置賜地域介護現場における方言翻訳システムの研究

山形大学 工学部 情報エレクトロニクス学科 二瓶友岳

1. はじめに

昨今、日本の介護業界における人手不足は大きな課題であり、2025 年までに 37.7 万人の介護人材が不足すると言われている[1]。この課題に厚生労働省と経済産業省は、ロボット技術の介護利用における重点分野を策定し、「高齢者等とのコミュニケーションにロボット技術を用いた生活支援機器」もその中に含まれている[2]。医療・介護現場では、患者の症状を詳しく伝えるために方言の使用機会が多く、方言による発言をしっかり受け止めることが安心感や信頼感に繋がる。人間関係・信頼関係の構築のために、方言によるコミュニケーションは、超高齢化社会において注目されている分野である[10]。また、医療・介護現場の山形の方言に対して県外出身者や外国人はどのように方言を理解するかという調査では、県外出身者が 25.2% , 外国人が 83.1%の割合で理解ができないと示されている[3] [4]。方言の理解難易度に関しては、難しい方言として山形方言が 4 位・5 位というインターネット上の調査があり、多方言の機械翻訳に関する研究では、48 方言中 11 番目に翻訳難易度が高いとされている[7][8][9]。従って、人間も機械も理解することが難しい方言のうちの 1 つであると言える。

従って、コミュニケーションロボットが方言を認識できるようにすることや、県外出身の介護従事者が利用者との信頼関係の構築をしやすくするために、方言の機械翻訳を目指す。執筆者は、置賜地域においてコミュニケーションロボットを活用するプロジェクトに参加していることから、本研究では山形県の置賜地域で話される方言について機械翻訳のためのシステムについての研究を行う。

2. 置賜方言の特徴

山形方言は、村山・置賜・庄内・最上の 4 地域に大きく分類され、それぞれ地域特有の方言が使われている。置賜方言は、南東北方言に属すると言われている。地理的に隣接する関東方言や南東北諸方言の影響も及んでいる。以下に簡単に文法の特徴を述べる。[11]

2-1.動詞の特徴と活用

サ変動詞についてラ行五段化の傾向がみられる。活用語の主な特徴は、主として後続の助動詞による。基本的には共通語と同じであるが、ネ(打消)、ラセル(使役)などで特徴がある。[11]

2-2. 助詞の特徴と活用

格助詞は、主格・対格が省略される特徴があり、方向・受け手・場所・目的には、(サ)が使われる。接続助詞には、(カラ・コンダラ・コンジャ・ケンドモ・タテ)が使われている。終助詞には、(ケロ・ケンニガ・ダズ・ゴデ) が挙げられる。[11]

2-3. 濁音化・促音便化・省略

共通語で使われている語彙の一部が濁音化・促音便化・省略されることが頻繁に起きる。

3. 関連研究

3-1.形態素解析システム Sudachi [19] [20] [21] [22]

本研究で使用する形態素解析システム Sudachi は、2017 年に公開された新しい日本語形態素解析システムである。株式会社ワークスアプリケーションズのワークス徳島人工知能 NLP 研究所が開発した商用利用可能なオープンソースの形態素解析器であり、今後 10 年にわたり継続的に開発、更新していくことをプロジェクトの目標としている。主な特徴は Sudachi に標準搭載されているシステム辞書として同社が開発している大規模辞書の Sudachi Dict である。

[SudachiDict] 従来の形態素解析辞書では、継続的な更新や新語への対応が不十分などの理由から実用で使うことが難しかった。SudachiDict は、形態素解析器 Sudachi のための辞書であり、汎用的に使える大規模かつ高品質な言語資源を目指して開発されている。国立国語研究所による超大規模なコーパス「NWJC」を利用して学習した 258 億語規模のコーパスにて学習を実施し、数ヶ月に 1 度専門家により継続的に更新している特徴がある。また、正規化がサポートされていることも特徴の 1 つであり、送り仮名・字種・異体字・誤用といった表記の違いを正規化する機能がある。規模の大きさによって、small・core・full と 3 つの辞書が提供されているが、本研究では最も語

彙数が多い full を使用する。

4. 翻訳システム

置賜方言から共通語への翻訳処理は、次の流れで行う。

[step 1] 形態素解析

テキスト入力された文書を形態素に分割する。形態素の分割は SudachiDict と SudachiDict にない未知語を自由に追加できるユーザー辞書の情報に基づいて、Sudachi によって行われる。ユーザー辞書に置賜方言の情報を追加することで、置賜方言を含めて正確に形態素に分けることを行う。

[step 2] 語彙変換

形態素が方言の場合、辞書引きにより語彙変換を行う。語彙変換をするために、各々の置賜方言の形態素に対して対応する共通語の情報をユーザー辞書に追加する。

[step 3] 形態素結合

形態素を結合し、出力する。

[追加機能 : Step2+] 格助詞補完

名詞,代名詞の次に動詞,形容詞が続いた場合は、格助詞が省略されたとみなし、頻繁に省略される主格・対格・目的の格助詞（が・を・に）を補完する手法をとる。本研究では、Step2 の次の処理として、試験的にランダムで補完し、どの程度性能が変化したか評価することでより今後の研究に活かしたいと考え実装した。

5. 辞書

置賜地域で現在も使用されている方言と介護現場で使われている方言をインターネット上の個人ブログやサイト [11] - [18] を中心に収集を行い、収集した置賜方言の 1199 語に品詞情報を含む注釈付きコーパス(以下、置賜 Dict)を作成した。注釈の付与ルールは、unidic-mecab 2.1.2 と Sudachi ユーザー辞書作成方法[20]に準拠し、介護現場で使われることの多い活用語には優先的に活用情報を付与した。以下に簡易版を示す。実際には 1 つの単語あたり{0 見出し, 1 左連接 ID, 2 右連接 ID, 3 コスト, 4 見出し（解析結果表示用), 5 品詞 1, 6 品詞 2, 7 品詞 3, 8 品詞 4, 9 品詞（活用型), 10 品詞（活用形), 11 読み, 12 正規化表記, 13 辞書形 ID, 14 分割タイプ, 15 A 単位分割情報, 16 B 単位分割情報}の 17 情報を付与した。

見出し	連接 ID	品詞 1	方言情報	共通語
おしょうしな	5687	感動詞	方言	ありがとう
やめる	5160	形容詞	方言	痛い
あぐど	4785	名詞	方言	かかと

Fig.1 置賜 Dict の例

6. 結果と評価

6-1. 形態素解析結果

介護に関係する置賜方言を含む文章を 170 文収集し、形態素に分割した結果の一部を以下に示す。出力結果①は SudachiDict のみで分割した結果、出力結果②は置賜 Dict を追加して分割した結果とする。

入力文	出力結果①	出力結果②
おしょうしな	['お', 'しょうし', 'な']	['おしょうしな']
あだまやめる	['あ', 'だ', 'ま', 'やめる']	['あだま', 'やめる']

Fig.2 置賜 Dict の追加による分割結果の違い（一部）

6-2. 形態素の分割性能評価

形態素の分割性能について評価するために、それぞれに人手にて正解とする分割情報を付与し、出力結果と比較することでどの程度正確に分割が行えているかを精度と再現率にて計算し、その 2 つの調和平均である Fscore にて判断することとする。以下に、それぞれの計算式の定義を示すが、分割結果が過剰に分割されていると精度が小さくなり、分割不足があると再現率が低い数値となるため、それらの調和平均である Fscore を算出することで単語抽出の統一された評価基準として用いる [21]。

$$精度 = \frac{正しく抽出された単語数}{システムが出力した単語数} \quad -(6.1)$$

$$再現率 = \frac{正しく抽出された単語数}{正解の単語数} \quad -(6.2)$$

$$Fscore = \frac{2 * 精度 * 再現率}{再現率 + 精度} \quad -(6.3)$$

以下に、170 文それぞれを SudachiDict のみで分割した結果(以下、結果①)と置賜 Dict を追加して分割した結果（以下、結果②）の F 値の分布をヒストグラムで示した。

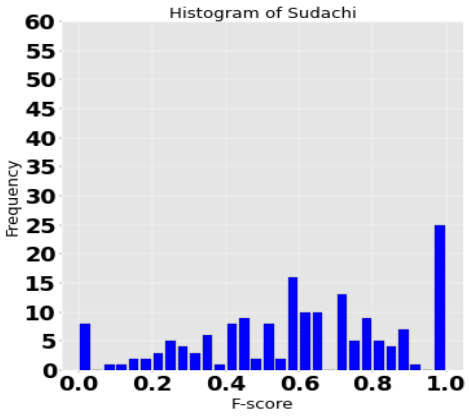

Fig.3 結果①の F 値の分布

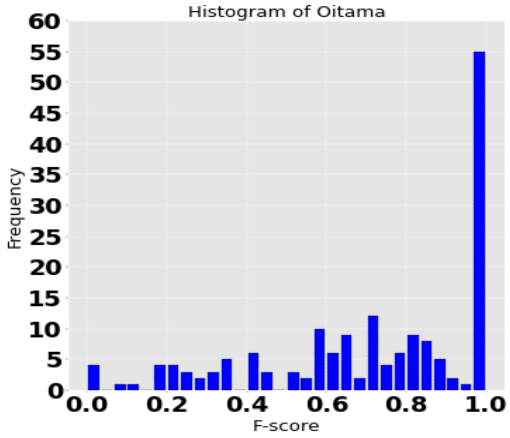

Fig.4 結果②の F 値の分布

	結果①	結果②	改善率
平均値	0.60	0.71	18.3%
中央値	0.60	0.77	28.3%

Fig.5 結果①・結果②の平均値・中央値と改善率

結果より、100%正しい Fscore=1 となる分割が行われた文が 2.2 倍となり、全体として平均値 18.3%，中央値 28.3%の改善が行われたことがわかる。

6-3. 置賜方言翻訳システムの結果

以下に、介護に関係する置賜方言を含む 170 文を第 4 章の方法にて翻訳を行なった結果の一部を示す。結果③は追加機能を含めない翻訳結果で、結果④は追加機能を追加して翻訳した結果とする。

入力文	結果③	結果④
あだまやめる	あたま痛い	あたまが痛い
いだいがら薬もらいだい	痛いから薬もらいたい	痛いから薬をもらいたい

Fig.6 置賜方言翻訳システムによる結果（一部）

6-4.a 品詞出力による自動評価

システムから出力された翻訳後の文章（以下、参照文）と正解とする共通語の文章（以下、正解文）がどの程度一致しているか評価をする。それぞれの文に対し、SudachiDict を元に形態素分割し、品詞情報を付与し、単語と品詞情報が一致した単語の数によって、正しく抽出されたと判断することとした。以下に定義の計算式を示す。前述した[6-2. 形態素の分割性能評価]と同様の理由から Fscore を使用する [21]。

$$精度 = \frac{正しく抽出された単語と品詞情報の一致数}{システムが出力した単語と品詞情報の一致数} \quad -(6.4)$$

$$再現率 = \frac{正しく抽出された単語と品詞情報の一致数}{正解の単語と品詞情報の一致数} \quad -(6.5)$$

$$Fscore = \frac{2 * 精度 * 再現率}{再現率 + 精度} \quad -(6.3)$$

以下に、結果③・結果④それぞれの F 値の分布をヒストグラムにて示し、追加機能を実装した場合の改善率も表した。

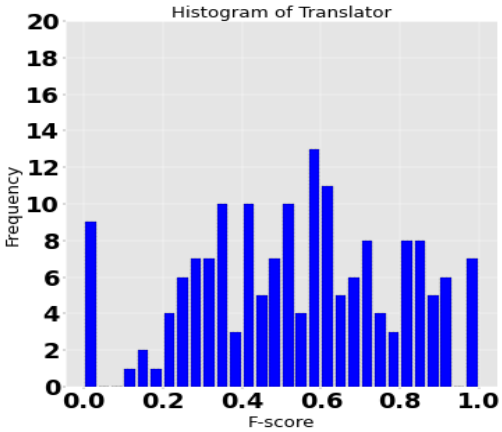

Fig.7 結果③の F 値の分布

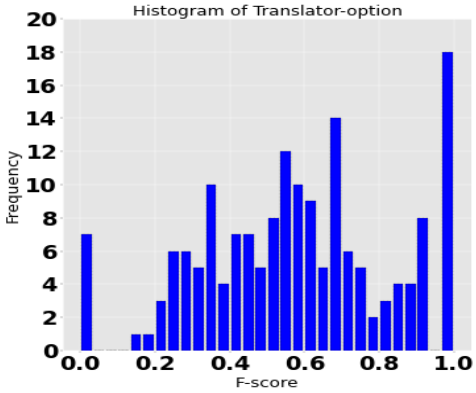

Fig.8 結果④の F 値の分布

	結果③	結果④	改善率
平均値	0.54	0.58	7%
中央値	0.55	0.57	4%

Fig.9 結果③・結果④の平均値・中央値と改善率

6-2.b 人手による評価

人手評価においては、文法性・共通語らしさ・意味の普遍さを人手によって判断するもので、日本語ネイティブ話者 1 名の評価者によって評価することとした。文法性・共通語らしさは、1~3 点(3 が満点)、意味の不変さは変化している(1)・変化していない(2)の 2 択で評価を行う。各々について平均値を算出し、以下に示す。また、どれだけ正しい翻訳が行えたかを調べるためすべての指標が満点の割合も併せて算出した。 [23]。

	文法性	置賜方言らしさ	意味の不変さ	正しい翻訳率[%]
結果③	2.49	2.22	1.74	32.9%
結果④	2.60	2.24	1.73	36.5%

Fig.10 人手による評価の平均値

格助詞の補完機能により、自動評価と人手評価の結果からそれぞれ精度が向上したといえる。正しい翻訳が行えている文が増えた一方で、本来格助詞がなくても意味が通じた文に、補完すべきでない格助詞が保管されてしまった文がいくつか見受けられた。そのため「意味の不変さ」を見ると格助詞の補完機能は、悪影響を受けていると言える。

7. まとめ

本研究では、介護現場で使われている置賜方言の例文を中心に、置賜方言を共通語へ変換するシステムを研究した。置賜 Dict の構築により形態素分割の性能が平均 28%向上し、機械翻訳による正解データとの精度・再現率からなる F 値は、0.55 となった。

現在の置賜 Dict の登録単語数は 1199 語であるが、置賜方言辞書[12]に収録されている語彙は 8000 語程度であるため、あと 7000 語ほど置賜 Dict に追加をすれば更なる精度向上が見込まれるが、全てに意味情報の付与を行うことは莫大なコストとなることが予想され、更にすべての動詞に活用情報を付与することは現実的ではない。本研究では、置賜方言翻訳システムの性能を向上させるための 1 つの施策として、格助詞をランダムに補完する手法を考案・実装し、F 値を平均 3%の向上がされ、有効であった。格助詞の補完については、前後の品詞関係からのみ欠落を予測して、ランダムで補完を行ったが、機械学習の手法によるアプローチを合わせることによって、ランダムではなく品詞以外の要素の前後関係からの予測による最適化が期待できる。第 2 章で述べた [2-3. 濁音化・促音便化・省略]について本研究では、頻繁に使われている語彙を辞書に登録することで、共通語への修正をおこなっていたが、発音する人によって濁音化などする基準が微細に異なるため、すべての語彙に対して、「濁音化・促音便化・省略」の有り得るすべての組み合わせを辞書へ追加することは困難であった。このような課題も含め、置賜方言の特徴を捉え、機械学習によるアプローチを行うことにより、低コストで方言翻訳の性能向上を目指すことができると考えるため、今後の研究に期待したい。

[謝辞]

置賜方言ネイティブ話者としてご協力いただいた指導教員の横山道央准教授、形態素解析器 Sudachi を開発する株式会社ワークスアプリケーションズの開発者のみなさまに深く感謝いたします。

[参考文献]

[1] 厚生労働省　2025 年に向けた介護人材にかかる需給推計（確定値）について
[2] 厚生労働省老健局高齢者支援課,経済産業省製造産業局産業機械課, ロボット技術の介護利用における重点分野,平成 24 年 11 月制定 , 平成 29 年 10 月改定
[3] 後藤典子, 介護士が介護現場で使用している山形方言の特徴
[4] 後藤典子, 医療・介護現場の方言を外国人はどう理解するか ,日本語教育学会 161 巻 , 研究ノート, P42-P49
[5] 後藤典子, 医療・介護のための山形方言検索の工夫, 日本語教育方法研究会誌, Vol.22 No.1 , P60-P61
[6]公益社団法人国際厚生事業団,株式会社光洋スクエア,一般社団法人国際交流&日本語支援 Y「外国人のための 会話で学ぼう!介護の日本語 第 2 版: 指示がわかる、報告ができる」中央法規出版(2020-05-11)
[7] 方言が難しい都道府県ランキング　1 位に輝いたのは北か南か… , https://sirabee.com/2017/05/18/20161131919/ , 入手日 2021.12.06
[8]「方言」が難解すぎる都道府県ランキング , https://ranking.goo.ne.jp/column/6333/ , 入手日 2021.12.06
[9] 阿部香央莉,松林優一郎,岡崎直観,乾健太郎,ニュートラルネットを用いた多方言の翻訳と類型分析,自然言語処理学会発表論文集(2018 年 3 月)
[10]日高貢一郎(2007), シリーズ方言学 3 方言の機能, 岩波書店
[11] 平山輝男(1997),日本のことばシリーズ 6　山形県のことば,明治書院
[12] 菊池直, 読む方言百科事典　置賜のことば百科（上・下）笹原印刷
[13]SWING GIRLS おきたま応援サイト,おきたま弁講座, http://swinggirls.jan.jp/okitama.html , 入手日 2021.12.06
[14] 山形県 しあわせ子育て応援部 しあわせ子育て政策課, 置賜地方・置賜弁, やまがた子育て応援サイト, https://kosodate.pref.yamagata.jp/dialect/okitama , 入手日 2021.12.06
[15] おきたま弁講座, SWING GIRLS おきたま応援サイト, http://swinggirls.jan.jp/okitama.html , 入手日 2021.12.06
[16] 投稿米沢の方言, http://ayrtonsports.com/hougen/hougen.html 入手日 2021.12.06
[17] ひとりごとダイアリー・アーカイブス,方言・米沢弁・米沢でしゃべらっちぇきた言葉集　おきたまのラジオマン編 , https://okitama-npo.sakura.ne.jp/, 入手日 2021.12.06
[18] 米沢弁 方言集 691 ワード, http://www.omn.ne.jp/~yamada-s/hoogen/yonezawa.htm,入手日 2021.12.06
[19] Sudachi: a Japanese Tokenizer for Business, Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida and Yuji Matsumoto
[20] Large Scale Dictionary Development for Sudachi ,Proceedings of Language Resources Workshop ,Miho SAKAMOTO,Noriko KAWAHARA,Sorami HISAMOTO,Kazuma TAKAOKA,Yoshitaka UCHIDA
[21]WorksApplications , https://github.com/WorksApplications/ , 入手日 2021.12.06
[22] 工藤拓(2018) 「実戦・自然言語処理シリーズ 第 2 巻 形態素解析の理論と実装」近代科学社
[23] 長谷川駿, 事前学習と汎化タグによる方言翻訳の精度向上