レポート評価支援システム構築に向けた文章の特徴量の抽出

Extracting feature values of sentence for support system for evaluating of the report

蘆田誠人[*1]，内田眞司[*1]
Makoto ASHIDA[*1], Shinji UCHIDA[*1]
[*1] 奈良工業高等専門学校情報工学科
[*1]Department of Information Engineering, National Institute of Technology Nara College
Email: {ashida, uchida}@info.nara-k.ac.jp

あらまし: 学生が実験や研究のレポート等を書く場合, 文章を何度も推敲することが望ましい. しかし初心者の場合, 文章を推敲するポイントがわからない, また推敲するポイントがわかったとしても評価基準がわからないために, 修正してよいかどうかわからないといった問題がある. 本研究では, 初心者を対象としたレポート評価支援システムを構築することを前提として, 文章の推敲ポイントの評価基準となる特徴量を, 査読付き学術論文の文章から抽出する.

キーワード： テキストマイニング，文章の特徴量，段落間情報量

1. はじめに

自分の伝えたい意図を文章で伝えるためには，対象となる文章を推敲する必要がある．例えば，学生がレポートを作成する際は，書いた文章を何度も推敲しながら完成させる．わかりやすい文章を目指して推敲するが，初めてレポートを作成する学生にとって推敲という作業は難しい．そもそも，推敲するポイントが分からないし，基準もあいまいである．一般的には，推敲ポイントは文章の長さ，段落の長さ，段落間の関係など多岐にわたる．また，推敲ポイントについて修正するか否かを判断する基準が存在しない.

そこで本研究では，初心者を対象とした推敲支援を行うシステムを作成するために，推敲ポイントの特徴量を抽出する，世間に公開されている学術論文の特徴量の抽出を行う．予め定義した特徴量を学術論文から抽出し，その特徴量を学生が作成したレポートの特徴量と比較する．そして，その特徴量の有用性を導く.

2. 研究手法

本研究では，1 文あたりの長さが長いと読みにくいことから文章の長さに関する特徴量に着目する．また，論文はトップダウン構造を用いることが望ましいとされているので段落間の関係に関する特徴量に着目する.

文章の長さに関する特徴量

- 1 文あたりの文字数
- 1 文あたりの文節数
- 1 段落あたりの文節数

上記 3 点の特徴量は，1 文あたりの長さが長いと読みにくいという観点より抽出する[1].

段落間関係に関する特徴量

- 段落間の情報量

上記に示した 4 項目について，情報処理学会論文誌に掲載されている査読付き論文の特徴量を抽出する. そして, 学生が作成したレポートの特徴量を各々比較し，その特徴量の有用性を探り,考察する．抽出には統計解析向けのプログラミング言語である「R」を用いたテキストマイニングを用いて抽出する．テキストマイニングとは，定型化されていない文章の集まりを単語やフレーズに分割し，それらの出現頻度や相関関係を分析して有用な情報を抽出する手法やシステムのことである.

文字数，文節数，段落文節数は「R」を用いて抽出を行う．段落間の情報量は，1 段落の文節数を抽出し，その中から 1 つ前の段落の文節と重複する文節の数をカウントし, 定義した公式で情報量を導く. 定義した公式を式(1)に示す.

$$I_i = -\log \frac{A_i}{T_i} \quad (1)$$

I_i ： i段落と$(i-1)$段落間の情報量

A_i: i段落と$(i-1)$段落間で重複している文節数

T_i: i段落の総文節数

今着目している段落をiとする. I_iは着目している段落とその前の段落間の情報量を表している．A_iは着目している段落とその前の段落で，重複している文節の数を表している．またT_iは，着目している段落の総文節数を表している．これらを基に，今着目している段落の文節が前の段落の文節をどの程度用いているのかを確率で求め，それを-log にかけることにより情報量が求まる．各段落の情報量の推移から，その文章がどの構造なのかを判断する．本研究では，トップダウン構造とボトムアップ構造を取り上げる．トップダウン構造とは，論文の最初に大構造を決め，詳細な下部構造を作る構造であり，読者にとって流れが把握しやすい構造である．ボトムアップ構造は，テーマに沿ったアイデアを次々に書いていく構造である．しかし，読者が議論の流れを

把握しにくく，論文を作成するのにこの構造は向いていない．

3. 研究結果

情報処理学会論文誌に2004年から2014年に掲載された査読付き論文 50 本と，学生が作成した本校講義「情報工学実験Ⅰ」のレポートの特徴量の抽出を行い比較した．学生のレポートについては，レポートの評価が高いレポートと低いレポートそれぞれ 2 冊を準備した．文章の長さに関する特徴量について図 1 に示す．

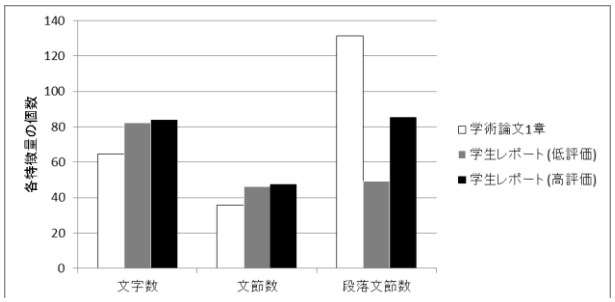

図 1　文章の長さに関する特徴量

図 1 は左から順に，1 文あたりの平均文字数， 1 文あたりの平均文節数， 1 段落あたりの平均文節数を示している．各々の特徴量内は左から順に，学術論文，低評価の学生レポート，高評価の学生レポートである．文字数，文節数のグラフに着目すると，学術論文が最も低い数値となった．学生が作成したレポートの文字数と文節数は，高評価，低評価の間で大きな差は見られなかった. 論文を作成する際は，文が短い方が良いという観点より，特徴量としてシステムに取り入れることができる[1].

一方，段落文節数の結果に着目すると，大きな差が生じた．本研究では，論文の 1 章の各特徴量とレポートのアブストラクトの各特徴量を比較している．段落文節数の大きな差が出た要因の一つとして考えられる．

次に段落間の情報量の結果を図 2 に示す．図 2 のグラフは実線が高評価のレポート，点線が低評価のレポートの情報量の推移となっている．段落間の情報量を定義した公式(1)で求めた結果, 図 2 に示す通り，高評価の学生レポートは段落が進むにつれ情報量が少なくなっていることが結果として出た．これは，新しく使用されている単語の割合が徐々に少なくなっていることを意味している．よって，トップダウン構造がなされた論文だと言える[2]．学生が作成した，高評価を受けた 2 つのレポートは，どちらもトップダウン構造で作成されていた. 一方, 低評価を受けた 2 つのレポートは，段落が進むにつれ，情報量が上昇しているものや，増加したり減少したりする構造でトップダウン構造とは言い難いものであった．この結果を受け，トップダウン構造が目指すべき構造と考える．

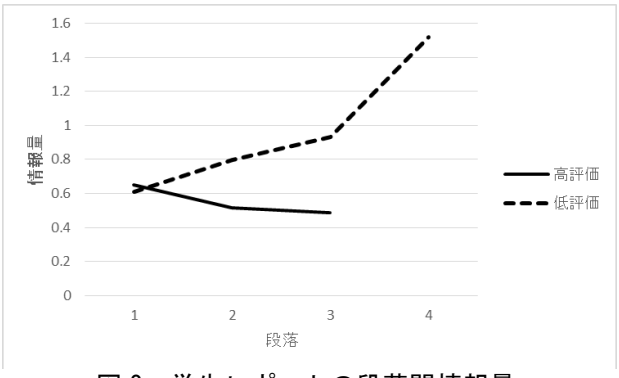

図 2　学生レポートの段落間情報量

4. まとめ

本研究では，レポート評価支援システム構築に向けた文章の特徴量の抽出を行った．特徴量の中の 1 文あたりの文字数，1 文あたりの文節数，段落間の情報量は，レポートと論文を比較した結果，論文がレポートの特徴量を下回ったため，所望の結果が求められたと言える．よって，システム構築の際に用いることができる特徴量であると考える．しかし，1 段落あたりの文節数は，論文の 1 章とレポートのアブストラクトを比較したため，大きな差が出たと考える．この特徴量で所望の結果を得るためには，論文とレポートの比較する対象を一致させることが一番の解決策として考える．

また，本研究では，段落間の情報量から構造を調べる際，トップダウン構造のみを対象としていた．レポート評価支援システムをより実用的なものにするためには，トップダウン構造の他にボトムアップ構造，トップダウンとボトムアップが混合している構造についても考慮しなければならないと考える．

5. 参考文献

[1] 清水康隆：”論文等における文の長さに関する検討”，日本教育工学雑誌 21, p1-p4(1997)

[2] 山手砂都美・砂山渡：”トップダウン・ボトムアップな文章構造作成のための推敲支援システム”，人工知能学会全国大会論文集 27 , p1-p4(2013)