特定の著者の文章的な特徴を学習して文章を生成するツール

東京情報大学　総合情報学科　情報システム学系

知能情報システム研究室　永井ゼミ

目次

- 研究の動機と目的
- 具体的内容
- ツール作成において使用するもの
- 手法
- 検証
- 計画

動機と目的

- 動機
  - テキスト分析に関する知識を深め、今後の文章作成に生かす。
  - テキストを利用したアプリケーションやツールを作りたい。
  - 機械学習を利用して自然言語処理がしたい。
- 目的
  - 何らかの著作物の傾向を掴んで

    文章校正をする。
  - 著名な人物の再現

具体的な内容

- テキストを入力して書き方を類似の文章にする。
- 例

アルバイトの初日の朝に電話がかかってきて、私は受話器のボタンを目覚まし時計より早く押した。

アルバイトの初日の朝に電話がかかってきたために、私は受話器のボタンをうるさい時計より早く押すこととなった。

類義語や言い換え、接続詞の使い方、文章構成の特徴。たくさん使う言葉など。

利用環境と技術

- Python3
  - mecab
    - 特徴語を抽出する。
  - 文字列置換アルゴリズムの作成
- N-gram（n=1,2,3,4,5)
  - 任意の文字列を連続したn個の文字で分割する
  - 類似度を調べて品詞を置き換える。（もとの文章データと比べて）
- 係り受け解析
  - 品詞のつながりを調べて文章の類似度や複雑さを見る。

検討手法

- データセット
  - 青空文庫で公開されている夏目漱石のテキストデータ
  - 地の文のみに加工
- やりたいこと
  - はじめにデータセットから特徴語を形態素解析用いて抽出する。（文章から単語や語尾などの特徴）
  - そこから言葉の置き換え表を作る。
  - 置き換え表から入力されたテキストに対して似た意味になるように言葉の置き換えを行い、結果を出力する。

検証内容

- 目標
  - まともな文章の構造をした出力ができる
  - 学習データに応じて変化がある
  - より似た文章にできる
- 評価の仕方
  - 評価関数を作って類似度を調べる。
    - 前後の文を比較して文脈を調べる。

課題と進捗

- 課題
  - 置換表の作成
  - 置換するかどうかの判断
  - 置換アルゴリズムの作成
  - 置き換えるだけで似た文章にできない可能性
- 進捗
  - 形態素解析して特徴語を抽出する
  - テキストを処理する基本的なプログラム
  - 入出力の方法

研究計画