pl-17

生成AIを用いた鹿児島方言生成 ―日琉諸語の低資源言語・方言の生成に向けた試み―

坂井美日 (鹿児島大学)

1. はじめに

➢ 本研究の目的：**低資源言語・方言**の生成

➢ 本発表の試み：**鹿児島方言**（鹿児島市）の生成

- 日琉諸語に属す⇒語順や承接順序等は標準日本語等と共通。
- 語彙、音韻、文法面で固有の特徴。

(1) へがわっぜふってもーびんたがしれ、はよなけにぐっど。

(2) he//hai//=ga　wazze　fur-Q=de
灰=nom　非常に　降る-NPST=CSL
moo　binta=ga　sire//siro-i//
感嘆　頭=NOM　白い-ADJS
hajo//haja-u//　nake//naka=i//　nigu-Q=do
はやい-ADV　中=DAT　逃げる-NPST=SFP
（灰がひどく降るので、もう頭が白い。はやく中に逃げるよ。）

- 低資源かつ消滅危機方言。機会学習可能な言語資源は乏しい。
- GPT-4のzero-shot→精度ほぼ0%。

➢ 標準語＝**同系高資源言語**の活用＆**言語知識**の補填

機械翻訳の知見を応用！
- 高資源言語の活用有効(Zoph et al. 2016)
- 言語間の類似度が性能向上に影響 (Martínez-García et al. 2021等)

初チャレンジ！
- 現段階のLLMは、言語分析が苦手。

↓

言語知識を与えれば精度向上？

【本発表の結論】
日琉諸語の低資源方言たる鹿児島方言の生成は、
• 標準日本語と対訳形式で整備されたデータ
• 当該方言の言語知識の説明
を入力することで生成の精度が上がる。

2. 言語知識の入力は精度を上げるか

➢ 実験：母音融合現象の処理

➢ 形容詞2タイプ（カ系／イ系）のうち、イ系で当該現象。

➢ 母音が融合かつ短縮した形が表層形になる。

(4) 形容詞2タイプ（例：甘い）
カ系：あまか／イ系：あめ（//amai//→//ame:i//→/ame/）
(7) イ系に見られる規則
Ⅰ. //ai//→/e/、Ⅱ. //ui//→/i/、Ⅲ. //oi//→/e/、Ⅳ. //i:i//→/i/

➢ タスク20問を設定。GPT-4を使用。4パターンを検証。

Zero-shot…タスク(5)のみ

(5)次の意味の形容詞を鹿児島方言でどのように言うか答えて下さい。
1. ひどい、2. ながい、3. 古い、4. 美しい、5. 高い、6. 安い、7. 速い、8. 遅い、9. 苦しい、10. 熱い、11. 冷たい、12. 重い、13. 軽い、14. まぶしい、15. 明るい、16. 寂しい、17. 強い、18. 弱い、19. 悲しい、20. 白い
【条件】語尾が「か」で終わる形式以外で答えて下さい。つまり「ひどか」「ながか」等、語尾が「か」で終わる形式以外の言い方を答えて下さい

Few-shot…例示(8)→(5)

(8)鹿児島方言の形容詞の例を挙げます。
標準日本語=鹿児島方言
a. 甘い=あめ、b. 寒い=さみ、c. 固い=かて、d. 嬉しい=うれし、e. 太い=ふて、f. 清々しい=清々し、、g. 悪い=わり、h. 美味しい=おいし、i. えぐい=えぎ、j. 黒い=くれ、k. 痛い=いて、l. 細い=ほせ
それでは、上記の例を参考に、鹿児島方言の法則を分析しつつ、
(―以下(5) を入力―)

言語知識（音素ベース）…(9)→(5)

(9)鹿児島方言では、母音が連続すると母音が融合します。
(法則)
1. ai→e、2. ii→i、3. ui→i、4. ei→e、5. oi→e
この法則は、形容詞の語末にも適用され、発音に反映されます。
(例) あまい（amai）→あめ
嬉しい（ureshii）→うれし
寒い（samui）→さみ
黒い（kuroi）→くれ
（―以下(5) を入力―）

言語知識（仮名ベース）…(10)→(5)

(10)仮名ベース鹿児島方言の形容詞は、次の手順で作ります。
1. 語末の「い」を削除する。
2. 語末から2番目の仮名の発音を変更する。
- ア段ならばエ段に変更。
- イ段ならば変更なし。
- ウ段ならばイ段に変更。
- オ段ならばエ段に変更。
(例) 甘い（あまい：「ま」はア段）→あめ
嬉しい（うれし：「し」はイ段）→うれし
寒い（さむい：「む」はウ段）→さみ
黒い（くろい：「ろ」はオ段）→くれ
（―以下(5) を入力―）

正解：1 ひで、2 なげ、3 ふり、4 うつくし、5 たけ、6 やし、7 はえ、8 おせ、9 くるし、10 あち11 つめて 12 おめ、13 かり、14 まぶし、15 あかり、16 さびし、17 つえ、18 よえ、19 かなし、20 しれ

➢ 各パターン10回生成。正答1点／誤答0点で得点化。

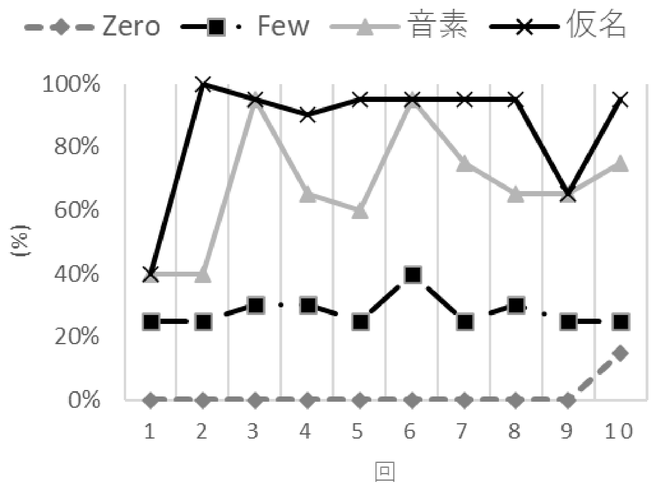

図1：母音融合の処理

回	Zero		Few		仮名		音素	
	正答	%	正答	%	正答	%	正答	%
1	0	0%	5	25%	8	40%	8	40%
2	0	0%	5	25%	20	100%	8	40%
3	0	0%	6	30%	19	95%	19	95%
4	0	0%	6	30%	18	90%	13	65%
5	0	0%	5	25%	19	95%	12	60%
6	0	0%	8	40%	19	95%	19	95%
7	0	0%	5	25%	19	95%	15	75%
8	0	0%	6	30%	19	95%	13	65%
9	0	0%	5	25%	13	65%	13	65%
10	3	15%	5	25%	19	95%	15	75%

【平均】
Zero：0.3（1.5%） Few：平均5.6（28%）
音素：13.5（67.5%） 仮名：17.3（86.5%）

- Zero：ほぼ0%（デフォルトでの生成は困難）
- Few：30%以下（例から言語法則を見出すのは不得意）

↓

Few <*1 言語知識入力（音素 <*2 仮名）

*1：Few vs 音素… t(18)=−6.405、p値約4.97e-06<0.05
*2：音素 vs 仮名： t(18)=2.242,p=0.0378<0.05

• 言語知識を与えることで、精度が向上。
• 仮名ベースの分析が精度が高い傾向。

【補足】動詞活用の処理

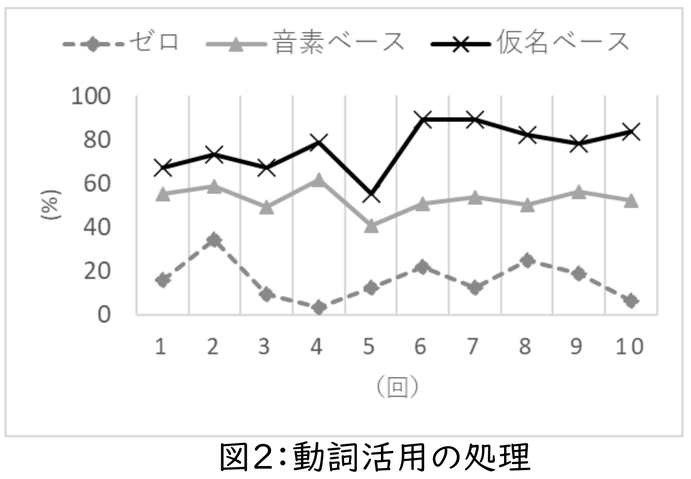

図2：動詞活用の処理

平均得点率
仮名：76.4%
音素：52.9%

やはり仮名ベースのほうが精度高い（p<0.05）

3. 鹿児島方言の生成

タスク「鹿児島方言で東京スカイツリーについて説明して」

➢ a. 会話文、b. 語彙集、c. 言語知識をプロンプト入力。

方言会話約20分。（スカイツリーの話はない）標準日本語と対訳セット。

方言固有語彙の標準語対訳120セット(aに出現)。動詞は活用種類と語幹情報。

文法概説。格、形容詞、動詞、文末、接続。

(16) 生成結果
東京スカイツリーは、東京のスミダクにあるどーC。このBタワーは、ふっとかで、世界ん中で一番高かタワーじゃった。スカイツリーんてっぺんから見っと、東京ん景色が一望できっと、よか眺めじゃらいね。特に夜ん景色は、光がきらきらしてB、きれいかどー。スカイツリーん中にはBおみせもずんばいあって、おきゃっさーが買い物Bしたり、食事したりできっと。また、そこにはB展望台もあってB、わっぜ高かところからB東京を見がでくっど。天気のよか日にゃ、富士山も見えっど。スカイツリーへの行き方は、電車で行くのがB便利じゃっど。東京スカイツリー駅や押上駅からちかいからB、そっから歩いていけっど。

評価	基準	点数
A	問題無し	1点
B	不十分（方言文脈の中にあっても良いが、より伝統的な言い方がある。）	0.3点
C	不自然（意味はわかるが、当該方言として違和感がある）	0.1点
D	非文（文法的、意味的に不可）	0点

➢ 59文節
➢ D判定なし
➢ 50.4／59点
➢ 精度85.4%

参考文献：木部暢子. 2001. 鹿児島方言に見られる音変化について. 音声研究. 第5巻第3号. pp. 42-48.／橋口満. 2004. 鹿児島方言大辞典（上）（下）. 高城書房／平塚雄亮. 2017. 鹿児島県鹿児島市方言. 方言文法 研究会編. 全国方言文法辞典資料集(3)活用体系(2). pp. 107-116.／Martínez-García Antonio, Badia Toni, and Barnes Jeremy. 2021. Evaluating morphological typology in zero-shot cross-lingual transfer. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3136-3153, Online. Association for Computational Linguistics.／Torrent Timponi Tiago, Hoffmann Thomas, Almeida Lorenzi Arthur and Mark Turner. 2023. Copilots for Linguists AI, Constructions, and Frames. Cambridge : Cambridge University Press.／Zoph Barret, Yuret Deniz, May Jonathan, and Knight Kevin. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1568-1575, Austin, Texas. Association for Computational Linguistics
謝辞：本研究は令和５年度鹿児島大学地域活性化研究支援事業の助成を受け、かつ科研費19K13213、22H00007、23H00007Jの成果の一部を反映したものである。

SAKAI LAB