8L－06　中国人学生日本語作文 n-gram モデルによる特徴抽出

高 建斌* 馬 火玄* 西野 順二** 小高 知宏** 小倉 久和**
*福井大学工学研究科　**福井大学工学部

1　はじめに

日本語を母国語とする日本人が日本語で書いた文章でも人それぞれの特徴がある [1]。外国人の書いた日本語文章は、日本人の書いた文章とまた異なる意味での特徴を持つと考えられる。しかし、これに関する研究報告はまだ見られない。ここでは、中国人の日本語学習者の日本語文を対象に、n-gram モデルによる特徴抽出を行う予備的な結果を報告する。

n-gram モデルは確率的言語モデルの中で最も基本的な方法で、文章の特徴的な抽出によく使われる [2][3]。本報告では n 文字パターンとその使用頻度を用いて、外国人の日本語文章の特徴を抽出することを試みている。この方法の利点としては、不自然な日本語で書いた、形態素解析をしにくい外国人の日本語文章であっても、その特徴が複雑な計算をせずに抽出できることである。

2　n-gram モデルによる特徴抽出

学生の作文にはいろいろな特徴がある。同じテーマであってもずいぶん異なる書き方をする。この特徴を n=3 の 3-gram、３文字パターンの出現頻度を分析することによって検討する。

学生作文と日本語コーパス、日本人作家の作品との間の類似度を次のように定義する。ある一人の中国人学生作文 G のすべての３文字パターンの組を G_3、日本語コーパス C のすべての３文字パターン組を C_3、ある日本人作家の作品 S のすべての３文字パターン組を S_3 とする。G_3の大きさを N_{G3}とする。また、G_3、C_3、S_3のうち、出現頻度の高い上位 200 個の３文字パターンをそれぞれ G_3^{200}、C_3^{200}、S_3^{200}とする。G_3と C_3の間のマッチング数 (一致する文字列の数) を $M_{G,C}^{3,3}$とし、G_3と C_3の間の類似度を $R_{G,C}^{3,3} = M_{G,C}^{3,3}/N_{G3}$とする。また、$G_{200}$と C_3との間の同様の類似度を $R_{G,C}^{200,3}$とする。$R_{G,C}^{200,200}$なども同様に定義する。G-S、C-S についても同様に類似度を定義する。本報告では、主として G-C、G-S 内の類似度に基づく分析結果を示す。

Characteristic extraction of Chinese student composition by the n-gram model
Jianbin Gao* Xuan Ma* Junji Nishino** Tomohiro Odaka** Hisakazu Ogura**
*Graduate School of Engineering, Fukui University
**Faculty of Engineering, Fukui University

3　実験と実験データ

本報告では、中国人学生の書いた志賀直哉の『城の崎にて』の粗筋文を対象に、EDR 日本語コーパス、及び志賀直哉自身の小説原文を使用して、実験を行った。

3.1　実験データ

[中国人学生の作文]

19 人クラスの中国の外国語大学日本語コース三年生が書いた志賀直哉の『城の崎にて』の粗筋のまとめ文で、19 人分の文章の総計文字数は 18164 文字である。この作文には解釈的な書き方もあれば感想的な書き方もあるし、500 文字以下のものもあれば 1500 文字に近いものもある。本報告では作文と称す。

[日本語コーパス]

約 16MB のデータ量を持つ EDR のコーパス中の例文 (約 22 万文) を抽出して、1469286 パターンの 3 文字パターンを作成した。この中には、英文字だけのような文字パターンは取り除いてある。

[志賀直哉の小説原文]

学生たちが読んだ作品は現代仮名使いである。ここでは井上靖他編の『日本の短篇・上』に載せられた現代仮名使いの『城の崎にて』を使用した。本報告でいう「原文」はこの 5321 文字の作品を指す。

3.2　実験

図 1 は $R_{G,C}^{3,3}$、$R_{G,S}^{3,3}$、$R_{G,C}^{3,200}$、$R_{G,S}^{3,200}$の結果である。横軸は 19 人学生の作文を表す番号で、縦軸は類似度である。ただし、作文の番号は G と S の類似度 $R_{G,S}^{3,3}$の大きさの順につけた。

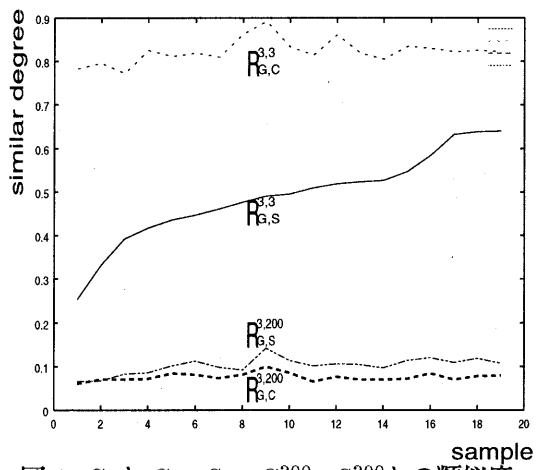

図 1: G_3と C_3・S_3・C_3^{200}・S_3^{200}との類似度

$R_{G,S}^{3,3}$について、値が大きいことは学生作文において原

文からの抜き出しが多くなっていることを推測させる。また、$R^{3,3}_{G,C}$が$R^{3,3}_{G,S}$と比べて、個人差は多少あるものの日本語コーパスとよく似ていることから学生作文の日本語らしさはその書き方にあまり影響されないこととが推測できる。$R^{3,200}_{G,C}$と$R^{3,200}_{G,S}$にいては一致度は低い。

図2は$R^{200,200}_{G,C}$、$R^{200,200}_{G,S}$、$R^{200,3}_{G,C}$、$R^{200,3}_{G,S}$の結果を示す。横軸は図1と同じ学生作文を表す番号であり、縦軸はそれぞれの類似度を示す。

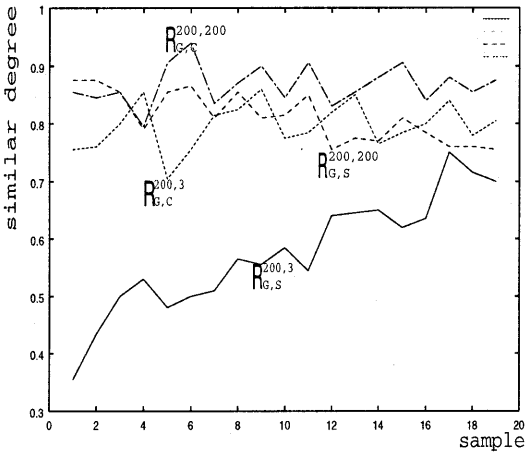

図 2: G^{200}_3とC^{200}_3・S^{200}_3・C_3・S_3との類似度

$R^{200,200}_{G,S}$について、番号が大きくなるにつれその値が小さくなる傾向があることから、図1の$R^{3,3}_{G,S}$と比較すると、番号の小さい方が原文のキーワードを掴んだ度合の高いことが推測できる。また、$R^{200,200}_{G,C}$は図1の$R^{3,3}_{G,C}$と同じく、学生作文の日本語らしさの程度が低くないことを示唆する。$R^{200,3}_{G,C}$と$R^{200,3}_{G,S}$から図1の$R^{3,3}_{G,C}$と$R^{3,3}_{G,S}$の傾向性があるが、何かによって変るところもあることが考えられる。

4 考察

『城の崎にて』の粗筋についてまとめた19の中国人学生の作文を読んで検討した結果次のように分類できた。

- 感想文・解説文
  原文を引用しながら自分の感想、或は作者の主張を説明する文 (1、2、3)。
- サンドイッチ式文
  作文の前後、或は前か後ろかに概説、まとめ的なものを入れ、真ん中部分は原文を引用する文。(4、7、15)
- 接続文
  殆んど原文を引用するが、文章段落間の接続を配慮して接続的な言葉で文章を繋がる文。(5、8)
- キーワード置換文
  抜き出し文に似ているが、原文のキーワードに少しだけ手を入れた文。(9、10、11、14、16)
- 抜き出し文
  殆んど原文そのままを抜き出して、小説に現れた順に繋がってできた文。(6、12、13、17、18、19)

図1の$R^{3,3}_{G,S}$から分かるように、番号が大きくになるにつれその作文は原文からの引用、あるいは抜きだした部分が多くなる傾向があることが分かる。もちろん本報告に使用した学生作文は日本人作家の作品の粗筋文であることに関係がある。$R^{3,3}_{G,C}$から分かるように中国の外国語大学日本語コース三年生は比較的日本語らしい表現ができると考えられているが、それでも作文の中に表現的な不足や誤りが少なくない。$R^{3,3}_{G,C}$だけではこのような表現的な誤りは抽出が困難である。また、SやCのような文章は学生作文一人一人より何倍も長いから、$R^{3,200}_{G,C}$と$R^{3,200}_{G,S}$でそのような作文の特徴を抽出するのは無理かもしれない。

$R^{200,3}_{G,S}$は$R^{3,3}_{G,S}$より学生作文の特徴を抽出できたことについて、作文5を例として挙げる。接続文に分類されたこの作文は短い7段落からなり、2段落目の書き方が感想文・説明文に近い。図2の$R^{200,3}_{G,S}$の中で、5の所は4より高くなると図1の$R^{3,3}_{G,S}$と一致するが、低くなっている。作文5の原文引用率は4より高いが、文の長さは長いサンドイッチ式文の作文4の半分ぐらいである。これは図2の$R^{200,3}_{G,S}$において作文5が作文4より低い原因であると考えられる。つまり、$R^{200,3}_{G,S}$は引用率が高いが文の長さが短ければその文章の値が高くなることが分かる。また、$R^{200,200}_{G,S}$のから、作文の原文のキーワードの使用度合はその学生の原文へ理解度と文章のまとめ方のうまさがうかがえる。

本報告の実験で$R^{3,3}_{G,S}$、$R^{200,3}_{G,S}$、及び$R^{200,200}_{G,S}$により述べた文章の書き方の特徴を抽出できた。これらの指標は原文からの引用の少ない感想文・解説文であることを示す。また、$R^{3,3}_{G,S}$と$R^{3,3}_{G,C}$、$R^{200,200}_{G,S}$と$R^{200,200}_{G,C}$、$R^{200,3}_{G,S}$と$R^{200,3}_{G,C}$から、学生作文一人一人の個性を推測できると同時に、これらの指標を学生作文への評価に使うことも期待できる。

参考文献

[1] 松浦司・金田康正著「近代日本小説家8人による文章の情 n-gram 分布を用いた著者判別」報処理学会研究報告 2000-NL-137,PP.1-8.

[2] 下畑さより・杉尾俊之著「隣接文字情報を用いた n-gram 抽出文字列からの名詞句の自動抽出」情報処理学会研究報告 96-NL-114,PP.13-18.

[3] 近藤弓末他著「日本語コーパスを使用した文章完成テストの表層的な解析」電子情報通信学会論文誌 A Vol.J80-A No.6 pp1038-1041,1997.