言語処理 100 本ノックと自然言語処理を用いた作文指導

発表者：E1433 藤本 恭子 指導教員：奥村 紀之

1 はじめに

我々は、課題研究で自然言語処理に関する基礎学習と卒業研究に対する事前調査を行った。2 節では言語処理 100 本ノックについて、3 節では卒業研究と事前調査について、4 節ではこれからの活動について示す。

2 言語処理 100 本ノック

プログラミング、データ分析、研究のスキルを習得することを目的に、言語処理 100 本ノック 2015 に取り組んだ。

2.1 活動の概要

言語処理 100 本ノック[1]とは、主に言語処理に関する全 10 章 100 問からなる問題集で、問題を解くのに必要なデータ・コーパスと共にインターネット上で公開されている。また、言語処理だけでなく、統計や機械学習、Web アプリケーションなどの内容も扱っている。現在第 1 章から第 8 章までの実装が完了しており、その全てで汎用の高水準言語である Python を用いた。また、第 7 章までのコードレビューも完了している。

2.2 python による問題解決の手法

言語処理 100 本ノックを通じて、Python の基礎・主となる解析器やライブラリの使用方法・問題解決に繋がる情報収集の手法を学んだ。言語処理 100 本ノックには、日本語の文章を係り受け解析し、その結果を用いて文中の全ての名詞句のペアを結ぶ最短係り受けパスを抽出せよという問題がある。抽出した最短係り受けパスは、指定された書式に従って出力する。係り受け解析とは、形態素 (言語で意味を持つ最小の単位) に切り分けられた文章を定義された文法に従って解釈し、文の構造を明確にする作業である。この問題では CaboCha[2]を用いて係り受け解析を行った。係り受けパスとは、係り受け解析の結果を木構造の有向グラフとするとき頂点の列のことであり、この問題では、名詞に分類される頂点同士の最短経路を最短係り受けパスとしている。図 1 に名詞に分類される形態素を含んだ係り受けパスの例を示す。この図に置ける根は”記憶している”である。

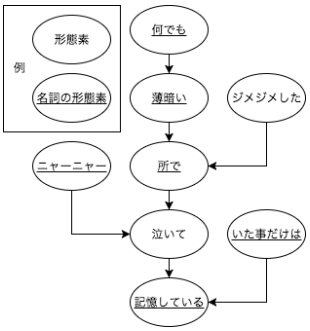

図 1: 可視化された係り受けパスの例

最短係り受けパスは以下の 2 通りが考えられ、これらをそれぞれ抽出するには場合分けなどの工夫が必要である。

場合 A 形態素 i から木構造の根に至る経路上に形態素 j が存在する場合

場合 B A 以外で、形態素 i と形態素 j から木構造の根に至る経路上で共通の形態素 k に交わる場合

場合分けの手段として重複要素の無い集合を表現するデータ型である set 型を用いたパスの重複チェックを行った。有向グラフの根からペアとなる目的の形態素までの最短経路をそれぞれ X,Y とすると、$X \supset Y$ の時**場合 A**、$X \not\supset Y$ の時**場合 B** であることが分かる。場合分け・抽出した後、出力した結果を図 2 に示す。

```
Xでも -> 薄暗い -> Yで
Xでも -> 薄暗い -> 所で | Y | 泣いて
Xでも -> 薄暗い -> 所で -> 泣いて | Yだけは | 記憶している
Xでも -> 薄暗い -> 所で -> 泣いて -> Yしている
Xで | Y | 泣いて
Xで -> 泣いて | Yだけは | 記憶している
Xで -> 泣いて -> Yしている
X -> 泣いて | Yだけは | 記憶している
X -> 泣いて -> Yしている
Xだけは -> Yしている
```

図 2: 最短係り受けパスの出力結果

[1] http://www.cl.ecei.tohoku.ac.jp/nlp100/
[2] https://taku910.github.io/cabocha/

3　自然言語処理を用いた作文指導

児童・生徒が作文を書く力を身につける為には、多くの作文を書き、フィードバックを受ける必要がある。その一方で学校教育で作文指導 [2] に当てられる時間はそう多くない [1]。そこで、人手をかけず児童・学生へのフィードバックを生成する「言語処理を用いた日本語作文の指導」を卒業研究のテーマとした。本項では、研究に向けて主に作文を採点する為の観点について調査した内容を示す。

3.1　文章の読みやすさ

建石らの研究では、日本語文の読みやすさの評価が行われた [3]。これは日本語文の表面の情報から、構文や意味によらないでその文章の読みやすさを評価する式を求めるものである。読みやすさと関係のある表面情報は、以下の 4 種類である。

(1) 文字の平均の長さ
(2) 各文字種についてその文字種達の相対頻度
(3) 文字種ごとの連の平均の長さ
(4) 終点の数の句点の数に対する比

建石らは 4 種類の情報と文章の読みやすさには相関関係があると結論している。特に (3) 文字種ごとの連の平均の長さは特に相関があることが分かっている。

3.2　語彙の多様性

語彙の多様性を表す指標として、ユール [4] の K 特性が様々な研究 [5] で用いられている。ユールの K 特性は、文章中に n 回現れた語の個数を f[n] で表すとき、以下の式で表される。

$$K = \frac{T - S}{S^2} \times 10,000$$

ただし

$$S = \sum_{n=1}^{n\ の最大} (n \times f[n]), T = \sum_{n=1}^{n\ の最大} (n^2 \times f[n])$$

とする。S は語の出現回数の 1 次モーメントである T は語の出現回数の 2 次モーメントであるが，n を 2 乗しているため，出現回数の合計が同じであっても，出現回数が偏っている程，T の値は大きくなる．

3.3　英文における自動添削システム

世界最大のテスト期間である Educational Testing Service: ETS の Burstein らの研究グループが開発した E-rater というシステムがある [6]。これは、英語の文に対し 5 つにカテゴリー分けされる 12 の言語上の特徴量のスコアに基づいて、6 点満点で採点を行うシステムである。この 12 の特徴量を以下に示す。

(1) 総語数に対する文法エラーの割合
(2) 総語数に対する語の使用法についてのエラーの割合
(3) 総語数に対する手順エラーの割合
(4) 総語数に対するスタイルについてのエラーの割合
(5) 必要とされる談話要素の数
(6) 談話要素における平均語数
(7) 語彙の類似度が 1 番近い点数
(8) 最高点を得た作文との語彙の類似度
(9) 総語数に対する異なったワードの種類の割合
(10) 単語頻度指標に基づく語彙の困難度
(11) 平均の単語の長さ
(12) 単語の総数

この 12 の特徴量は論題に寄らず固定であり、E-rater では如何なる文章でも単一の基準により添削することを目指している。

4　今後の活動

言語処理 100 本ノックに関しては、第 10 章までの実装とコードレビューを引き続き進めていく。また卒業研究の基礎学習について、自動採点・添削技術に関する論文を読むとともに、ニューラルネットワークの学習も並行して行っていく。

参考文献

[1] 慶松 勝太郎 (2011)「我が国における作文教育の問題点」,『LEC 会計大学院紀要』9,LEC 会計大学院.

[2] 文部科学省 (2009)「補習授業校教師のためのワンポイントアドバイス集」,http://www.mext.go.jp/a_menu/shotou/clarinet/002/003/002/010.htm(参照 2018-1-25)

[3] 建石 由佳ほか (1988)「日本文の読みやすさの評価式」,『文書処理とヒューマンインタフェース研究会報告』18, 情報処理学会.

[4] Yule, G.U.(1944)『The Statistical Study of Literary Vocabulary』,Cambridge University Press.

[5] 石岡 恒憲・亀田 雅之 (2002)「コンピュータによる日本語小論文の自動採点システム」,『電子情報通信学会技術研究報告』, 電子情報通信学会.

[6] 石岡 恒憲 (2004)「記述式テストにおける自動採点システムの最新動向」,『行動計量学』31(2), 日本行動計量学会.