Redpen を使った文書自動添削ツール

PM コース　矢吹研究室　1442031　氏名 小山 隆太郎

1. 研究の背景

ソフトウェアエンジニアはプログラムを組むだけではなく, たくさんの技術文書を書く. そして専門的なチュートリアル, マニュアル等を読み手に理解してもらうことが大切である. ソフトウェアを開発する際には多くの実装テストをする必要がある．そのため CheckStyle や lint 等の静的解析ツールを導入することで, フォーマットのエラーを自動で検知することが出来る. しかし静的解析ツールは文書のチェックを行えるものがなく, 文中のミスを修正することに時間を割く等で作業の本質に取り組む時間以外に無駄が生じてしまうことがある．このことはエンジニアにとって悩みの種である．このような状況に対して Redpen と呼ばれる文書静的解析ツールが開発され, 日々エンジニアが改良を加えている [1]．Redpen は一部の静的解析ツール (CheckStyle,lint 等) に相当する機能を文書に与えるものであり, 文書作成でも最低限の検査を自動で行いたいという動機の下, 改良が進んでいる．大学の授業等で文書を提出することがあるため,Redpen の開発に着目することでこれからの文書作成の質が向上するのではないかと考えた.

2. 研究の目的

Redpen を利用し, 文書作成に割く時間を短縮できるようにすることで, もっと大きな粒度の問題 (実作業, 着目すべき文や質等) に集中できるようにすることが目的である．また, 文中のミスを少なくした状態で文書を提出できるようにする [2].

3. 研究の手法

- 1,Redpen が動作する環境を構築する [3].
- 2, 動作することを確認したら, 日本語の文書”ja”が添削できるように設定ファイルを書き換える.
- 3, 参考文献のサンプル文 [4] で添削を行い, 添削結果が本の改善文に準拠するよう設定ファイルを書き換える.

4. 研究の結果

Redpen の設定ファイルを書き換え, 以下の表現を添削する機能にした.

- SuccessiveWord 同一の単語が連続して使用されていないかを検査する.
- KatakanaEndHyphen カタカナ単語末尾の長音検査する.
- KatakanaSpellCheck カタカナ単語のスペルチェックをする.
- InvalidExpression 顔文字や感情的な表現など技術文書で利用するべきでない句の使用を検査する.
- JapaneseStyle ですます調, である調が混じっていないかを検査する.
- DoubleNegative 二重否定を検知する.
- FrequentSentenceStart 同じ文頭表現が過度に利用されていないか検査する. 例えばどの文頭も”私は〜”からはじまる等.
- JapaneseAmbiguousNounConjunction 格助詞の「の」+ 名詞連続 + 格助詞の「の」というパターンを発見するとエラーを出力する.
- JapaneseNumberExpression 日本語の数値表現を検査する機能.「ひとつ, ふたつ」,「１つ, ２つ」などのゆらぎを検知する.
- LongKanjiChain 長過ぎる漢字の連続をみつけるとエラーを出力する.
- SuccessiveSentence 同一内容の文が連続して出現するとエラーを出力する.

- DoubledConjunctiveParticleGa 一文に二回以上, 接続助詞または二重否定の「が」が出現するとエラーを出力する.

Redpen がコマンド上で動作し, 文中ミスを正しく抽出できたので以下に実行結果を載せる.

図 1

図 1 は 2 ヶ所の文中ミスを指摘している.ValidationError[CommaNumber] The number of commas (5) exceeds the maximum of 3. at line は文中の句点 (コマ数) が多いことを指摘している. この場合” 摂取量、安全、嗜好”の句点を”・”にすることで見やすい表記になる.ValidationError[DoubledConjunctiveParticleGa], Found multiple conjunctive particle: "が" at line は文中に二重否定があることを指摘している. この場合”野菜ジュースは濃度が上がるほど嘱目繊維が豊富になりますが 、消化機能が低下した人が誤嚥をする危険が増えますが、”と二重否定が使われていることがわかる.

図 2

図 2 は 2 ヶ所の文中ミスを指摘している.ValidationError[SuccessiveWord], Found word "た" repeated twice in succession. at line:は文末の”た”が 2 つあることを指摘している.ValidationError[SuccessiveSentence], Found similar two sentences in succession は文中に同じパラグラフが 2 つあることを指摘している. この場合”ポケモン Go はおもしろい。”が 2 つあることを指摘している.

5. 考察と今後の計画

Redpen には設定ファイルが設けられており, 設定を書き換えるごとに出力結果も変わる. その為きちんとした設定が確立されておらず, 現在も研究が続いている.Redpen のファイルの設定を確立し, 文中のミスを抽出することは出来たが, 中には不要である間違いを指摘されることがあった. これを取り除く設定を確立できれば, より精度の高い文書添削ツールになり得るだろう. 今後は皆が共通して使えるように添削マシンを拡張する予定である.

参考文献

[1] Redpen を使って技術文書を気軽に校正しよう. http://gihyo.jp/lifestyle/serial/01/redpen.

[2] Redpen でお手軽文書校正. http://tech.respect-pal.jp/try-redpen/.

[3] 阿部紘久. シンプルに書く　伝わる文章術. 飛鳥新社, 2012.

[4] Redpen1.7 ドキュメント. http://redpen.cc/docs/latest/index_ja.html.