

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・生徒が作文を書く力を身につけるためには、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(300 140 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、人手をかけず児童・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、作文の表面の情報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) 文字種ごとの連の平均の長さ

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、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、

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

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あるとしている。また、文中に含まれるBig Wordの割合と読み手に与える印象には関係性が見出される。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、言葉が具体的に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、今後と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>