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，開発には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．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る．また係り受け解析とは，形態素(言語として意味を持つ最小の言語単位)に切り分けられた文章を日本語の文法に従って解釈し，文の構造を明確にする処理である．抽出した最短係り受けパスは，以下のように，指定された書式に従って出力した．なお，抽出元の例として日本語文「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．

```
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
```

この問題ではCaboCha[3]を用いて係り受け解析を行った．また，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．

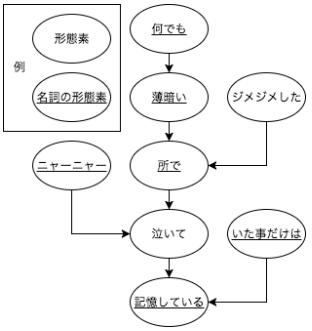

図 1: 可視化された係り受けパスの例

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

[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/

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．

この問題の他にも，正規表現，データベース，機械学習などの問題を解いた．言語処理 100 本ノックを通して，日本語文を解釈し言語処理を行うプログラムの手法を学習できた．

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．慶松の研究では，小中学校における作文の授業は，国語の授業時間数の約 3 分の 1(35 時間から 105 時間) 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].

卒業研究では，人手による添削や指導を模したシステムを開発し自動添削を行うための「言語処理を用いた日本語作文の指導」をテーマとする．本節では，研究に向けて主に作文を指導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を評価している [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，日本語の文法や文章の意味に依存せず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) 1 文あたりの平均文字数
(2) 1 文あたりの読点の数
(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，文字種ごとの出現頻度
(4) 連続した同一文字種の長さ

例えば「吾輩は猫である．名前はまだ無い.」を考えた時の表面情報の数値は，(1)7 文字，(2)0 個，(3) ひらがな 64%，漢字 36%，(4) ひらがな 3 文字，漢字 1.7 文字であると言える。

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 (4) 同一文字種の平均の長さは負の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 n を根拠として S, K が求められる．Yule の K 特性は，S, K を用いて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 回現れた語の種類数を $f[n]$ で表すとき，

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

とする．K 特性値は，語彙が一様な程大きくなり，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．語彙が大きい文章ほど理知的であるとされ，社説やコラムの解析結果は筆者自身の印象と一致した．

3.3　Big Word の割合

Big Word とは，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，作文指導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, p1-14,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, p1-8, 情報処理学会.

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

[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」,『日本計算機統計学会シンポジウム論文集』16, p153-156, 日本計算機統計学会.