ぎなた読みの自動生成の試み

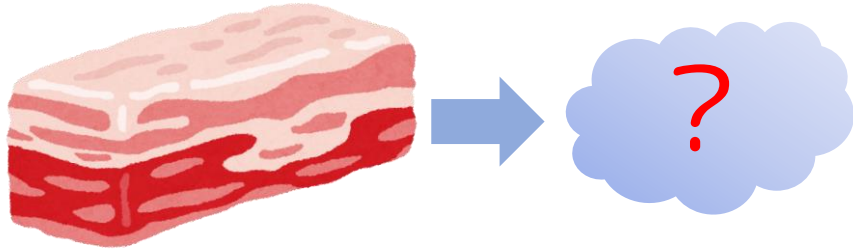
2018/3/16

@形態素解析WS(岡山コンベンションセンター)

林部 祐太 （無所属）

面白い形態素解析誤り

✓ |長いも|と|豚肉|の炒め物
✗ |長い|もと|豚肉|の炒め物



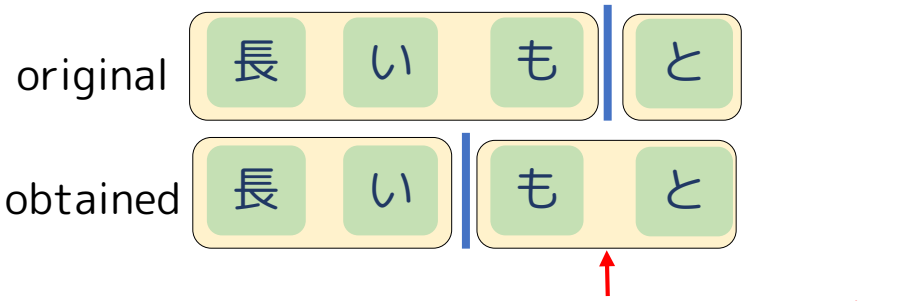
✓ 好き嫌いは|あまり|ない|ほう|です
✗ 好き嫌いは|あまり|な|いほう|です

面白い誤りをもっ
と堪能したい！

※本例は https://github.com/FairyDevicesRD/FairyMaCorpus より抜粋

ぎなた読み候補の生成

- ウェブコーパスから形態素n-gramを得る(n=2, 3, …)
- 頻度が一定以上あるn-gramから，異なる形態素分割を1つ得る

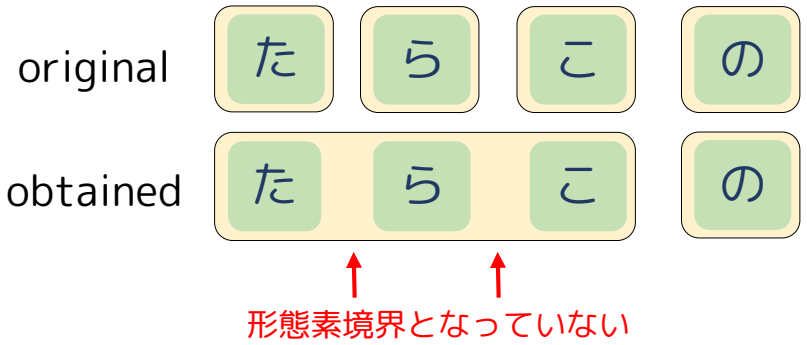


- 大量の異分割が得られる
  - 大半は不自然（高コスト）or大して面白くない
  - なんらかの基準でフィルタリングやスコアリングが必要

ぎなた読み…文意が変わってしまう形態素分割

異分割の簡易フィルタリングと加工

- 簡易フィルタリング
  - 元分割と異分割のコスト差が少ない
  - 異分割のみに「体-生産物-食料」や「体-主体-家族」が出現
    (分類語彙表を利用)
  - 人手での精選
- N-gramに人手で前後文脈を追加



例: 着膨れたらこの方法で対処してね

得られたぎなた読みの例1

- 外国ですしねぇ
  - |外国|です|し|ねぇ|
  - |外国|で|すし|ねぇ|

- エブリデイロープライス
  - |エブリデイ|ロー|プライス|
  - |エブリデイ|ロープ|ライス|

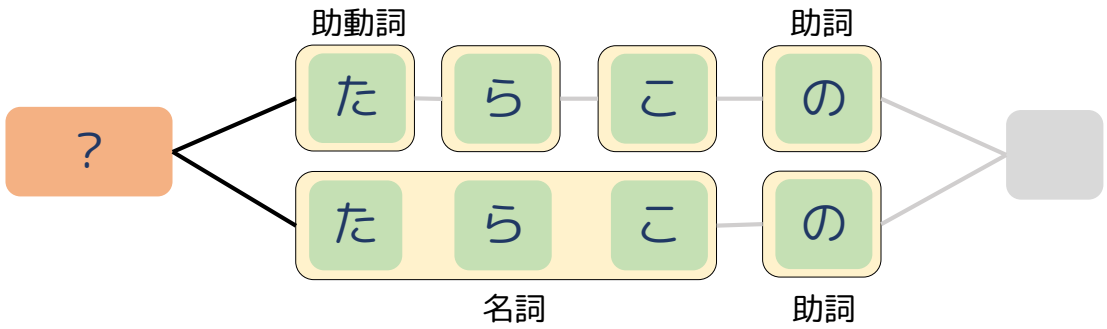
得られたぎなた読みの例2

- あったかいとこが好き
  - |あったかい|とこ|が|好き|
  - |あったか|いとこ|が|好き|

- 女の子連れ去った
  - |女の子|連れ|去った|
  - |女|の|子連れ|去った|

今後の課題

- 元分割と異分割の組にスコアを与える
  - 言語モデルを考慮する
  - 前後文脈の共有しやすさを考慮する



- 前後文脈の候補を自動生成する

まとめ

- ぎなた読みを自動で作ってみた
  - 楽しい
  - 制約付き解析の利用
- 今後の課題
  - コスト計算の改良
  - 前後文脈の候補自動生成
- hayashibe.jp
  - 詳細な実験情報・データ・プログラムなど
  - 求人のお問い合わせ等もこちらへ

Appendix

関連研究: 類音文変換システム[金久保13]

- 類音文への自動変換
  - それ見た事か　　→ それに蛸とか
  - 真っ赤な嘘　　→ まっカワウソ
  - おくびにも出さない　　→ おう,ビキニもダサい
- システムの概要
  - ルールベースの形態素解析
    - 形態素候補の制限
    - 形態素間・品詞間接続規則の制限
  - 類音も形態素候補に加える

参考文献

- [金久保13]「形態素解析手法と通俗的単語群に基づく類音文変換システム」情報処理学会論文誌，Vol.54, No.7, pp.1937-1950, 2013
  [Morita+15] "Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model", EMNLP 2015
- [林部17]「日本語部分形態素アノテーションコーパスの構築」, 情報処理学会第231回自然言語処理研究会, NL-231-9, pp.1-8, 2017