seq2seqによる部首の意味を考慮した漢字生成システム

藤井 諒[1*]○，舟山 弘晃[1]，北山 晃太郎[1]，阿部 香央莉[1]，Ana Brassard[2,1]，三田 雅人[2,1]，大内 啓樹[2,1]
* r-fujii@ecei.tohoku.ac.jp 1. 東北大学 2. 理研AIP

概要

- 表意文字である「漢字」が持つ**部首**の意味を考慮するシステムを作成
- 「へん」「つくり」部分を生成するモデルを**別々に学習**
- より質の良い漢字生成のための分析

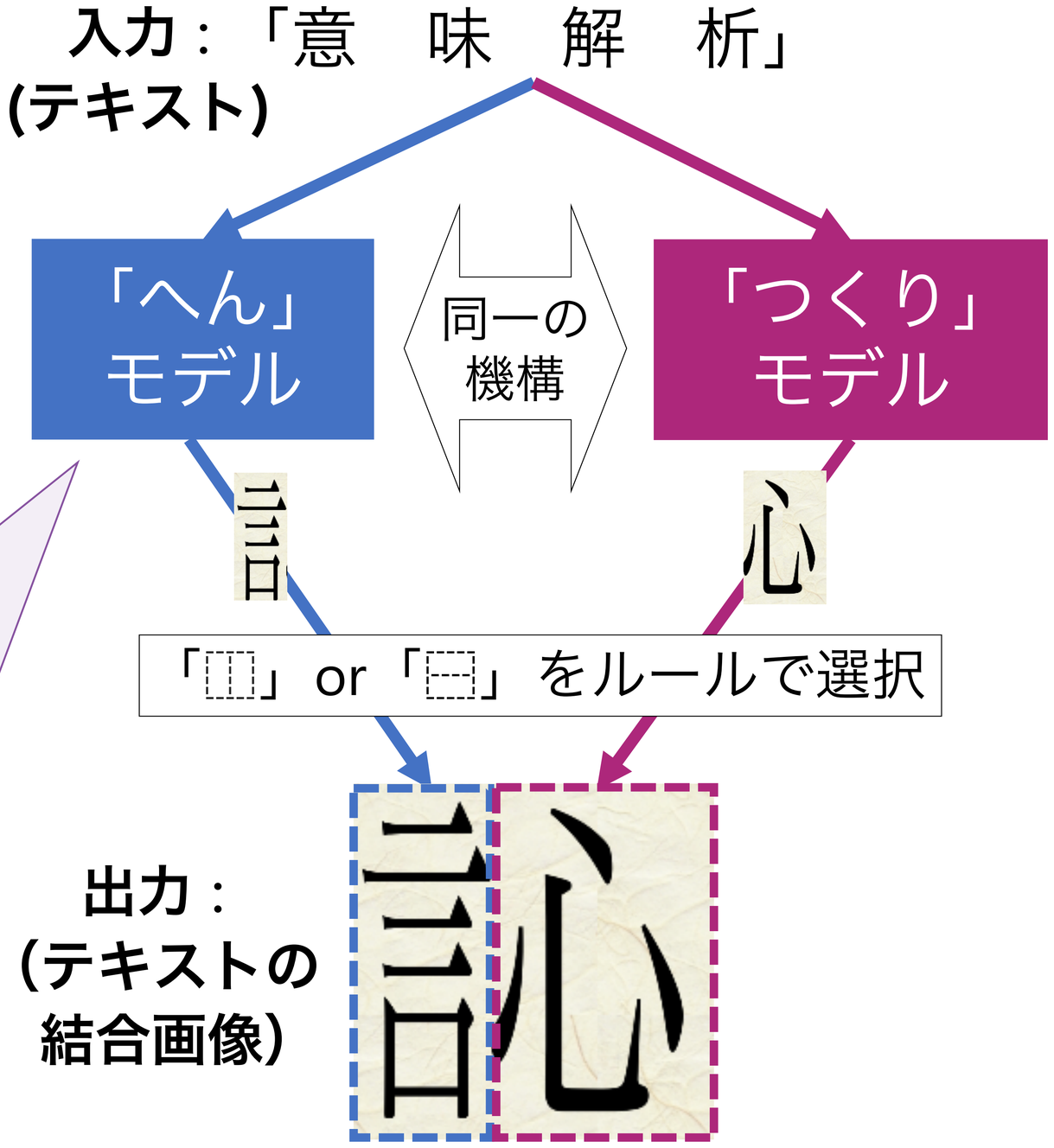
学習用データ

- 常用漢字1文字についての(辞書的意味, 構成要素) の対をseq2seqで学習
  例 : (みだりに, ⿰舌し)
- 二字熟語の擬似データ追加
  例 : (<Two_C> 一生涯, 一 <sep> ⿰イ弋)
  ↑ 擬似データを表すタグ

モデル

- fairseq[Ott, 2019]* のLSTMを使用
  * https://github.com/pytorch/fairseq

モデル



- へん・つくりモデルの出力を結合
- ビームサーチの上位6文字も提示

分析

① 出力の近傍事例

Q. 訓練データ中のどの例が出力の根拠となったか？
→ 出力と類似度の高い訓練事例を編集距離に基づいて検索

「言」を「へん」に持つ漢字の近傍事例

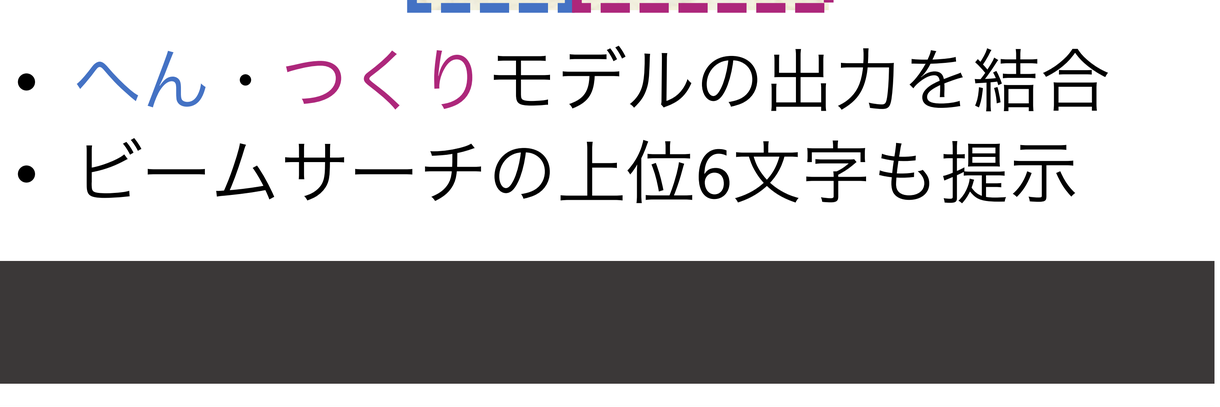
類似度	入力 :「意味解析」の近傍事例
0.375	**意味**を解きほぐす
0.286	**意味**、また理由
0.286	**意味**を読み取る

「心」を「つくり」に持つ漢字の近傍事例

0.500	**意味**
0.143	**意**図するところ
0.125	<Two_C> おもな**意味**や考え

② Attention可視化

Q. 2つのモデルがどこを見ているか？
→ Attentionを確認



各モデルのAttention可視化 (左 : へん, 右 : つくり)

今後の課題

- 「へん」：形態素 / BPE単位の事前学習済みベクトルの利用検討
- 「つくり」：漢字の読み（音）を取り入れたモデルの検討