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 節では研究課題の事前調査について，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

本授業では，プログラミングやデータ分析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[1] に取り組んだ．

2.1　言語処理 100 本ノックの概要

言語処理 100 本ノックとは，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 Python3 を用いた．また、第 8 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 本ノックには,「文中のすべての名詞を含む文節に対し，その文節から構文木の根に至るパス (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，後に示す出力結果のように，指定された書式に従って出力する．抽出元には日本語文「何でも薄暗いじめじめした所でニャーニャー泣いていた事だけは記憶している.」[2]を係り受け解析した結果を用いた．係り受け解析とは，形態素 (意味を持つ最小の言語単位) 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[3]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．

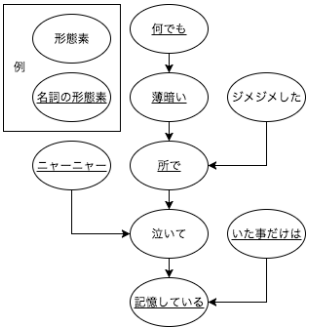

図 1: 可視化された係り受けパスの例

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

図 1 を例にすると，〈何でも〉と〈記憶している〉の関係は**場合 A** と考えられ，〈何でも〉と〈ニャーニャー〉の関係は**場合 B** と考えられる．

場合分けの手段として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，$X \supset Y$ で**場合 A**，$X \not\supset Y$ で**場合 B** であることが分かる．場合分け・抽出した後，出力した結果は以下の通りである．

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

[1] http://www.cl.ecei.tohoku.ac.jp/nlp100/
[2] 『吾輩は猫である』(夏目漱石, 1905)
[3] https://taku910.github.io/cabocha/

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

3 自然言語処理を用いた作文指導

児童や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，その作文に対して添削や指導を受ける必要がある．その一方で添削の仕事量の多さから，学校教育で作文指導に当てられる時間は多くない．実際に慶松の研究では，小中学校における作文の授業は，国語の授業時間数 (300 から 140 時間) の約 3 分の 1 であり，長時間の訓練を要する作文の学習時間を鑑みると，この数字は少ないのではないかという指摘がある [1]．

そこで，人手をかけずに児童や学生への添削や指導を生成する「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．本節では，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述べる．

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 [2]．この研究では，作文を構成する文字列から文体の特徴を抽出し，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ことで，作文の採点を行なっている．文章の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) 文あたりの平均文字数
(2) 文あたりの平均の句の数
(3) 各文字種 (アルファベット，ひらがな，漢字，カタカナ) について，文字種ごとの長さの平均
(4) 各文字種の連の平均の長さ

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づけている．特に (3) 文字種ごとの連の平均の長さは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，Yule[3] の K 特性がある．Yule の K 特性は，単語の出現頻度がポアソン分布に従うと仮定した時，単語が用いられている回数を根拠として計算される．Yule の K 特性は以下の式で定義される．

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

ただし文章中に n 回現れた語の個数を $f[n]$ で表すとき，

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

とする．K 特性値は，語彙が集中しているほど大きくなり，語彙が多様な程小さくなる．石岡らの研究 [4] では，毎日新聞の社説[4]を解析すると K の中央値が 87.3 であり，コラム[5]を解析すると中央値は 101.3 であったと述べられている．

3.3 Big Word の割合

Big Word とは，~~非常に~~抽象的であり様々な解釈を生むキーワードのことである．石岡ら [4] は，Big Word は多くの人が連想しやすいが，文章の本質が伝わらないキーワードで，状況によって Big Word を高い割合で含む文章は文章として低水準であるとしている．以下は，Big Word を含む文と含まない文の例である．

(1) シナジーを意識して早めに対処する
(2) A と共同で作業し 1 週間以内に上司に結果報告する

(1) の文章は，何をすれば「シナジーを意識」したことになるのか・「早めに」とは一体どれくらいの時間を指すのか，といった点で解釈がぶれる可能性が高い．一方で (2) の文章は，言葉が具体的で聞く側の認識も統一される．

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．また卒業研究について，自動採点の新たな観点を提案する為，今後も様々な論文を読んで基礎学習を進めていく．

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] 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.

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

[4] 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.

[4] http://www.asahi.com/news/editorial.html
[5] http://www.asahi.com/rensai/featurelist.html