平成 22 年度第 6 回情報処理学会東北支部研究会（山形大学）
資料番号 10-6-B33

年代差を考慮した方言翻訳システム

柴田直由　　横山晶一　　井上雅史
（山形大学工学部）

1. 研究概要

方言はその地方の人間との会話をより円滑にし、より良い人間関係を築くためには必要不可欠なものである。そこで、方言を理解する有効な手段として、方言機械翻訳システムを利用した学習が挙げられる。

平成 18 年度には、共通語から村山方言への機械翻訳システムのプロトタイプ[1]が構成され、平成 20 年度には付属語の拡張[2]による精度の向上、平成 21 年度には方言辞書の拡張[3]による精度の向上が図られた。これらの研究で、付属語、特に助詞や助動詞、または文末表現の方言は世代を通してほぼ共通して使われており、名詞や動詞といった方言が若い年代になるにつれて使われなくなっている傾向があることが判明している。また平成 21 年度の研究で、新しい形態素解析用辞書(NAIST Japanese Dictionary[4])を導入することにより、より正確な形態素解析が可能となった。

以上を踏まえ本研究では、先行研究である共通語から村山方言への機械翻訳システムについて、拡張と改良を行った。主に村山方言特有の言い回しや文末表現の改良を行うことと、年代別に訳し分けが可能になるよう従来のシステムの拡張を行った。その結果、より精度の高い方言翻訳文の生成が可能となった。今回実施したアンケートの結果から、約 84%の正答率が得られた。

2. 研究背景

2. 1. 村山方言概説

山形県は、庄内・最上・村山・置賜の四地域に分かれ、それぞれの地域で特有の方言が使用されている。

村山方言は、山形市を中心とした、山形県内陸部の中央に位置する村山地方の方言である。この地方は、歴史的な背景から、音韻や文法で南奥羽方言的ではあるが、語彙では北奥羽方言的な要素を持ち合わせたりする[5,6]。このため村山方言は、山形県の他の地方の方言である庄内方言(北奥羽方言)、置賜方言(南奥羽方言)とは語彙や文法で異なる面がある。

2. 2. 音声面での特徴

(1)有声化(濁音化)

村山方言では、語中や語尾にあるカ行・タ行に濁音が付き、ガ行ダ行の濁音に変化する。これは東北地方全体に見られる特徴である。この特徴により、共通語よりも濁音が多くなり、それが「ズーズー弁」と呼ばれる原因の 1 つになっている。

(例) スイカが食べ**たい**。 ⇒ すいがん食**だい**。
明日晴れ**だと**いいな。 ⇒ 明日晴れ**だど**いいな。

(2)促音化

村山方言では 3 音節からなる形容詞は促音化することが多い。また、語尾がラ行の動詞の場合、その語尾が促音化することが多い。

(例) この商品は**たかい**。 ⇒ この商品**たっがい**。
何**見る**と思う？ ⇒ 何**見っ**ど思う？

2. 3. 助詞

(1)主格「が」、「は」

村山方言では、主格を表す「は」、「が」は省略されることが多い。

(例) 雪**が**止んだ。 ⇒ ゆぎ_止んだ。

(2)場所・方向「へ」、「に」

共通語で場所や方向を表す「へ」、「に」などを、村山方言では区別せず「さ」で表す。

(例) 山**に**登る。 ⇒ 山**さ**登る。

この他にも、受け手「に」が変化した「がら」、目的「を」が変化した「ば」、逆接仮定「なら」が変化した「ごんたら」等、数多くの助詞がある。

2. 4. 文末表現・述語の文法的カテゴリー等

(1)否定表現

(例) 行か**ない**。 ⇒ 行が**ね**。

登**れない**。 ⇒ 登**らんね**。

(2)推量・意志・勧誘表現

(例) 買い物に行**こう**。 ⇒ 買い物に行**ぐべ**。

暗いから帰**ろう**。 ⇒ 暗いがら帰**っぺ**。

(3)義務・当然表現

(例) 勉強**しなくてはならない**。 ⇒ 勉強**さんなね**。

行か**なければならない**。 ⇒ 行が**んなね**。

その他にも、希望、依頼、禁止、尊敬、受け身など、数多くの言い回しや表現が存在する。

2. 5. 年代による方言の違い

先行研究より、若い世代、特に 20 代以下の年代の多くは、名詞や動詞、形容詞といった方言語彙を知らないということがわかっている[3]。一方で、活用形方言語彙の場合、語幹は共通語と同じ語を取るが、活用部分や活用型が村山方言文法に倣うようである。例えば共通語「持つ(タ行五段動詞)」の場合、村山方言では「たがぐ(ガ行四段動詞)」と言うが、若い世代では「持づ(ダ行四段動詞)」、というように活用部分が有声化し活用型も変わる。

以上の理由により、訳文の精度を向上させるためには、年代による方言語彙の訳し分けを行う処理が必要であるといえる。

3. 翻訳システムと実行例

3. 1. 処理概要

処理の流れは以下の 4 ステップで構成される。

Step1. 形態素解析

入力された文章を単語に分割し、品詞を付与する。

Step2. 語彙トランスファ

年代を考慮した辞書引きにより語彙変換を行う。

Step3. 意味解析

原言語のもつ意味内容に合致する語彙変換を行う。

Step4. 形態素構造生成

変換された単語を連結し、訳文を生成する。

同じ日本語内の翻訳であり、語順に相違が見られないため、構文解析などのプロセスは省略できる。基本的なシステムの流れは従来のシステムを踏襲した。翻訳処理システムの全体の流れを以下の図に示す。

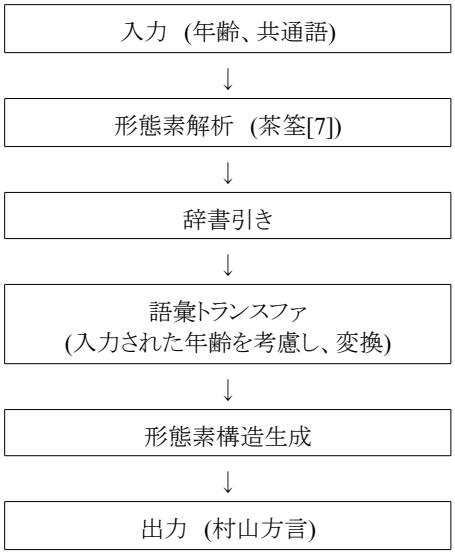

図 1 共通語から村山方言への翻訳処理の流れ図

3. 2. 語彙辞書

方言辞書に記述する基本パラメータは、「共通語語彙－見出し」「共通語語彙－読み」「方言語彙－見出し」「方言語彙－読み」である。活用語の場合はこれに加え、「基本形－見出し」「基本形－読み」「語幹－見出し」「語幹－読み」がパラメータとして記述してある。今回新たに、年代別に翻訳が可能になるよう、「年齢 1」「年齢 2」という年齢パラメータを基本パラメータとして追加した。辞書のデータ形式は、各値をカンマで区切ったデータ形式の CSV(Comma Separated Value)形式でまとめてある。現在は 1800 語程度の語彙が登録されている。

名詞辞書の一部を以下の表 1 に示す。なお、太字部分が今回新たに追加したパラメータを表わしている。

表 1 名詞辞書一部

共通語	共通語	方言語彙	方言語彙	**年齢 1**	**年齢 2**

語彙－見出し	語彙－読み	－見出し	－読み		
氷柱	ツララ	ぼんだら	ボンダラ	**50**	**200**
同じ	オナジ	おんなす	オンナス	**30**	**200**
尻尾	シッポ	おっぱ	オッパ	**30**	**200**

3. 3. 年代別語彙変換処理

共通語を入力する前に、年齢を入力するコマンドラインを追加した。辞書引きから語彙変換を行う際、この入力された年齢と、語彙辞書に記載されている年齢 1、年齢 2 のパラメータをそれぞれ比較する。「年齢 1 ≦ 入力年齢 < 年齢 2」という条件を満たす、共通語語彙に対応する方言語彙を検索する。存在するならばその方言語彙に変換し、そうでない場合は、変換を行わずそのまま共通語語彙として出力し、次の語彙変換処理に移る。

年代別に出力される例を以下に示す。

(例 1)

入力年齢：**20**

共通語　：ちょっと 財布を 探してくれ。

村山方言：**ちょっと 財布ば 探してくれ。**

(例 2)

入力年齢：**50**

共通語　：ちょっと 財布を 探してくれ。

村山方言：**ちぇっと 財布ば たねでけろ。**

4. アンケート評価

山形県立寒河江高等学校の 1、2 年生とそのご家族の計 1200 人を対象に、紙媒体で以下のようなアンケートを実施した。その中で回答が得られた 255 人分のデータを対象に評価・統計を行った。上記の高校は筆者の母校であり、村山地方の各市町から生徒が集まってくるため、方言の年代差、地域差の調査を行う対象としてふさわしいと考えたため、以上を対象として選択した。

4. 1. アンケート内容

共通語 26 文について、それぞれ比較的高齢の方が使うと思われる方言 1 と、比較的若い世代が使うと思われる方言 2 の 2 つの出力を用意し、その両方について「その方言が正しいかどうか」を問い、「Yes/No」で判断してもらう。「No」である場合や掲載した翻訳文以外の言い回しや表現等がある場合は、「その他」の欄に任意で答えてもらい、出力された方言が間違っている場合には、任意で出力文を添削して頂くというものである。このアンケートには本システムが出力した翻訳文を使用した。

4. 2. 評価と考察

総合評価として、各設問の正答率の平均値を計算し、約 84%という結果が出た。中でも方言 1 の評価から、方言語彙(特に名詞・動詞・形容詞)を使うにあたって年代差がみられた。特に、以下に記した設問 2 の内容が、年代差の現れた例である。この設問では、方言 1 の正答率が 50 代以上では 80%を超えるのに対し、30 代以下では 30%を下回る傾向がみられた。逆に、方言 2 では 30 代以下の正答率が 70%を超える傾向が見られ、方言語彙の使用有無の区切りがはっきり分かれた。

設問 2　共通語 (ちょっと 財布を 探してくれ。)

方言 1 (ちぇっと 財布ば たねでけろ。)

方言 2 (ちょっと 財布ば 探しでけろ。)

逆に、以下に記した設問 20 の内容は、年齢差が表れなかった例である。この設問では、方言 1、方言 2 の両方の正答率がどの年代を通しても 70%程度と高かった。これは、共通語形容詞「重い」の方言「おもだい」、動詞「持つ」の方言「たがぐ」を、年代問わず使用しているからだと考えられる。

設問 9

共通語 (無理して 重いものを 持たなくても 大丈夫だよ。)

方言 1 (無理して おもだいものば

たがぐんたって 大丈夫だよ。)

方言 2 (無理して 重いものば 持たねたって 大丈夫だよ。)

全体を通して、20 代以下では方言語彙の使用頻度が低いことがわかった。しかし、方言特有の言い回しや文末表現については若干の変化は見られるが、どの年代、地域でも共通して使用されている。これが村山方言の特徴として確固たる地位を築いているようである。

今回のアンケートだけでは、年代差、地域差を見積もる

情報が少ないので、さらに広くデータを収集するためにWeb 上でのアンケートを行う予定である。アンケート内容は今回行ったアンケートと全く同じ内容で、研究室のホームページよりアクセスして頂く形を考えている。

5. 問題点と今後の方針

アンケート結果より、村山地方の中でも細かなニュアンスの違いが非常に多く、方言語彙にあたる共通語の明確な決まりが存在しないため、一意的な翻訳は不可能に近いという問題が出た。精度向上を目指すのであれば、今後訳文候補を多数出力できるような処理等をシステムに組み込むことが必要となってくる。

また、方言の地域差について詳細なアンケート分析を行っていないため、今後アンケートについて特に地域差に着目した詳細な分析が必要である。

本システムでは、形態素解析の結果が正確であるということを前提に翻訳を行っている。そのため、形態素解析性能が優れている和布蕪[8]にシステムを移し変えることで、翻訳の精度向上が図れると考えられる。

今後の方針として、アンケート評価に基づき、現在語彙辞書に設けている年齢パラメータを正確な値に修正する必要がある。この際、年齢を絶対年齢表記で表すことで、時間遷移による年齢パラメータ値の変化を回避する。

本システムは、出力された訳文が方言話者に納得して頂けるような訳文になるよう、より自然で柔軟な翻訳結果を目指している。今後は本システムを基に、村山方言から共通語への翻訳システムの開発や、音声認識分野と結び付けることも視野に入れ、拡張に努めたい。

6. おわりに

本研究により、文末表現、方言特有の言い回しを増強するとともに、年代別に語彙の訳し分けを行うことができるようシステムを拡張することで、訳文の違和感を抑え、より柔軟な訳文を出力することが可能になった。

一方アンケートより、方言の活用語尾の変化、有声化や撥音化などは人により様々であり、共通語には置き換えられないニュアンスのみの方言や、活字には表せない方言が多々あることがわかった。これが方言の翻訳を困難にしている要因であると考えられる。

このことから、以上の問題を克服し、方言語彙、文法を完備させ、条件に合わせて柔軟な出力が可能であるシステムを利用することで、より効果的に方言を学習することが可能であるといえる。

参考文献

[1]尾形真美：共通語から村山方言への機械翻訳シス テム，山形大学卒業論文(2007)

[2]工藤翔陽：付属語に着目した村山方言翻訳システム，山形大学卒業論文(2008)

[3]兼子真弓：共通語から村山方言への機械翻訳システム，山形大学卒業論文(2009)

[4]Hideki YAMANE、Masayuki ASAHARA：形態素解析用辞書「NAIST Japanese Dictionary」(http://sourceforge.jp/projects/naist-jdic/memberlist)

[5]森下喜一：標準語引東北地方方言辞書，桜楓社(2008)

[6]平山輝男：山形県のことば、日本のことばシリーズ 6，明治書院(1997)

[7]奈良先端科学技術大学院大学：形態素解析システム「茶筌」

[8]京都大学情報学研究科-日本電信電話株式会社コミュニケーション科学基礎研究所共同ユニット：形態素解析システム「和布蕪」、http://mecab.sourceforge.net/ ：「和布蕪とは」という項目を参照(2010 年 12 月 1 日アクセス)