文脈化埋め込みを用いた言語学習者のための語義別例文検索システム

ジョシュア・ターナー（東大生研・ワシントン大学)、吉永直樹（東大生研）

概要

- 既存の辞書は数多い語義の中から対応する語義を学習者自身で見つけないといけないため、第二言語の語彙学習には使いづらく不便
- 文中の単語に対し、文脈に応じた語義と他の語義を表す用例集合を学習者に提示することで語彙学習を簡易化するシステムを提案

研究背景

言語学習において意味のわからない単語に遭遇することは多い

I joined theater club so I could [**play**] Romeo.

Merriam-Webster[1]でのplayの定義文は動詞だけで60以上ある

1. a) “to engage in sport…”
   b) …
   ⋮
3. a) ...
   b) (1) “to act in a dramatic…” ←
      (2) ...

16番目の3-b-1は見つかりづらい

既存の辞書のよくある問題：

- 対応する語義を見つけるのが困難
- 定義文や限られた用例が語義の説明として不十分

着想：実データから語義の理解に役立つ用例を抽出できないか？

- "...Jha to play yoga guru in TV series."
- "In addition, the actor played in cine-magazine."

1.https://www.merriam-webster.com/dictionary/play

語義別用例検索システム

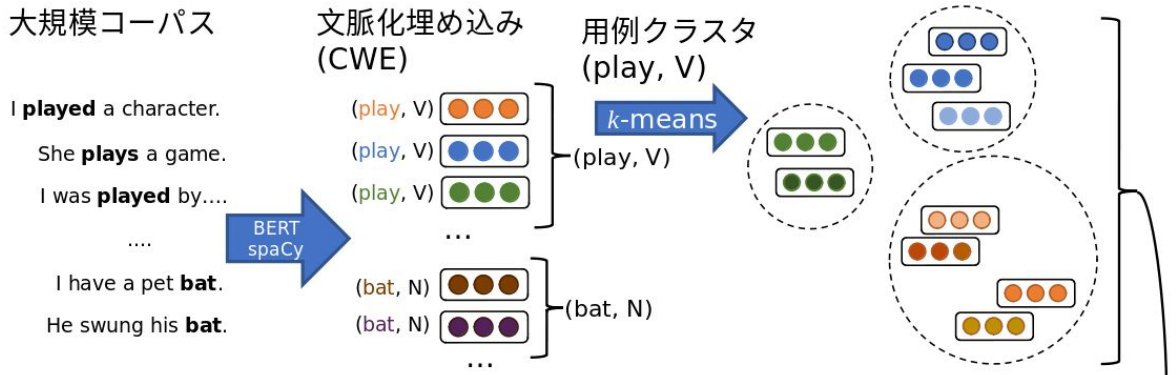
ユーザが語義を理解したい文中の単語に対し、大規模コーパスから入力された文脈に応じた語義の用例、及び他の語義の用例を探索できるシステムを構築

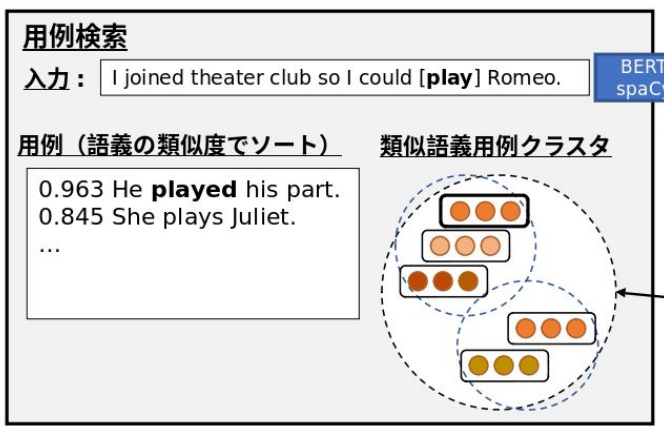
事前処理：

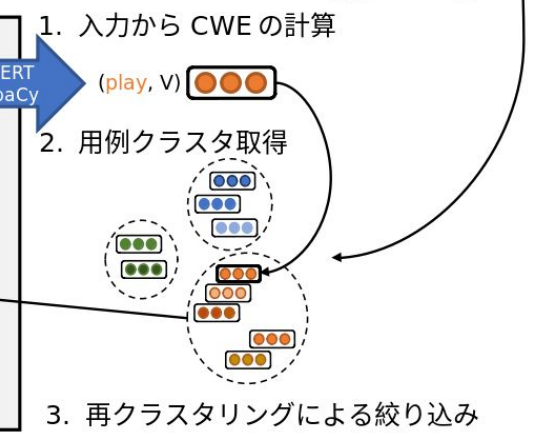
- コーパス中の各単語をBERTで埋め込み、spaCyで品詞タグ付けと見出し語化し、文に紐付け
- 語義数を抑え語義の見通しを良くするために荒い粒度でクラスタリング

オンライン処理：

- 入力された文中の単語を埋め込み、対応する語義の用例クラスタに分類し提示
- ユーザの要求に応じて再クラスタリングすることで、任意の粒度で用例を探索







議論したいこと

用例クラスタ数はどう決めれば良いか

- クラスタ数が単語によって異なる
- 辞書の定義数は多すぎる

どのようなコーパスを使えば良いか

- 学習者向けで平均難易度が大事
- 頻出度の低い単語も必要

どのようにシステムを評価するか

- 単語類似度とユーザー評価?
- 語義を理解するまでの操作数？

今後の展望

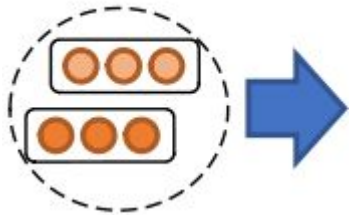
入力された単語の語義に近い違う単語の語義を表示

- 最近傍検索を用いる

学習者の母国語の用例との対応づけ

- mBERT+最近傍検索でできそう
- [Cao+2020]の手法で他言語の類似する語義をさらに近づけることが可能

用例クラスタと辞書からの定義文、典型用例を生成



“To play a character in a film or...”

“The actor played his character well.”