語大規模コーパスと比べて極端に小さいため，やさしい日本語コーパスのみで追加学習を行う反復型学習フローにより，やさしい日本語へのドメイン適応を強化していく．

後者の課題に対しては，大規模コーパスで汎用的な日本語を生成可能な翻訳モデルを構築した後に，やさしい日本語コーパスによる学習を反復させる学習フローにより，ドメイン適応を強化する．追加学習を行い，やさしい日本語翻訳モデルの精度を向上させていく．

未知語の削減率と BLEU 値を用いて，やさしい日本語コーパスのみだけで訓練した場合と日本語大規模コーパスも併用した場合とで比較し，提案手法の有効性を検証した．本研究の貢献は以下の通りである．

語彙数の拡大

やさしい日本語コーパスで学習した場合と日本語大規模コーパスを併用して学習した場合では，語彙数を増加させることに成功した．

さらに,出力と入力での未知語はやさしい日本語コーパスのみを用いた場合と比較して，90 件から 0 件まで削減することに成功した．

ドメイン適応

日本語大規模データを利用し構築された汎用的な日本語を出力できるモデルと，やさしい日本語コーパスのみで学習を繰り返す反復フローによって，ドメイン適応を強化した場合では，BLEU スコアは，0.5793 から 0.6002 向上させ，21 万回の追加学習を行うと，BLEU 値に変化がないことがわかった．

Translation Model Generation for Easy Japanese Based on Transformer

Koki Kuroyama

Abstract

In recent years, the number of foreign residents in Japan has been increasing. The current number of foreign residents is 2,961,969, a three-fold increase over the past 30 years. The Japanese language proficiency of 80% of these foreign residents is limited to the conversational level necessary for daily life, and the importance of "Easy Japanese" is increasing.

The Immigration and Residence Agency and the Agency for Cultural Affairs have issued guidelines for Easy Japanese. In accordance with these guidelines, there are services that diagnose whether the Japanese is easy or not, and services that point out what is not easy about the input text.

However, existing services are limited to adding ruby to the original text and adding explanations of terms, and are not able to convert the text into easy-to-understand phrases.

This research introduces a neural machine translation model designed to convert Japanese into a more comprehensible, easy-to-understand version of the language. To implement this, a translation model is developed utilizing a corpus comprising simple Japanese texts (SNOW T15: Easy Japanese Corpus, SNOW T23: Easy Japanese Extended Corpus) and leveraging Open-NMT, a Transformer-based neural machine translation framework. To actualize this approach, two primary challenges must be surmounted.

Vocabulary Expansion

Because the Easy Japanese Corpus is small in size, training neural machine translation only on the Easy Japanese Corpus will increase the number of unknown words, which will cause a decrease in translation accuracy. Therefore, it is necessary to increase the number of words that can be processed by neural machine translation to reduce the number of unknown words.

Domain Adaptation

When translating from Japanese to Easy Japanese, only some difficult terms and phrases need to be converted to easy terms and phrases. Therefore, it is

necessary to use a model that outputs general Japanese as a base, and adapt the base model to Easy Japanese by using a small corpus.

For the former issue, in addition to a bilingual corpus of 80,000 Japanese and Easy Japanese, a large Japanese Corpus was used as a second corpus. Since Easy Japanese is also Japanese, each sentence in the large-scale Japanese corpus was replicated to build a bilingual corpus in which both source and target languages are Japanese. Subsequently, a lexical dataset was formed by combining the large Japanese corpus with the Easy Japanese corpus. Due to the notably smaller size of the Easy Japanese corpus, domain adaptation to Easy Japanese will be improved by conducting supplementary training exclusively on the Easy Japanese corpus.

For the latter task, after building a translation model that can generate general-purpose Japanese from a large corpus, the domain adaptation was strengthened through a learning flow that iterates training on the Easy Japanese corpus. Additional learning performed to improve the accuracy of the Easy Japanese translation model.

Using the reduction rate of unknown words and BLEU values, the effectiveness of the proposed method was verified by comparing the results of training on an Easy Japanese corpus alone with those of training on a large Japanese corpus as well. The contributions of this study are as follows.

Vocabulary Expansion

An increase in the vocabulary size was achieved through learning with both the Easy Japanese Corpus and the large Corpus of Japanese. Additionally, compared to using only the Easy Japanese Corpus, the count of unknown words was effectively diminished from 90 to 0.

Domain Adaptation

When domain adaptation was enhanced by using a model that can output generalized Japanese and an iterative flow that repeats training only with an Easy Japanese corpus, the BLEU score increased from 0.5793 to 0.6002, and after 210,000 additional training cycles, the BLEU value remained unchanged.

Transformer に基づくやさしい日本語の翻訳モデル生成

目次

第1章 はじめに

近年，グローバル化が進み，日本では外国人の観光客や在留外国人の人数も増えてきている．実際に，出入国在留管理庁の調査の図 1 によると，令和 4 年 6 月末の在留外国人数は，296 万 1969 人で，前年末に比べると 20 万 1334 人（7.3%）増加している．毎年，外国人の人数は増えてきており，国籍の多様化も進んでいる．図より，平成 24 年から比べると約 1.5 倍になっている．

図 2 の在留外国人の国別人数によると，中国が 744,551 人で構成比は 25.1%，ベトナムが 476，346 人で 16.1%，続いて韓国が 412，340 人で 13.9%である．上位 10 か国はアジア諸国が中心であるが，アメリカやブラジルも含まれている．さらに，上位 10 か国・地域ではいずれも前年末に比べ増加している．

そのような現状の中で，外国人が生活する中では，言葉の壁が存在してしまう．さらに日本は他の国と比べても災害が多い国であり，緊急時などで外国人が情報弱者となってしまう．そこで，本研究で扱う「やさしい日本語」が注目されている．

「やさしい日本語」とは，普段使われている言葉を，外国人にも分かるように配慮した簡単な日本語を指す．日常的な場面や身近な話題で使われる日本語を「ある程度」理解できる人が使うレベルである．

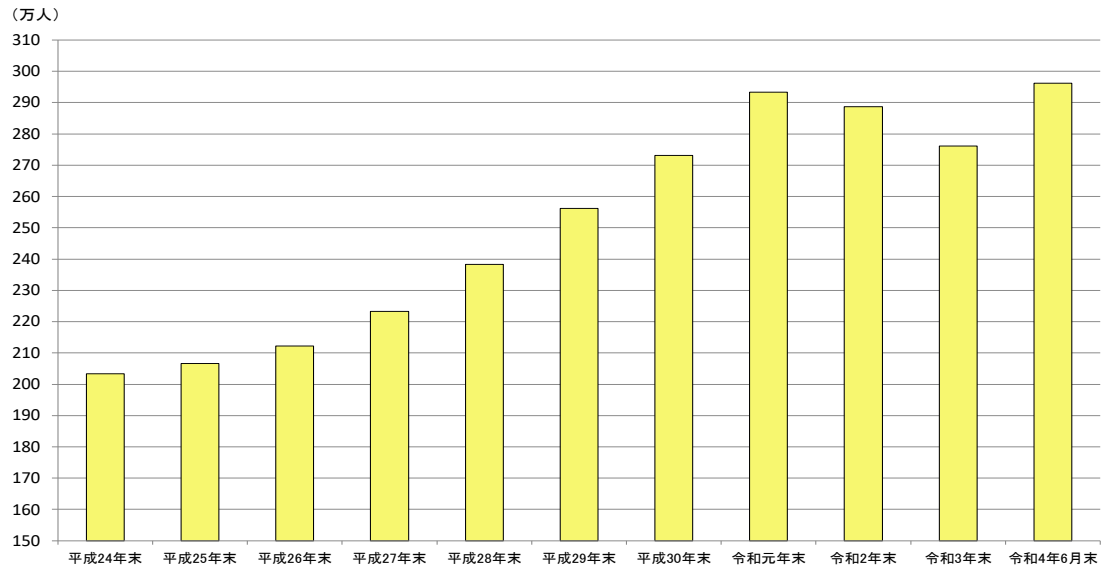

図 1 ：在留外国人数の変化

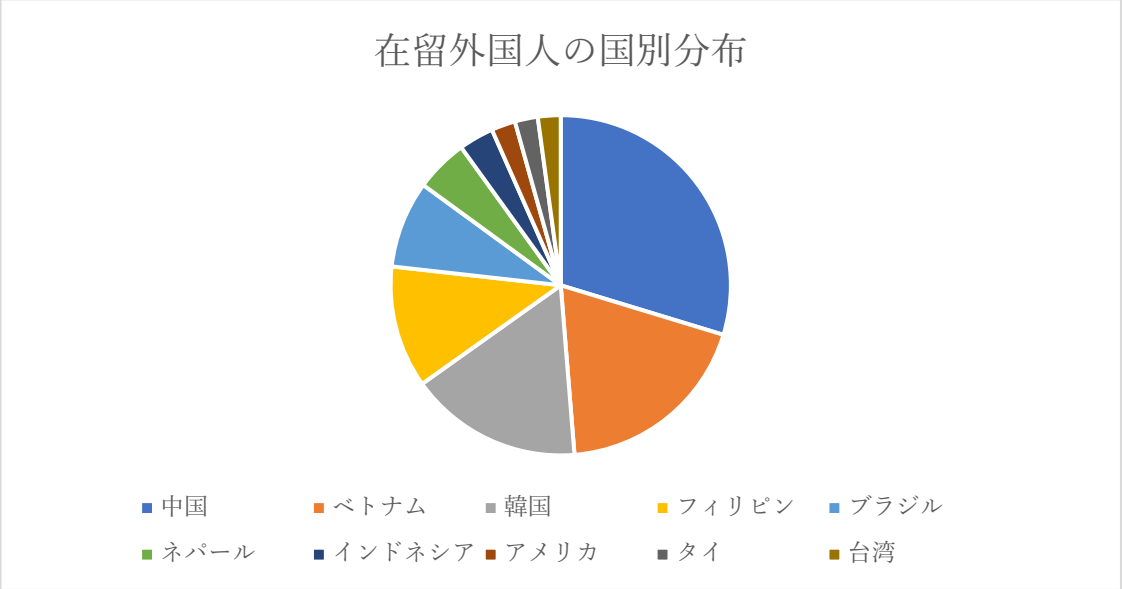

図２：在留外国人の国別人数

脚注：出入国管理庁ホームページ：令和４年度在留外国人に対する基礎調査：https://www.moj.go.jp/isa/policies/coexistence/04_00017.html （令和５年９月公表）

やさしい日本語は，阪神淡路大震災の時から注目がされ始め，研究が進められてきた．この震災のとき，日本人の死傷者は約 1%であったが，外国人の死傷者は 2%以上であった．外国人の中には地震を経験したことのない人もおり，災害が起きた時にどのような行動をすればいいのかわからず，混乱し，情報弱者となってしまう．それが原因で，被害者が増えてしまう結果となってしまった．その後，新潟県中越地震(2004 年)や東日本大震災(2011 年)を経て，災害時のやさしい日本語での発信の取組が全国に広がった．ただ，熊本地震（2016 年）の時には，避難先での生活でトラブルが起きていた事例もあった．

やさしい日本語を活用していくことは，外国人が日本で生活する中で，必要なことである．令和４年度在留外国人に対する基礎調査での，公的機関が発信する情報を入手する際の困りごとでは，「多言語での情報発信が少なかった」（20.4%）が最も多く，「やさしい日本語での情報発信が少なかった」（13.0%）が３番目に多いなど，言語に関する困りごとが多いことがわかる．さらに，母語以外の情報提供を望む言語は「日本語」が 52.7%で最多であり，「英語」が 37.6%，「やさしい日本語」34.2%と続く．このように，外国人は言語に対して問題が発

生している.

災害時の外国人に対する情報提供のための日本語表現とその有効性に関する試論によると，ニュース文をやさしい日本語に変換することは，外国人が情報を理解するに当たって，大幅に理解度が高くなることが確認され，災害時の情報を円滑に伝える上で，とても有効な伝達手段であることが確認された.

「やさしい日本語」は，外国人に向けてのみに限らず，小さな子供や高齢者にも伝わりやすい言語としても注目されていて，その中で問題を解決する方法として，有効的な手段であると考えられている.

現在の取り組みとしては，出入国在留管理庁と文化庁は，やさしい日本語ガイドラインを作成しており，各自治体が「やさしい日本語」での情報発信をする取り組みがされている．さらに，サービスとして，やさしい日本語であるか評価を行う「やさにちチェッカー」や外国人の方などにも分かるように伝える「NEWS WEB　EASY」などがある.

しかしながら，既存のサービスは，原文に対するルビ振りや用語説明の付記などに限定されており,分かりやすい言い回しへの変換ができていない.例えば,「緻密な」という日本語はそのまま「緻密な」という日本語のまま出力されており,やさしい日本語に翻訳されていない.公開されている言語資源としては,「やさしい日本語対訳コーパス」で作成されたやさしい日本語コーパス（SNOW　T15:やさしい日本語コーパス）と拡張されたコーパス（SNOW T23:やさしい日本語拡張コーパス）しか存在せず，少数資源言語となっている．日本語から英語への翻訳と比べると，すべての難しい日本語を網羅している現状ではないので，翻訳の向上をしていく必要がある.

そこで，本研究では，Transformer に基づくやさしい日本語翻訳モデルを生成する.

具体的には，やさしい日本語コーパス（SNOW　T15:やさしい日本語コーパス，SNOW T23:やさしい日本語拡張コーパス）と Open-NMT を利用して，ニューラル機械翻訳で翻訳モデルの精度を向上させていく．本手法の実現にあたり，取り組むべき課題は以下の 2 点である.

語彙数の拡大

やさしい日本語コーパスは一般的な日本語コーパスと比べて,サイズが小規模なため,やさしい日本語コーパスのみでニューラル機械翻訳を学習すると，未知語が増え，翻訳精度を低下させる原因となる．したがって，ニューラル

機械翻訳が処理可能な語彙数を増やし，未知語を減らす必要がある．

ドメイン適応

日本語からやさしい日本語に翻訳する際に，言語が同じため大部分は入力文のまま出力するものの，一部の難しい用語や言い回しのみをやさしい用語や言い回しに変換する必要がある．したがって，一般的な日本語を出力するモデルをベースとして用い，小規模なコーパスによりベースモデルやさしい日本語に適応させる必要がある．

前者の課題に対しては，日本語とやさしい日本語の対訳コーパス８万件に加えて，第２コーパスとして日本語大規模コーパスを併用する．やさしい日本語も日本語のため日本語大規模コーパスの各文を複製して原文と訳文のペアを作成し，翻訳元言語，翻訳先言語をともに日本語とした対訳コーパスを構築する．次に，日本語大規模コーパスとやさしい日本語コーパスを統合して語彙データを作成する．ただ，やさしい日本語コーパスのサイズが日本語大規模コーパスと比べて極端に小さいため，やさしい日本語に対応させる必要がある．そこで，後者の課題の提案手法であるドメイン適応を用いることで解決していく．

後者の課題に対しては，大規模コーパスで汎用的な日本語を生成可能な翻訳モデルを構築した後に，やさしい日本語コーパスによる学習を反復させる学習フローにより，ドメイン適応を強化する．追加学習を行い，やさしい日本語翻訳モデルの精度を向上させていく．

以下本論文では，第２章で関連研究とやさしい日本語の定義について，第３章で提案手法 1 つ目の大規模日本語コーパスの併用による語彙数の拡大，４章で提案手法 2 つ目のやさしい日本語コーパスの反復フローによるドメイン適応について述べている．第５章では，評価として語彙数の拡大，ドメイン適応について述べている．

第2章 日本語からやさしい日本語への変換

2.1 やさしい日本語

「やさしい日本語」とは，普段使われている言葉を，外国人にも分かるように配慮した簡単な日本語を指す．日常的な場面や身近な話題で使われる日本語を「ある程度」理解できる人が使うレベルである．実際に政府や自治体でも，看板，掲示物，文書等で現在利用が進められている．

出入国管理庁が出している「やさしい日本語ガイドライン」には，やさしい日本語に変換するときのポイントを挙げている．

さらに，やさしい日本語であるかを確認するサービスや，やさしい日本語に翻訳するサービスも存在するが，すべての難しい日本語を網羅していることはない．

表 1 と図 3 に日本語からやさしい日本語に変換する例と実際にやさしい日本語を利用した掲示物の例を挙げる．

表１：やさしい日本語に変換する例

日本語	やさしい日本語
彼はその事件の真相を追求している	彼はその事件の本当のことを詳しく調べている
その国は，経済的に困難な状況に直面している	その国は，経済的に難しい状況になっている
その絵画は，複雑なテクニック用いて描かれている	その絵は，難しい技術を用いて描かれている
ご用件は何ですか？	どうしましたか？
高台に避難してください	高いところに逃げてください

図３：やさしい日本語を利用した掲示物の例（引用：北海道十勝音更町ホームページ)

2.1.1 やさしい日本語の区別

日本語からやさしい日本語に変換するに当たって，変換するときのパターンの区別を行う．

以下には，変換するときのパターンを 3 つに分けて，表２に表す．

１つ目は語句の変換によってやさしい日本語に翻訳をするが，単語数は変化しないものである．例であれば，「テクノロジー」が「技術」に，「返品してください」が「返してください」に翻訳されており，やさしい日本語への翻訳はされているが，単語数は変化しない．

２つ目は語句の変換によってやさしい日本語に翻訳しているが，単語数が変化している．例であれば，「レストラン」が「食事を提供する店」に，「ラストオーダー」が「最後の注文」に翻訳されており，やさしい日本語への翻訳はされているが，単語数が変化している．このパターンに関しては，1 つの単語に対して，説明が付与されている例である．

３つ目は，表現全体が変化するものである．例であれば，「ご用件は何ですか？」が「どうしましたか？」に，「土足厳禁です」が「靴を脱いでください」に翻訳されており，やさしい日本語への翻訳がされるときに，表現全体が変化している．

評価を行う際には，この 3 つのパターンに分けて，どのような翻訳のされ方であれば，精度が高く翻訳できるのか，実際に検証を行い，比較を行う．

表２：やさしい日本語の区別

	原文	やさしい日本語
語句の変換（単語数が変化しない）	彼は，新しいテクノロジーを使う	彼は，新しい技術を使う
	この商品を返品してください	この商品を返してください
語句の変換（単語数が変化する）	私たちはレストランを利用した．	私たちは食事を提供する店を利用した
	ラストオーダーです．	最後の注文です．
全体の表現が変化する	ご用件はなんですか？	どうしましたか？
	土足厳禁です	靴を脱いでください

2.1.2 やさしい日本語に変換するときのルール

日本語からやさしい日本語に変換する際に，やさしい日本語の明確な正解は現在定義されていない．WEB 上ではやさしい日本語辞書も存在しているが，明確なルールがないのが現状の課題の 1 つでもある．そこで，出入国管理庁では「やさしい日本語ガイドライン」を出しており，その中で具体的にやさしい日本語に変換するときのポイントを挙げている．そこで下ではそのポイントを説明する．

まず，日本人に分かりやすい文章に書き替える．情報を整理し，１つの文章を短くすることで分かりやすい文章に置き換え，外来語は基本的に使わないようにする．次に，外国人にわかりやすい文章に置き換える．いくつかポイントが存在している．１つ目は二重否定を使わないことである．「～でないことはない」のような文章は外国人にとっては，分かりにくい文章であるため，変換する必要がある．２つ目は，受身形や使役表現を必要最低限の使用で抑えることである．主語に対しての行動を示す内容が分かりにくくなる可能性があるため変換をする必要がある．３つ目は簡単な言葉を使う点である．和語を使わないことや抽象的な言葉は使わないことで，難しい表現を簡単な表現に変換する．４つ目は，ルビ振りを行うことである．漢字に対しては，ひらがなのルビ振りを付与することで，漢字を読むことができない外国人にも理解できるように行う．他にも具体的なポイントは存在するが，上に挙げた４点が大きなポイントである．そして，最後にやさしい日本語であるかを確認する．このようにして，外国人にとって分かりにくい日本語をやさしい日本語に変換する．

2.2 関連研究

2.2.1 統計機械翻訳を用いたやさしい日本語への自動変換

熊野らは，NHK から公開される「NEWS WEB EASY」を制作する業務を効率化するために，統計機械翻訳を利用してニュース文をやさしい日本語に自動変換するシステムを制作した．このシステムはこれまで制作する業務の時に，作成された日本語の教師による書き換えの例から学習することによって，自動で翻訳されるシステムを実現する．具体的に概要を説明する．まず，句ベース統計機械翻訳によるやさしい日本語への自動変換について述べる．句ベース統計機械翻訳とは，機械翻訳を行うために，まずは翻訳先言語の文とその翻訳先の言語の文のペアから構成されるコーパスを使って，単語やフレーズの翻訳に関する統計的なモデルを作成する．次に，翻訳先の言語のテキストだけを集めたコーパスを用

いて，単語がどのように並ぶかについての統計的なモデルを学習する．これらのモデルを組み合わせて，機械翻訳を実現させる．その中で学習コーパスの小ささを克服するために，3 つの対策を行っている．1 つ目は，「default 翻訳」の対訳コーパスへの追加である．対訳コーパスの翻訳元側に出てくる異なる訳のそれぞれについて，「その 1 語からなる文」と「同じ文対からなる文」の文章のペアを生成する．そして，これらを「default 翻訳」として対訳コーパスに追加する．2 つ目は，品詞言語モデルの併用である．「Factored language model」を利用し，言語モデルとして表層 n-gram モデルと品詞 n-gram モデルの 2 つを利用する．3 つ目は，語の汎化である．複数の異なる語を 1 つの類似語に一般化し，出力にこれらの似ている語が現れた場合には，規則を利用して，語の表層を生成する．具体的には，数の汎化と機能語の汎化を行った．評価は BLEU スコアと書き換え数と正解数で行った．学習コーパスの小ささへの対策によって対策を行い，書き換えを行うシステムの品質向上に一定の効果があったことが認められた．

2.2.2 災害時の外国人に対する情報提供のための日本語表現とその有効性に関する試論

松田らは，災害が起きた時に外国人にどのように必要な情報を提供するべきなのかについて研究し，「ある程度日本語を理解できる外国人に伝わる日本語を用いた災害情報の表現の方法と有効性」について検証を行った．実験の概要を具体的に説明する．

聴解実験用のニュース文として，震災が起きた時に実際に放送された NHK のニュースの原稿を記録したテープ A と，やさしい日本語表現に言い替えた，ほぼ同じ内容であるテープ B を用意した．テープの A と B は 5 つのニュース文からなっている．ニュース原稿とニュース用案文，それぞれの平均文字数は，A が約 80 字，B が約 64 字である．A は 1 分間に約 400 字で音読し，B は 1 分あたり約 200 字で音読している．表３にニュース文の例を挙げる．

被験者は，日本語レベルが初級後半から中級前半程度の日本語話者ではない人である．そして，そのニュースの文章を聞いた後に，理解できているかどうか試すためにテストを行い，点数を算出し検証を行った．

実験の結果を示す．

14 点満点のテストに対して,A のグループの平均点は 4.1 点の正答率は 29.3％であった．

B のグループの平均点は 12.7 点の正答率は 90.7％であった．

表３：ニュース文の例

A	避難していた住民たちが自宅に帰宅し，充満していたガスに気づかず，夕食の準備や暖房のスイッチを入れ，新たな出火が趨きているという情報があり，消防ではガス漏れにも十分気をつけるよう，呼びかけています.
B	部屋でガスがくさい時は火を使わないでください 電気もつけないでください すぐ口窓をあけてください ガスが部屋に漏れていると火事になるかもしれません ガスに気をつけてください

外国人のための提案された日本語でのニュース文章の理解度は，普段放送されているニュースよりも，高くなることが確認された．災害が起こった時に，情報を円滑に伝える手段として，とても有効なものであることが確認された.

2.2.3 やさしい日本語対訳コーパスの構築

山本らは，やさしい日本語が低資源言語である点に注目し，やさしい日本語コーパスの構築を行った．具体的に，概要を説明する.

書き換えを行っていくための使用テキストは，small_parallel_enja: 50k En/Ja Parallel Corpus for Testing SMT Methods である．このコーパスは日本語から英語の機械翻訳のために作成された規模が小さい対訳コーパスである．そして，このコーパスの 5 万文をすべてやさしい日本語に変換する．作業は人手で，5 人で行った．作業手順を具体的に説明していく.

まず，最初の簡単な日本語語彙として，BCCWJ の UniDic 高頻度 2000 語を選択する．次に，入力文に対して単語の解析を行い，やさしい日本語ではない単語を含む場合は書き換えを行う．書き換えは単語単位ではなく，文単位で行うようにし，やさしい日本語の語彙のみでできるだけ同じ意味となるように努める前提で，原文中に含まれている情報の一部の情報がなくなることは許し，作業の途中に，作業者にやさしい日本語の語彙の追加や削除を許す．ある時点で追加した語句や削除した語句を集め，やさしい日本語の語彙の定義の修正を行い，一時的にやさしい日本語の語彙が 2000 語以上になることを許す．やさしい日本語の語彙の定義の修正をした場合は，ステップ 2 から作業を繰り返す.

このようにして作業を行い，そして，５万文の（やさしい日本語，日本語，英語）の対訳コーパス，2000 語のやさしい日本語辞書，やさしい日本語チェッカーを公開した.

第3章 大規模日本語コーパスの併用

3.1 大規模日本語コーパス

やさしい日本語コーパスは一般的な日本語コーパスと比べて，サイズが小規模なため，やさしい日本語コーパスのみでニューラル機械翻訳を学習すると，未知語が増え，翻訳精度を低下させる原因となる．したがって，ニューラル機械翻訳が処理可能な語彙数を増やし，未知語を減らす必要がある．

未知語となった語句は，やさしい日本語コーパスのみで学習させた場合では，未知語は 90 件存在しており，大規模日本語データ併用することによって，未知語となる語句を減らしていく．

具体的には，やさしい日本語の対訳コーパスである約８万件に加えて，第 2 コーパスとして日本語大規模コーパスを併用する．この日本語大規模データは JParaCrawl から入手してきた文章のノイズを除去したデータの約 110 万件を利用する．やさしい日本語も日本語のため日本語大規模コーパスの各文を複製して原文と訳文のペアを作成し，翻訳元言語，翻訳先言語をともに日本語とした対訳コーパスを構築する．そして，この２つのコーパスから，語彙データを作成し，

表 4：やさしい日本語コーパスの例

	翻訳元言語	翻訳先言語
やさしい日本語コーパス	彼は自分の考えを書き留めた.	彼は自分の考えを書いておいた.
	多くの動物が人間によって滅ぼされた.	多くの動物が人間によって殺された.
	もう夕食は済みましたか.	もう夕食は食べ終わりましたか.
	ここでは自由に振る舞っていいですよ.	ここでは自由にしていいですよ.
	この事実を心に留めておいて下さい.	この事実を覚えておいてください.

表 5：日本語大規模複製コーパスの例

	翻訳元言語	翻訳先言語
日本語大規模複製コーパス	宿泊中はご自由にお使いください.	宿泊中はご自由にお使いください.
	しかし，これらの観測が成功するかどうかは確かではありません.	しかし，これらの観測が成功するかどうかは確かではありません.
	もちろん，すべてテンプレートによって異なります.	もちろん，すべてテンプレートによって異なります.

ニューラル機械翻訳で学習させることにより，語彙数の拡大を行う.

表 4 と表 5 に，やさしい日本語コーパスと日本語大規模複製コーパスの例の一部を挙げる.

この 2 つのコーパスとコーパスから作られた語彙データを Transformer に基づくニューラル機械翻訳である Open-NMT で学習させることによって，汎用的な日本語を翻訳することができる翻訳モデルを構築する.

3.2 語彙の拡張

やさしい日本語コーパスのサイズが日本語大規模コーパスと比べて極端に小さいため，低頻度のやさしい日本語が語彙データに収録されない問題がある．そのままの翻訳モデルを学習した場合，やさしい日本語の語彙が反映されないため，やさしい日本語コーパスでのドメイン適応を行うことによって，やさしい日本語に変換されるモデルの精度の向上を行っていく．ドメイン適応に関しては，第 4 章で論ずる.

第4章 反復型学習フローによるドメイン適応

4.1 ドメイン適応

ドメイン適応（Domain Adaptation）は，機械学習と特に自然言語処理（NLP）で用いられる手法の 1 つである．この手法は，ある特定のドメインやタスクで学習したモデルを，データは異なっているが，関連しているドメインやタスクに適用する際に用いる．ドメイン適応を行う目的としては，モデルが新しいドメインのデータに対してもうまく機能するようにすることである．

ドメイン適応には，いくつか手法の種類があり，本研究では「ファインチューニング」という手法を用いる．

ファインチューニングとは，既に学習が行われたモデルを，新たなデータを使って追加学習を行うというものである．ファインチューニングを行う際には，学習率を既存の学習モデルよりも，下げて行うため，影響力は比較的小さくなる．

4.2 やさしい日本語コーパスによるドメイン適応

日本語からやさしい日本語に翻訳する際に，翻訳をしようとする言語が同じため大部分は入力文のまま出力するものの，一部の難しい用語や言い回しのみをやさしい用語や言い回しに変換する必要がある．日本語大規模コーパスを併用した場合，やさしい日本語コーパスの語彙データが小さくなってしまい，やさしい日本語に翻訳されない文章が出てきてしまう．したがって，一般的な日本語を出力するモデルをベースとして用い，小規模なやさしい日本語コーパスによりベースモデルをやさしい日本語に適応させる必要がある．

本研究では，大規模コーパスで汎用的な日本語を生成可能な翻訳モデルを構築した後に，やさしい日本語コーパスによる学習を反復させる学習フローにより，ドメイン適応を強化する．

具体的に説明をしていく．

1. 日本語大規模データを複製することによって，日本語大規模コーパスを作成する
2. 日本語大規模コーパスを併用し，2 つのコーパスから語彙データを作成する
3. Open-NMT でバッチサイズ 4，ステップ数 300000 回，エポック数 1 で学習

を行う

4. 汎用的な日本語を生成することができる翻訳モデルを作成する
5. 構築された翻訳モデルをやさしい日本語コーパスのみで追加学習を行うことにより，ドメイン適応を強化する
6. 最終的に完成したモデルは翻訳結果が適切であるかを確認する
7. そうでなければもう一度追加学習を行い，さらに，ドメイン適応を強化していく

図 4 では，学習の流れを示すフローチャートを表す.

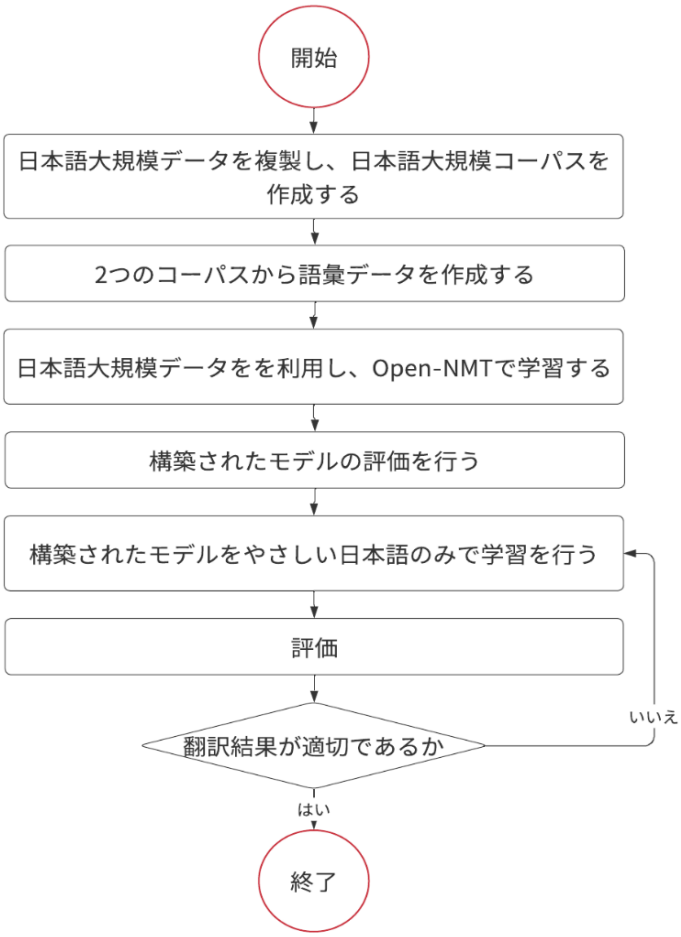

図４：学習の流れを示すフローチャート

第5章 評価

5.1 語彙数

語彙数に関しては,ランダムに生成された日本語の入力文120個に対して「unk」と出力される割合と実際の語彙データに含まれる語句の数で比較を行う.

実際の語句データをやさしい日本語コーパスのみで学習させた場合と日本語大規模データを併用した場合とで比較する.

実際の語句のデータ数は，表６のとおりである．語彙データはやさしい日本語コーパスのみで学習させた場合では，SOURCE は 20875 件，TARGET は 6227 件であったが,日本語大規模データを併用した時では,SOURCE は,354938 件,TARGET は 353640 件まで上げることに成功した.

表７では，120 文に対する入力と出力での unknown と表示される個数の変化を示す.やさしい日本語コーパスのみで学習させた場合では，入力と出力の合計で90件unknownと表示されていたが,日本語大規模データを併用した場合では,入力と出力の合計でも０件まで減少させることに成功した.

表６：語彙データに含まれる語句の個数

	日本語	やさしい日本語
やさしい日本語コーパスのみ	20875	6227
日本語大規模データを併用した場合	354938	353640

表７：入力と出力で未知語となる語句の個数

	日本語	やさしい日本語
やさしい日本語コーパスのみ	90 件	0 件
日本語大規模データを併用した場合	0 件	0 件

表 8：BLEU 値に用いた日本語とやさしい日本語の正解データ

日本語	やさしい日本語
彼の論文は，非常に深遠な思想を含んでいる．	彼の論文はとても深い思想を持っている
彼女は百貨店で洗練された服を選んだ．	彼女はいろいろなものを売っている店で品が良い服を選んだ
私たちはその問題を解決するための効果的な戦略を練っている．	私たちはその問題を解決するための効果的なものを作っている．
その新しいテクノロジーは，我々の生活を大きく変える可能性がある．	その新しい技術は，私たちの生活を大きく変える可能性がある．
私の親友は，非常に緻密な性格を持っている．	私の仲の良い友達は，とても詳しく良い性格を持っている．

5.2 正解データの作成

翻訳された日本語がやさしい日本語であるか評価を行うために，BLEU の正解データとして，やさしい日本語のデータを作成する．

やさしい日本語の作成は，手動で行うものとするが，出入国管理庁から出されている「やさしい日本語ガイドライン」に従い，作成するものとする．しかし，本研究で用いるやさしい日本語コーパスでは，語句の変換が中心となっており，本研究では，語句の変換を重視し，やさしい日本語の正解データを作成する．

実際に利用した日本語とやさしい日本語の正解データ５つを，上記の表８に表す．

5.3 ドメイン適応

やさしい日本語コーパスと日本語大規模データ併用し学習させた場合と学習させた翻訳モデルをやさしい日本語コーパスのみで再学習した場合とで比較を行った．さらに，再学習を行った回数によってどのようにドメイン適応がされていったのかを比較する．

さらに，やさしい日本語の区別を行い，どのようなパターンの日本語が正しく翻訳されるのかの検証も行う．入力文はランダムに作成された文章 60 個を用いる．出力に用いたデータは，基本的に最も値が高いものを利用する．

もし，値が高いものが意味の通っていない場合は，上位 3 つを出力し，その中

表９：BLEU 値の変化

	BLEU 値
汎用的な日本語を出力することができる翻訳モデル	0.5793
再学習を行った学習モデル（最高値）	0.6002

から最も日本語の意味が通っているものを参照するデータとして利用する．

表９は，BLEU 値を評価したものである．

1 つ目は汎用的な日本語を出力することができる翻訳モデルであり，BLEU 値は，0.5793 となっている．一方，再学習を行った学習モデル（最高値）は，0.6002 となっており，翻訳の精度の向上が見られた．

次の図 5 と図 6 の図は，構築されたモデルの追加学習を行った，反復フローの回数による BLEU 値の変化を表した図である．図 5 は追加学習の学習率が 0.1，図 6 は 0.05 である．全体的に評価した際に，ドメイン適応による BLEU 値の変化を見ると，一定の評価が上がっていることが分かる．

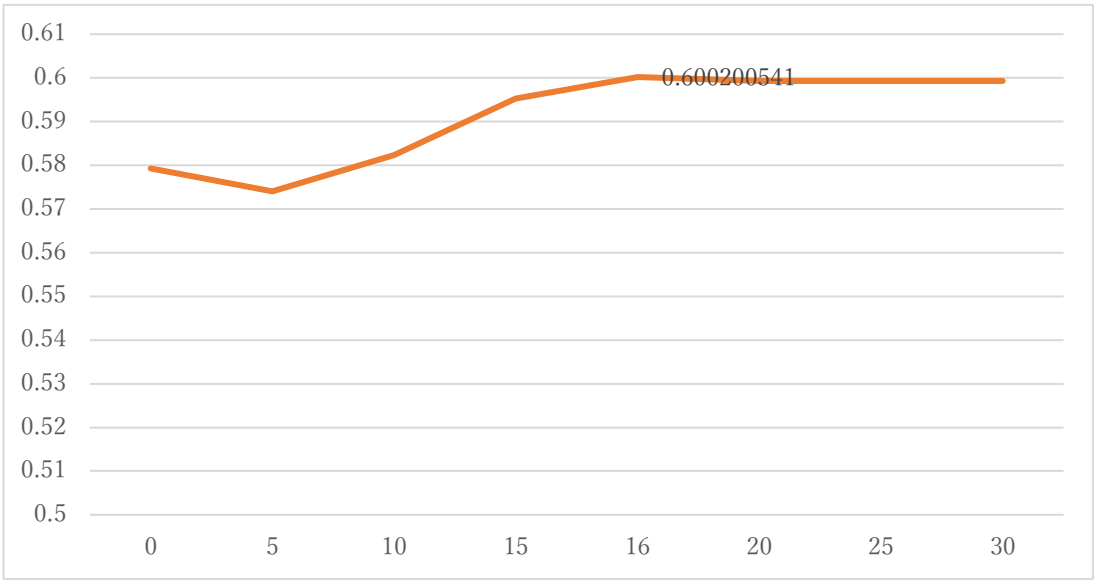

図５：学習率 0.1 で追加学習を行った反復回数による BLEU 値の変化

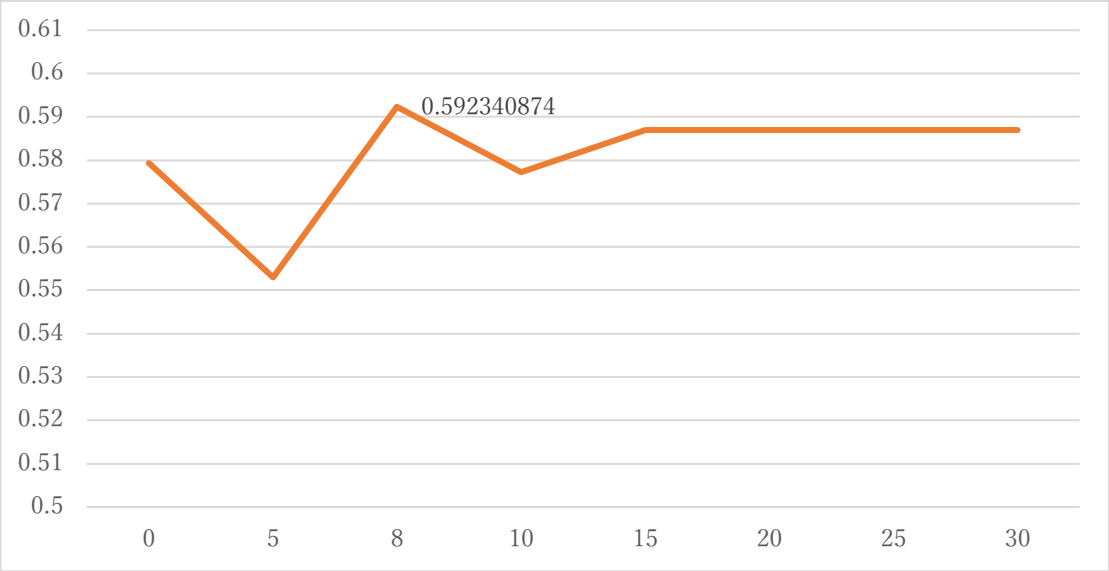

図 6：学習率 0.05 で追加学習反復回数による BLEU 値の変化

5.3.1 やさしい日本語の区別のパターンによる評価

やさしい日本語の区別のパターンによって，翻訳はどのようになっているのか検証し，評価を行う．

1. 語句の変換（語句数が変化しない）

入力文章が「その絵画は，複雑なテクニックを用いて描かれている．」では，出力文章は「その絵は，難しい技術を用いて描かれている．」というように翻訳されており，「絵画」が「絵」に，「複雑なテクニック」が「難しい技術」というようになっている．表１０に実際の翻訳結果を表す．

このことから，今回の方法では語句数が変化しない語句の変換は，比較的正しく翻訳されることが分かる．

2. 語句の変換（語句数が変化する）

入力文章が「私の親友は，非常に緻密な性格を持っている．」では，出力文章は「私の仲の良い友達は，非常に多くの友達を持っている．」というように翻訳されており，「親友」が「仲の良い友達」となっている．次の「このレストランでは，地元の食材を活かした料理が提供されている．」では，「この食事をする店では，世界の石油を作った料理が提供されている．」と翻訳されている．「レストラン」は，「食事をする店」となっているが，後の文章はドメイン適応を行ったことによって，意味が通らない文章になる結果となった．表１１に実際の翻訳結果を表す．

3.　表現全体が変化する

例に挙げた「ご用件は何ですか？」,「土足厳禁です.」という 2 つの文章は，前者は,「ご連絡はなんですか？」となり,「ご用件」という語句が「ご連絡」という語句に変換されており，正しく翻訳されていない．後者の「土足厳禁です.」という文章は,「赤ちゃんの中で遊んでいます.」という結果となっており，日本語の意味的にも全く異なる文章が出力される結果となってしまった．表１２に実際の翻訳結果を表す.

このように，表現全体が変化するやさしい日本語への翻訳は，今回の研究で構築したモデルでは，翻訳ができないことが分かった.

表１０：実際の翻訳結果

入力文章	出力文章
その絵画は，複雑なテクニックを用いて描かれている.	その絵は，難しい技術を用いて描かれている.
その映画は，人間の心理を巧みに描いている.	その映画は，人間の心理をうまく描いている.
彼女はその事故について詳細な報告をした.	彼女はその事故について詳しく報告した.
彼はその問題について深く考察した.	彼はその事件について深く考えた.
その国は，経済的に困難な状況に直面している.	その国は，経済的に難しい状況に困っている.

表１１：実際の翻訳結果

入力文章	出力文章
私の親友は，非常に緻密な性格を持っている.	私の仲の良い友達は，非常に多くの友達を持っている.
このレストランでは，地元の食材を活かした料理が提供されている.	この食事をする店では，世界の石油を作った料理が提供されている.

表１２：実際の翻訳結果

入力文章	出力文章
ご用件は何ですか？	ご連絡は何ですか？
土足厳禁です.	赤ちゃんの中で遊んでいます.

第6章 考察

本研究では，日本語大規模データを併用したことにより，やさしい日本語コーパスの語彙数を拡大することは成功した．日本語からやさしい日本語に翻訳する際に，言語が同じため大部分は入力文のまま出力するものの，一部をやさしい用語や言い回しに変換する必要があったため，やさしい日本語コーパスのみで，ドメイン適応を行った．しかし，翻訳の精度は一定の精度向上は見られたものの，BLEU 値では，0.6002 までしか上がらない結果となった．

実際のやさしい日本語への翻訳では，単語数が変化しない変換が，比較的翻訳の結果が良くなる結果となった．翻訳結果では，「その絵画は，複雑なテクニックを用いて描かれている．」という日本語は，「その絵は，難しい技術を用いて描かれている．」というやさしい日本語の翻訳になった．

しかし，単語数が変化する変換では，一部は翻訳ができているものの，その後の翻訳が悪い結果となった．翻訳結果では，「私の親友は，非常に緻密な性格を持っている．」という文章が「私の仲の良い友達は，非常に多くの友達を持っている．」という翻訳になった．ドメイン適応がうまくいかなかった原因としては，「親友」が「仲の良い友達」に翻訳された時に，語句数の変化があったことにより，次の語句を予測するときに間違った翻訳が出力されたと考えられる．

さらに，表現全体が変化する変換では，翻訳ができない結果となった．実際には，「土足厳禁です．」という文章が，「赤ちゃんの中で遊んでいます．」という翻訳になった．本来であれば，「靴を脱いでください．」という文章に翻訳されるのが正解であるが，意味が通らない結果となった．翻訳ができない原因は，今回の翻訳モデル構築で利用したやさしい日本語コーパスの中に，表現全体が変化するデータがなかったことが原因と考えられる．

翻訳精度を上げるためには，いくつか策が考えられる．それは，やさしい日本語コーパスの拡張である．やさしい日本語コーパスは，統合させたとしても約８万文しかなく，日本語の文章を翻訳するにはデータが少なすぎるため，ドメイン適応を行う時にも，すべての表現を翻訳できることはなかった．もし，やさしい日本語コーパス自体を拡張することができれば，どの方法を行うにしても，役に立つことであると考える．さらに，表現全体が変化するやさしい日本語の変換は，翻訳ができない結果となった．この問題を解決するためには，表現全体が変化するやさしい日本語への変換は，定型文である場合が多いため，いくつかの用例翻

訳を準備することで，そこから対訳を探す案も有効的だと考える．

第7章 おわりに

本研究では，Transformer に基づくやさしい日本語翻訳モデル生成の手法を提案し，日本語からやさしい日本語への翻訳の精度向上を行った．しかし，やさしい日本語には，大規模なコーパスや具体的な定義も存在せず，やさしい日本語の翻訳精度は向上の余地がある．現在は，グローバル化が進み，外国人の数が増えている中で，災害時などでやさしい日本語の注目度は高くなっている．現時点では，すべての語句をやさしい日本語には翻訳できておらず，人手の修正が必要であるのが現状である．すべての人が，やさしい日本語に簡単に翻訳することができれば，簡単に情報を出すことができるのではないかと考える．そのようにするためにも，やさしい日本語のニューラル機械翻訳の精度向上は不可欠なことである．

本研究の貢献は以下の通りである．

語彙数の拡大

やさしい日本語コーパスで学習した場合と日本語大規模コーパスを併用して学習した場合では，語彙数を増加させることに成功した．

さらに，出力と入力での未知語はやさしい日本語コーパスのみを用いた場合と比較して，90 件から 0 件まで削減することに成功した．

ドメイン適応

日本語大規模データを利用し構築された汎用的な日本語を出力できるモデルと，やさしい日本語コーパスのみで学習を繰り返す反復フローによって，ドメイン適応を強化した場合では，BLEU スコアは，0.5793 から 0.6002 向上させ，21 万回の追加学習を行うと，BLEU 値に変化がないことがわかった．

謝辞

本研究を行うにあたり，熱心なご指導，ご助言を賜りました指導教官の村上陽平教授に深謝申し上げます．また普段からお世話になっている社会知能研究室の皆さまにも感謝の意を表します．

参考文献

［1］ 松田 陽子，前田 理佳子，佐藤 和之：災害時の外国人に対する情報提供のための日本語表現とその有効性に関する試論：日本語科学：７：pp. 145-159(2000-04-15)

[2] 山本 和英，丸山 拓海，角張 竜晴，稲岡 夢人， 小川 耀一朗，勝田 哲弘，髙橋 寛治長岡技術科学大学： やさしい日本語対訳コーパスの構築:言語処理学会 第 23 回年次大会 発表論文集:pp. 753-755(2017 年 3 月)

［3］ 中町礼文，梶原智之：事前訓練済み系列変換モデルに基づくやさしい日本語への平易化：情報処理学会第 83 回全国大会:2:pp. 607-608（2021）

［4］ 熊野正，後藤功雄，田中英輝 NHK 放送技術研究所：統計機械翻訳を用いたニュース文のやさしい日本語への自動変換：映像情報メディア学会年次大会講演予稿集: pp. 32D-2-（2015）

［5］ 庵 功雄：「やさしい日本語」研究の現状と今後の課題，一橋日本語教育研究，2 号，pp. 1 12（2013）