深層学習と
日本語アクセント研究の可能性

東京医科歯科大学
野口大斗

自己紹介 (https://researchmap.jp/h_noguchi)

- 学歴
  - 2023年4月- 上智大学大学院 言語科学研究科 言語学専攻 博士後期課程
  - 2017年4月 - 2019年3月 東京大学大学院 総合文化研究科 言語情報科学専攻 修士課程
  - 2013年4月 - 2017年3月 同志社大学 文学部 英文学科
- 職歴
  - 2022年10月 - 現在 東京大学 教養学部 非常勤講師
  - 2022年4月 - 現在 国士舘大学 理工学部 理工学科 非常勤講師
  - 2021年4月 - 現在 江戸川大学 メディアコミュニケーション学部 情報文化学科 非常勤講師
  - 2020年4月 - 現在 千葉商科大学 国際教養学部 非常勤講師
  - 2020年4月 - 現在 工学院大学 教育推進機構国際キャリア科 非常勤講師
  - 2019年4月 - 現在 東京医科歯科大学 統合教育機構 (教養部) 非常勤講師
  - 2020年4月 - 2023年3月 国際医療福祉大学 成田保健医療学部 非常勤講師
  - 2019年10月 - 2021年3月 国立国語研究所 理論・対照研究領域 非常勤研究員

発表の流れ

1. 東京方言のアクセント
2. 言語学におけるアクセント研究
3. 言語処理におけるアクセント研究
4. 両分野で取り組まれてこなかったこと
5. これからの展望

1. 東京方言のアクセント(Kawahara, 2015)

- ピッチの高低（HL）パターン
- 単純語のパターン（東京方言）
  - a. ame（ga）　　LH(H)　　「飴（が）」
  - b. a’me (ga)　　**HL**(L)　　「雨（が）」
  - c. i’noti（ga）　　**HL**L(L)　　「命（が）」
  - d. koko’ro（ga）　　L**HL**(L)　　「心（が）」
  - e. atama’（ga）　　LH**H(L)**　　「頭（が）」
  - f. miyako （ga）　　LHH(H)　　「都（が）」
- アクセントの分布
  - 不均衡多クラス/long tail/正解ラベル複数

2. 言語学におけるアクセント研究 (Kawahara, 2015)

- ルールによる予測
  - 外来語
    - a. ku-ri-su’-ma-su　　LH**HL**L (-3)　　「クリスマス」
    - b. zya-a-na-ri’-zu-mu　　LHH**HL**L (-3)　　「ジャーナリズム」

    - c. pa-i-na’-p-pu-ru　　LH**HL**LL (-4)　　「パイナップル」
    - d. ka-re’-n-da-a　　L**HL**LL (-4)　　「カレンダー」
    - e. ma’a.ga.rin　　**HL**LLL (-5)　　「マーガリン」
    - f. a’.ma.zon　　**HL**LL (-4)　　「アマゾン」

2. 言語学におけるアクセント研究

- ルールによる予測
  - 複合語
    - a. undoo+kutu’ → un**do’o**+gutu　「運動靴」
    - b. hirosima+ke’n → hirosi**ma’**+ken　「広島県」
    - c. ko’o+ketuatu → koo+**ke’**tuatu　「高血圧」
    - d. kuti+yakusoku →　kuti+**ya’**kusoku　「口約束」

    - e. koosoku+ba’su → koosoku+**ba’**su　「高速バス」
    - f. ya’mato+nade’siko → yamato+na**de’**siko　「大和なでしこ」

- 制約による予測（Optimality Theory）…

2. 言語学におけるアクセント研究

- 課題
  - 予測力の低さ
  - 他の語種（和語・漢語）の取り扱い
  - 他の品詞の取り扱い
  - 意味の取り扱い

3. 言語処理におけるアクセント研究

- 未知語の予測
  - 人名の読み予測（Nakajima et al., 2005）
  - 複合語アクセント予測（Kuroiwa et al., 2007）
  - 読みを使った深層学習による予測（Tachibana & Katayama, 2020）

- その他
  - アクセント書き取りシステム（Bruguier et al., 2018）

- 課題
  - 表層形（品詞や意味、頻度）以外の情報の取り扱い

4. 両分野で取り組まれてこなかったこと

- 既知語の機械学習を用いた予測（Noguchi, 2022a）
  - 『NHK日本語発音アクセント辞典』（NKH hōsō bunka kenkyūjo, 2002）と『現代日本語書き言葉均衡コーパス短単位語彙表』（National Institute for Japanese Language and Linguistics, 2015）のデータを結合
  - モーラの構造と音節構造、語頭と語末から数えたアクセントの位置の付与（平板型のアクセントの位置は0）
  - アクセントパターンは、文字数と後ろから数えたアクセントの位置をアンダースコアで結合して表示
    - a. アイスコ’ーヒー（7_4）
    - b. アイスコーヒ’ー（7_2）
  - 書き言葉コーパスに出現しないものを除外
  - 同音異義語を除外

4. 両分野で取り組まれてこなかったこと

- 既知語の機械学習を用いた予測（Noguchi, 2022a）
  - 漢字の表記ゆれの問題から、かな表記で結合したため、同音のアイテムを除外
  - 1つの語しか所属しないアクセントパターンは除外
  - 文字数、品詞、音節構造、語種を用いてロジスティック回帰による多クラス分類
  - 数値以外のデータをワンホットベクトル化
  - 訓練データをオーバーサンプリング
  - シャッフルしたうえで、訓練データとテスト用のデータで正解ラベルの割合が等しくなるように設定し、75%対25%で分割
  - 音節構造を使ったところを文字とその出現位置に変えてパディングしたものでも確認
  - 正答率は5回の学習とテストの結果を平均

4. 両分野で取り組まれてこなかったこと

正答率		特徴量	
		言語学で指摘されているもの	音節構造の代わりに文字とその位置を入れたもの
データ	すべて	0.7015	0.9052
	和語	0.6516	--
	漢語	0.6443	--
	外来語	0.7107	--

4. 両分野で取り組まれてこなかったこと

- 意味の考慮（Noguchi, 2022b）
  - Noguchi (2022a)との違い
    - 同音異義語を追加（除外せず）
    - 語の分散表現（chiVe）(Hisamoto et al., 2020)を追加

4. 両分野で取り組まれてこなかったこと

正答率		特徴量		
		すべて	分散表現のみ	分散表現以外のみ
データ	同音異義語なし	0.87924	0.63401	0.69057
	すべて	0.79815	0.5696	0.63073

5. これからの展望

- 深層学習結果の解釈 (投稿中)
  - Attentionを利用したTransformerモデルを使用
  - それぞれのセグメントのその位置におけるアクセント予測への寄与度合い
  - 逆方向の翻訳（そのアクセントパターンに典型的なセグメントの並び）
- 東京方言と大阪方言を相互に予測できるか（投稿中）
  - 東京方言のアクセントパターンは大阪アクセントにほぼ影響なし
- （ユーザコンテンツからの）辞書作成
- 機械学習の仕組みと脳の仕組みの比較

参考文献

- Bruguier, A., Zen, H., & Arkhangorodsky, A. (2018). Sequence-to-sequence neural network model with 2D attention for learning Japanese pitch accents. *Interspeech 2018*, 1284–1287. https://doi.org/10.21437/Interspeech.2018-1381
- Hisamoto S., Yamamura T., Katsuta A., Takebayashi Y., Takaoka K., Uchida Y., Oka T., & Asahara M. (2020). chiVe: Towards Industrial-strength Japanese Word Vector Resources. *IEICE Technical Report*, *120*(166), 40–45.
- Kawahara, S. (2015). The phonology of Japanese accent. In *Handbook of Japanese phonetics and phonology* (pp. 445–492). De Gruyter Mouton.
- Kuroiwa, R., Minematsu, N., Den, Y., & Hirose, K. (2007). Accent labeling of a large-scale database by a single labeler and its use in statistical learning of accent sandhi. *Technical Report of IEICE*, *106*(614), 31–36.
- Nakajima, H., Nagata, M., Asano, H., & Abe, M. (2005). Estimating Japanese person name accent from mora sequence using support vector machines. *IEICE(D)*, *88*(3), 480–488.
- National Institute for Japanese Language and Linguistics. (2015). *Balanced corpus of contemporary written Japanese short unit vocabulary: Version 1.0*. http://doi.org/10.15084/00003218
- NHK hōsō bunka kenkyūjo. (2002). *NHK nihongo hatsuon akusento jiten* [NHK Accent Dictionary of the Japanese Language]. Nihon Hōsō Shuppan Kyōkai.
- Noguchi, H. (2022a, June 4). *Kikaigakushu wo mochiita tokyohogen tanjungo akusentopatan no yosokukanosei [The Predictability of Accent Patterns of Simplex Words in Tokyo Japanese Using Machine Learning]*. Spring Meeting 2022 (PhSJ).
- Noguchi, H. (2022b, August 24). *Word meaning and accent patterns in Tokyo Japanese: An examination using Word2Vec word embedding*. Phonology Forum 2022.
- Tachibana, H., & Katayama, Y. (2020). *Accent estimation of Japanese words from their surfaces and romanizations for building large vocabulary accent dictionaries*. https://doi.org/10.1109/ICASSP40776.2020.9054081