方言と標準語の違いを考慮した言語認識システムの開発

A Development of a Language Recognition System
which considers Difference between a Dialect and Standard Dialect

小林 聖也[*1]　　奥村紀之[*1]
Seiya Kobayashi　Noriyuki Okumura

[*1]国立長野工業高等専門学校 電子情報工学科
Nagano National College of Technology Department of Electronics and Computer Science

Most of present language recognition systems need standard dialect inputs. To construct more intelligent systems, we require functions which can reply to users' inputs flexibly even if dialects or accents are used. However, this system needs enormous number of data to tackle all of dialects and accents. Therefore, this research develops a system to a part of dialect. In addition, this paper entertains an expandability of proposal system.

1. はじめに

現在, 標準語を認識の対象とした言語認識システムは存在する. しかし, システムを使うユーザの中には, 綺麗な標準語を使う人もいれば, 方言やなまりといった言葉を使う人もいる. そのため, 標準語の意味を認識する言語認識システムがあることを前提に, 各地方独特の方言も認識の対象としていく必要がある. 本研究では, 方言の中でも長野県の方言に着目し, その意味を理解する言語認識システムの開発と評価を行う.

2. 言語処理技術の現状

2.1 形態素解析の問題

既存の言語認識システムの例として, 河岡らが開発したシステムを挙げる [1]. このシステムは, 入力された文や語の意味を理解するものであり, 認識の対象は標準語である. 例として感情判断 [2], 場所連想 [3] などのシステムが存在する. しかし, 入力文が標準語であったとしても, 形態素解析器の問題から, 認識が成功しないという問題がある. 感情判断システムにおいては, ある程度硬い文 (主述の関係があり, 目的語がはっきりしている文) を用いるなどの制約があった上での認識精度がおよそ80%程度である.

また, 本研究では, コンピュータ上で文における方言を認識し標準語へ変換することを目的としている. そのためには, 方言を含む文に対して形態素解析を行い, 文における方言を形態素として抽出する必要がある.

2.2 方言の調査について

現在, 方言に関する文献, ホームページなどは数多く存在する. そのひとつに国立国語研究所が制作した全国方言談話データベース [4] がある. これは村などの小さな地域別に方言の話し言葉としてのデータをまとめたものである. しかし, 本研究では, 方言の意味推測を目的としているため, 方言の単語としての情報に加えて, その語の頻度情報が重要となると考えている. そのため, アンケートを実施し, そのから実際に現在使われている方言, またその頻度を調査し, データベースを構築した.

連絡先: 奥村紀之，長野高専電子情報工学科，長野県長野市徳間 716，026-295-7133，noriyuki_okumura@ei.nagano-nct.ac.jp

3. 長野県の方言についての調査

15 歳から 20 歳までの男女約 200 名に対して, 長野県の方言に関するアンケートを実施した. その結果有効な回答は 130 部取得でき, 長野県の方言には次のような特徴が見られた.

- 独特の言い回しが存在する.
  - 「おやげない」,「かんます」など
- 標準語と同音の方言が存在する.
  - 「こわい」「ぼける」など

4. 各システムについて

本研究で提案するシステムの流れを図 1 に示す.

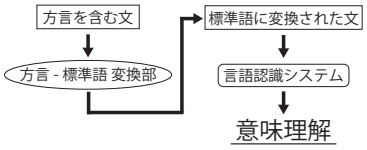

図 1: システムの処理の流れ

図 1 に示すように, 方言を含む文の入力に対して, 方言-標準語変換部により, 標準語のみで構成された文に変換を行う. その文を, 言語認識システムに入力文として与え, 出力結果を得る. 言語認識システムは, 認識の対象を標準語のみとしているが, 出力結果は方言を含む文の意味を認識した結果であると言え, 方言を含む文の言語認識システムとして成立する.

4.1 方言-標準語変換について

方言-標準語変換部では, 入力文に対して形態素解析を行い, 形態素と品詞に分割する. そして, 分割された形態素に方言があれば, 標準語との対応情報が記載されているデータベースから検索を行い, 方言を標準語に変換する. 変換後, 分割された形態素を結合し, 文として出力する. 解析器の辞書へ新たな単語を登録する際, ユーザが任意の情報を付与することができる. 今回, 方言には, それが方言であることを示すために, 品詞に加えて,「方言」という情報とその方言の標準語での意味情報を付与した. 方言を辞書に登録した後の解析形式は次のように出力される.

おやげない　　名詞, 方言, かわいそう
べちゃる　　動詞, 方言, 捨てる

検索によって方言に対応する標準語が見つかれば, そこで方言と標準語の変換を行う.

4.2 標準語と同音の方言の意味理解について

方言のアンケート結果 (3.) から, 標準語と同音の方言が存在することが確認されている. また, それらは同音であっても標準語とは異なった意味を持つ. 標準語と同音の方言の例を表 1 に示す.

表 1: 標準語と同音の方言の例

語	方言での意味	方言での用例
こわい	硬い	野菜がこわい
ぼける	すかすかする	りんごがぼける

語	標準語での意味	標準語での用例
こわい	恐ろしい	雷がこわい
ぼける	はっきりしない	論点がぼける

これらの場合, 独特の言い回しの方言とは違い, たとえ辞書に単語を登録したとしても, それが標準語の意味であるのか, 方言の意味であるのかを判断するのは難しい. そのため, 会話文という条件等の下で, 複数の文脈から意味を判断する必要があると考えられる.

「こわい」を例に, 標準語での「こわい」と方言の「こわい」, それぞれの意味で使われる会話の例を次に示す.

<例 1> (標準語の場合)
A: 私はこの野菜で食あたりになった。
B: あなたはこわい野菜を食べましたね。**(恐ろしいの意)**

<例 2> (方言の場合)
A: 私が食べた野菜は硬い。
B: あなたはこわい野菜を食べましたね。**(硬いの意)**

実際の処理の流れは図 2 に示す通りである. 図 2 に示す例は,<例 1>で提示した例文の場合の処理である. 例 1 の場合,A,B の文をそれぞれ感情判断システム [1] に与えると,「心配」という共通の感情が認識結果として出力される. 先述の通り, このシステムは認識の対象を標準語のみとしている. そのため,2 文の感情が一致したということは, 標準語での会話が成立していると推測でき, この文における「こわい」は標準語での意味として使われていると判断できる. また情緒の系図 [5] によると,「心配」の感情は,「恐怖」と類似した感情として定義されている. 文 A,B の認識結果が同じ感情であることから,B の文における「こわい」は標準語としての意味である「恐怖」を表していると推測できる.

次に,<例 2>で提示したような, 標準語と同音の方言が文に含まれていた場合の処理を, 図 3 に示す. 例 2 の場合,A,B の文をそれぞれ感情判断システムに与えると,A の文は「なし」,B の文は「心配」という感情がそれぞれ認識結果として出力される. ここでの言語認識システムは, 標準語を入力した場合のみ適切な判断を行う. そのため, 認識結果が異なるということは,「こわい」は標準語ではなく, 方言であると推測できる. そして「こわい」を含む文は,4.1 で説明した「方言-標準語変換部」に与え,「こわい」を標準語の意味である「かたい」に変換し, 出力する. 方言を標準語にした文に関しては再度感情判断システムに与え, 出力結果が入力文 A と一致または類似していることを確認する. これらの処理を行うことで, 標準語と同音の方言を含む文における, 変換の必要の有無を判断できる.

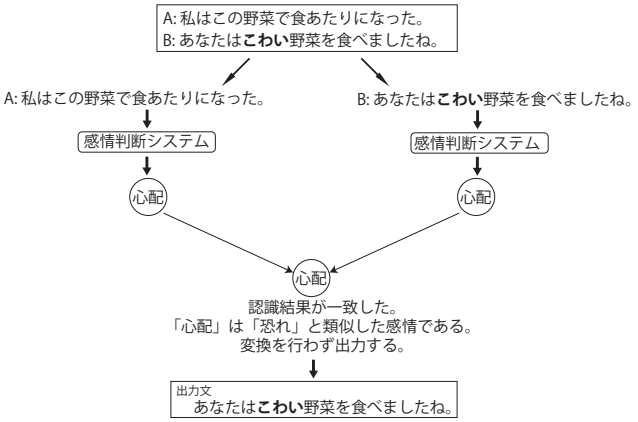

図 2: 標準語と同音の方言についての処理の流れ

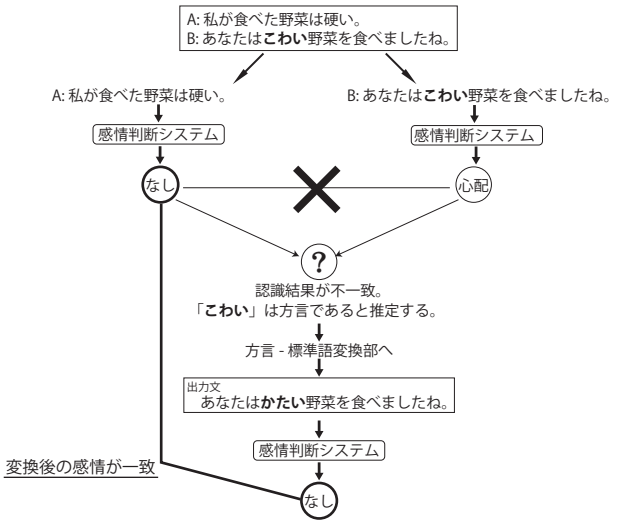

図 3: 標準語と同音の方言についての処理の流れ

5. 評価

MeCab により, 方言を含む文の形態素解析を行い, その結果から形態素解析の精度を検証した. 今回の調査によって得られた独特の言い回し 130 語, 標準語と同音異義の語 32 語の計 162 語について 1 文ずつ, 計 162 文の文を作成し入力文とした. 作成した入力文の例を表 2 に示す.

辞書に方言を登録する際に, 方言である語に対しては, それが方言であると明確になるよう「方言」という情報を付与した. 形態素解析の結果は 3 件出力するように設定した. 形態素解析が失敗である場合は, 方言が形態素として分割されていない. また形態素解析が成功である場合は, 方言が形態素として

表 2: 作成した入力文の一部

語	意味	文
ほんだら	そしたら	ほんだら明日行きます。
べちゃる	捨てる	ゴミをべちゃる。
かんます	かき混ぜる	牛乳をかんます。
おってな	一昨日	おってな友達に会った。
やまる	止む	雨がやまる。

分割され, 形態素に「方言」の情報が付与されている. 表示された結果 3 件のうち,1 件でも「方言」の情報が付与されている結果があれば成功とする. 形態素として分割されているものの,「方言」の情報が付与されていない場合は, 形態素解析は失敗であり, わかち書きが成功しているとする. これらの判断は目視により行う. この検証を, 方言を辞書に登録する前, 登録した後にそれぞれ行った. 方言を辞書に登録した後の「ほんだら」を含む文の形態素解析結果を示す.

```
ほんだら明日行きます。
ほんだら        *, 方言
明日    名詞, 副詞可能,*,*,*,*, 明日, アシタ, アシタ
行き    動詞, 自立,*,*, 五段・カ行促音便, 連用形, 行く, イキ, イキ
ます    助動詞,*,*,*, 特殊・マス, 基本形, ます, マス, マス
。      記号, 句点,*,*,*,*,。,。,。

ほんだら        *, 方言
明日    名詞, 副詞可能,*,*,*,*, 明日, アシタ, アシタ
行き    動詞, 自立,*,*, 五段・カ行促音便ユク, 連用形, 行く, ユキ,
ユキ
ます    助動詞,*,*,*, 特殊・マス, 基本形, ます, マス, マス
。      記号, 句点,*,*,*,*,。,。,。

ほんだら        *, 方言
明日    名詞, 固有名詞, 地域, 一般,*,*, 明日, アケビ, アケビ
行き    名詞, 接尾, 地域,*,*,*, 行き, イキ, イキ
ます    助動詞,*,*,*, 特殊・マス, 基本形, ます, マス, マス
。      記号, 句点,*,*,*,*,。,。,。
```

解析結果より, 方言「ほんだら」が形態素として認識され,「方言」の情報が付与されている. この場合は形態素解析が成功している.

標準語と同音異義の方言については, 分かち書きは成功する. そのため, それが標準語であるのか, 方言あるのかは, 辞書に「方言」の情報を付与することで判断する. ここでは「ぼける」を例に説明する.

```
りんご  名詞, 一般,*,*,*,*, りんご, リンゴ, リンゴ
が      助詞, 格助詞, 一般,*,*,*, が, ガ, ガ
ぼける  動詞, 自立,*,*, 一段, 基本形, ぼける, ボケル, ボケル
。      記号, 句点,*,*,*,*,。,。,。

りんご  名詞, 一般,*,*,*,*, りんご, リンゴ, リンゴ
が      助詞, 格助詞, 一般,*,*,*, が, ガ, ガ
ぼける  *, 方言
。      記号, 句点,*,*,*,*,。,。,。

りんご  名詞, 一般,*,*,*,*, りんご, リンゴ, リンゴ
が      助詞, 接続助詞,*,*,*,*, が, ガ, ガ
ぼける  *, 方言
。      記号, 句点,*,*,*,*,。,。,。
```

解析結果の 2 件目,3 件目での「ぼける」に「方言」の情報が付与されていることが分かる. この場合は「ぼける」の方言としての形態素解析が成功している.

5.1 独特の言い回しの方言について

独特の言い回しの方言について, 辞書登録前後の形態素解析の精度を図 4 に示す. 図 4 より, 辞書登録前は形態素解析・わかち書き共に精度は 0%である. 辞書登録語は, 形態素解析・わかち書き共に 100 %の精度を得た. この結果から, 辞書に未定義であった語で独特の表現の語を新たに登録することで, その語の形態素解析が成功することが確認できる.

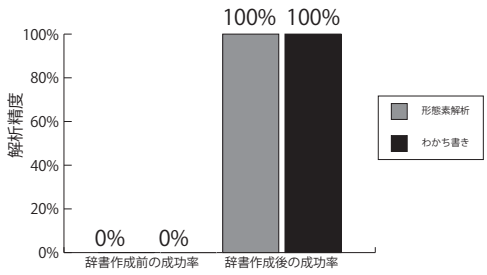

図 4: 辞書への登録前後での形態素解析精度の比較

5.2 標準語と同音異義の方言について

標準語と同音異義の方言について, 辞書登録前後の形態素解析の精度を図 5 に示す.

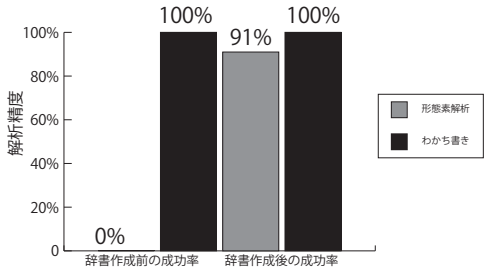

図 5: 辞書への登録前後での形態素解析精度の比較

図 5 より, 標準語と同音異義の方言については, 同じ音の標準語が存在するため, 辞書登録前はわかち書きの成功率が高い. しかし, 同音の標準語とは異なる品詞・意味を持つため形態素解析の精度は 0%である. 辞書へ登録後の解析結果は, 登録の際に品詞情報も付与しているため, 形態素解析の精度が 91 %まで上がった. 形態素解析が失敗した語には,「えらい」「しみる」「くれる」があり, これらが同音異義語に含まれる.

ここで, 標準語と同音異義の方言を辞書に登録したことによる標準語への影響について検証した. 検証方法は, 同音異義語を標準語として使用した場合の文を作成し, その文を入力文とし形態素解析を行った. 入力文は同音異義語 32 語について 1 文ずつ, 計 32 文を作成した. 作成した入力文の例を表 3 に示す.

表 3: 作成した入力文の一部

語	方言の意味	標準語の意味	入力文
おれた	私たち	「折れる」の過去形	骨がおれた
ねった	寝た	「練る」の過去形	生地をねった
やらず	やりましょう	「やる」の否定形	やらずに帰る
だべ	ーでしょ	「だべる」の活用形	みんなでだべる
つる	運ぶ	「釣る」の原形	魚をつる
ぼける	スカスカする	「はっきりしない」の意味	輪郭がぼける
こわい	硬い	「恐ろしい」の意味	幽霊がこわい

検証方法は, 形態素解析の結果を 3 件出力するよう設定し, 作成した入力文の形態素解析を行う. 出力された 3 件の形態素解析の結果のうち,1 件でも同音異義語が標準語として認識されていれば成功とし, 3 件の出力結果において同音異義語が方

言として認識されている (「方言」の情報が付与されている) 場合は失敗とする.

標準語を辞書に登録する前の形態素解析の結果を次に示す. 入力文は「こんなに」の意味である方言「こげん」と同音である標準語「焦げん (焦げないの意味)」のを含む文「餅がこげん。」とする。

```
餅      名詞, 一般,*,*,*,*, 餅, モチ, モチ
が      助詞, 格助詞, 一般,*,*,*, が, ガ, ガ
こげ    動詞, 自立,*,*, 一段, 未然形, こげる, コゲ, コゲ
ん      助動詞,*,*,*, 不変化型, 基本形, ん, ン, ン
。      記号, 句点,*,*,*,*,。 ,。 ,。
```

解析結果から「こげん」が「こげる」の活用として認識されていることが分かる. この場合は, 形態素解析・わかち書き共に成功とする.

次に「こげん」を辞書に登録した後の形態素解析結果を示す.

```
餅      名詞, 一般,*,*,*,*, 餅, モチ, モチ
が      助詞, 格助詞, 一般,*,*,*, が, ガ, ガ
こげん  *, 方言
。      記号, 句点,*,*,*,*,。 ,。 ,。
```

この解析結果から,「こげん」が方言として認識されていることが分かる. この場合,「こげん」のわかち書きは成功しているが, 形態素解析は失敗である.

図 6 に, 標準語と同音の方言を辞書に登録後, その同音の標準語について形態素解析を行った結果を示す.

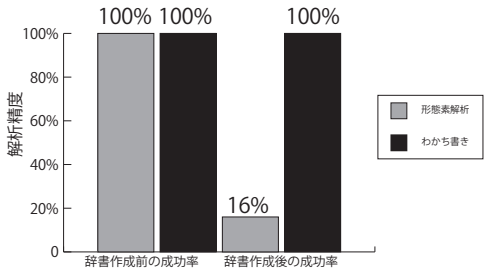

図 6: 辞書への登録前後での形態素解析精度の比較

図 6 より, 辞書登録前の結果はわかち書き・形態素解析ともに 100%であった. これは, 標準語に対して形態素解析を行っているため当然の結果であるが, 辞書へ同音異義の方言を登録した後の標準語における形態素解析精度は 16%と, 大幅に低下してしまった.

6. 考察

6.1 独特の言い回しの方言について

4.1 で述べた通り, 方言の形態素解析が成功することにより, 方言から標準語への変換が行えることを確認できた. また, 方言など標準語には存在しない語の形態素解析を行うためには, アンケートなどの調査により, その語についての辞書を充実させる必要があることが分かった. 一般ユーザからの情報収集を行うことで, 多くの語を得ることができ, また方言に関する新たな問題点を発見し, そのための解決策を検討できる.

6.2 標準語と同音異義の方言について

5.2 の図 5, 図 6 で示すように, 標準語と同音の方言を辞書に登録することで, 方言としての形態素解析の精度は向上するが標準語の形態素解析の精度は大きく低下してしまった. そのため,4.2 で述べた意味判断の手法により, 入力された同音異義語が方言であるか標準語であるかを判断・推測する必要性が高まった. また, 感情判断システムの精度に依存するものの, 主述の関係が明確である硬い文を用いることでの意味判断は行えるため, 今回提案した方法は正しいと考えられる. 本研究では, 入力文が表す感情を基に, 同音異義の語について方言であるか標準語であるかの判定を行った. しかし, 入力文に感情がない場合など, 感情判断システムでは同音異義語の意味推定は行えない. 言語認識システムとしては, 感情判断システムの他にも時間判断システム [6], 場所判断システムなど様々な分野に特化したシステムが存在する. これらのシステムを用いることで, さらに同音異義語における品詞・意味判定の精度が向上することが期待できる.

7. おわりに

本研究では, 方言を含む文の言語認識システムの開発について述べた. 長野県方言における独特の言い回しについては, 解析器の辞書を充実させることで, ほぼその意味を理解させることは可能であることを示した. また, 標準語と同音の方言については, 感情判断システムに大きな制約があり, それによって充実した検証を行うことができなかった. しかし, 硬い文を入力とすることでの意味判断は成功するため, この方法の有用性・可能性を示すことができた. 標準語を認識対象とした言語認識システムの精度向上に伴って, 方言の言語認識が更に実現化できると期待される.

また, このシステムを用いて音声認識システムへ発展させることも期待できる. しかし現在の技術では, なまりやイントネーションの認識が困難であるとされている [7]. これらの技術発展に伴い, 音声により方言を認識することができるようになれば, コンピュータと人間で音声により方言を用いた会話をすることも可能となる.

参考文献

[1] 土屋誠司, 渡部広一, 河岡司 : 常識的感情判断メカニズムの構築, 同志社大学理工学研究報告,Vol.43,No.1,pp.1-11,2002.4

[2] 土屋誠司, 吉村枝里子, 渡部広一, 河岡司：連想メカニズムを用いた話者の感情判断手法の提案, 自然言語処理,Vol.14,No.3,pp.219-238,2007.4

[3] 手原信太朗, 渡部広一, 河岡司：共起情報を用いた場所語未知語処理の精度向上, 電子情報通信学会 2008 総合大会講演論文集,D-5-4,2008.3

[4] 国立国語研究所：全国方言談話データベース 日本のふるさとことば集成, 国書刊行会,2007

[5] 九鬼周造：「いき」の構造，岩波書店，1991.

[6] 土屋誠司, 奥村紀之, 渡部広一, 河岡司：連想メカニズムを用いた時間判断手法の提案, 自然言語処理,Vol.12,No.4,pp.111-129,2005.10

[7] 河原達也, 李晃伸：連続音声認識ソフトウエア Julius, 人工知能学会誌,Vol.20,No.1,pp.41-49,2005