

³<https://taku910.github.io/cabocha/>

X->泣いて->Y している
X だけは->Y している

言語処理 100 本ノックを通して、日本語文を解釈して言語処理に関する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 (300 時間から 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]。

私は人手をかけずに添削や指導を行う「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、作文を書く力を向上させることを目標とした。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) 同一文字種の文字の一続きの平均の長さ

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 (4) 同一文字種の文字の一続き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 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、語彙が集中しているほど小さくなり、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、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。
- [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。

⁴<http://www.asahi.com/news/editorial.html>

⁵<http://www.asahi.com/rensai/featurelist.html>