ジンポー語の公開データに基づいた Google Colaboratory によるテキスト処理

田口智大

奈良先端科学技術大学院大学 博士前期課程2年
The University of Edinburgh, MScR Linguistics

2021年10月16日　ジンポー語第二回フォローアップミーティング

目次

- **自己紹介**
- 発表概要
- Google Colaboratory とは
- プログラミング言語 Python とは
- ジンポー語のテキスト処理
  - **スクレイピング**
  - **データ整理**: 頻出度順単語リストの作成
  - **自然言語処理**: 条件付確率場 (CRF) を用いた品詞分類

自己紹介

Ngai na mying gaw Chihiro re. Jinghpaw ga hku gaw La Ja re.
Na gaw Scotland kaw nga ai.

- 奈良先端科学技術大学院大学M2 自然言語処理学研究室
  - Universal Dependencies, Abstract Meaning Representations の研究
  - **少数言語の自然言語処理**

- エジンバラ大学言語学修士課程（研究コース）
  - タタール語（テュルク語族キプチャク語群）の **不定詞**にまつわる形態統語的現象の研究
    - 繰り上げ構文 (raising)、コントロール、例外的格標示 (ECM)、人魚構文 (mermaid construction)、ロシア語やウラル語族との言語接触、などなど ...

目次

- 自己紹介
- **発表概要**
- Google Colaboratory とは
- プログラミング言語 Python とは
- ジンポー語のテキスト処理
  - **スクレイピング**
  - **データ整理**: 頻出度順単語リストの作成
  - **自然言語処理**: 条件付確率場 (CRF) を用いた品詞分類

発表概要

倉部さんが PARADISEC にて公開しているジンポー語のデータを用いて、
簡単なテキスト処理の例を示す

- **Web** からのテキストの集め方
- **Python** を用いてテキストデータをいじる
- テキストデータを用いて**自然言語処理**っぽいことをする

目次

- 自己紹介
- 発表概要
- **Google Colaboratory とは**
- プログラミング言語 Python とは
- ジンポー語のテキスト処理
  - **スクレイピング**
  - **データ整理**: 頻出度順単語リストの作成
  - **自然言語処理**: 条件付き確率場 (CRF) を用いた品詞分類

Google Colaboratory とは

- Google が無料で提供しているサービス
  - Google Docs, Google Spreadsheet, etc.
- ブラウザで Python を実行できる
  - Google アカウントがあればすぐに使える
  - 面倒な環境構築が **不要**
- 実行結果をすぐに確認できる

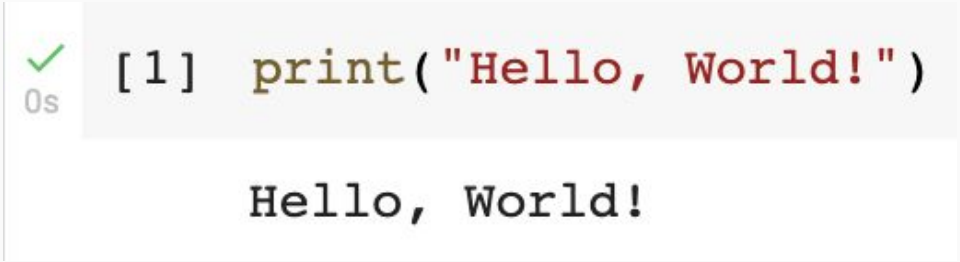
→ プログラミングの授業や単純なコーディングで広く使われている

目次

- 自己紹介
- 発表概要
- Google Colaboratory とは
- **プログラミング言語 Python とは**
- ジンポー語のテキスト処理
  - **スクレイピング**
  - **データ整理**: 頻出度順単語リストの作成
  - **自然言語処理**: 条件付確率場 (CRF) を用いた品詞分類

Python とは

- プログラミング言語の一つ
- 読みやすく、コードが簡潔
  - 比較的学びやすいため初学者に適している
- 機械学習、自然言語処理、データ分析などで広く用いられる言語
  - 近年人気上昇中
- Python のコード例:



目次

- 自己紹介
- 発表概要
- Google Colaboratory とは
- プログラミング言語 Python とは
- **ジンポー語のテキスト処理**
  - **スクレイピング**
  - **データ整理**: 頻出度順単語リストの作成
  - **自然言語処理**: 条件付確率場 (CRF) を用いた品詞分類

スクレイピング

スクレイピング:

- Webサイトなどからテキストデータを抽出すること
- 著作権やライセンスの扱いに注意

Colab 上のコード

頻度順単語リスト

マイナーだけどメジャーじゃない言語あるある:

- ニュースサイトなど資料はあるといえばあるが、単語帳のようなちょうど良い学習リソースがない

Python で作ってみよう！

コード

自然言語処理のタスク例: 品詞タグ付け

教師あり機械学習の**条件付き確率場**を用いて**品詞タグ付け**を行う

品詞タグ付け: 文中のそれぞれの単語が名詞なのか、動詞なのか……といったように品詞を機械に推定させるタスクのこと

教師あり機械学習: 問題と正解のペアを学習させて、類似の問題を与えられた時に正解を答えられるようにする

条件付き確率場 (Conditional Random Fields): 条件付き確率を利用したモデル。単語Xから品詞Yを推定したい時、単語Xの前後の情報（コロケーション）を考慮したりできる。品詞タグ付けのほか形態素解析にも用いられる。

自然言語処理のタスク例: 品詞タグ付け

- 品詞は Universal Dependencies で定義されているものを用いる
- 使用したテキストデータ:
  - 兄弟が湖を動かした話
  - 訓練データ 564単語
  - テストデータ 92単語
  - 各単語に対して手動でアノテーションを行い、CSV形式でデータを出力
    - 研究として使うにはあまりにも少ないが、
      今日のために手動で用意したのでご容赦願います
- コード

自然言語処理のタスク例: 品詞タグ付け

結果は...

- 訓練データの精度: 99.6%
- テストデータの精度: 88%

機械にとって未知のジンポー語の文に遭遇した時に、各単語に9割程度の精度で正しく品詞を判定できる

- たった562単語（33文！）でしか訓練していない割には、割とよくできている

まとめ

- 言語データを使ってプログラミングで遊んでみた
- 言語学と自然言語処理は近いようで遠い
  - しかし、問題設定によっては面白い貢献が可能
  - 特に、フィールド言語学者の持つデータは貴重
- 言語学者と自然言語処理研究者の交流が増えるべき
- ジンポー語の復習の良い機会になった
- 自然言語処理に転向して一年、それっぽいことができて嬉しい

今後の展望

ジンポー語に関して思いついた自然言語処理タスクのアイデア

- ジンポー語の Universal Dependencies
- ジンポー語正書法から、声調と声門閉鎖を予測して補う
- ジンポー語のニューラル機械翻訳（大量の対訳データが必要なため、難しい）

参考文献

- 倉部慶太. 2012. ジンポー語文法概要および民話資料ー兄弟が湖を動かした話ー.
- PARADISEC.
- Universal Dependencies.
- Ruthu S Sanketh. 2020. POS tagging using CRFs.

自然言語処理に興味のある言語学徒向け

- Dan Jurafsky and Martin, J.H. 2021. Speech and language processing (3rd ed.).
  - 言語学徒でも読みやすい自然言語処理の入門書です
- Aizu Online Judge.
  - プログラミングの入門から上級までの練習問題があります
- 言語処理百本ノック
  - 自然言語処理の練習問題100問です