n-gram による中国人日本語学習者日本語作文評価

3H−5

高 建斌 *　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

外国人日本語学習者の日本語学習支援システムの開発を目的とし、その一環として、我々は n-gram 分布による中国人日本語専攻学生作文の特徴抽出を試みている [1]。本報告では、中国人日本語学習者の日本語作文を対象に、n-gram 分布による作文評価を行い、その予備的な結果を報告する。ここで言う作文評価は一定の基準により作文を採点することを意味する。

n-gram 分布は確率的言語モデルの中で最も基本的な方法で、文章の特徴抽出によく使われる [1][2][3]。形態素解析や構文解析を行わず、膨大な品詞接続情報や意味情報などを必要としないため、不自然な日本語が含まれた外国人日本語文章の処理に適していると考えられる。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を評価する手法を提案する。

2　n-gram による作文評価

本報告では、3 文字パターンの出現頻度分布である 3-gram 分布を用い、テキスト間の類似度を計算し分析することによって、外国人の日本語作文を評価する方法を検討する。

実験用の中国人学生作文と日本語コーパス、日本人学生作文、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 S_x の 3-gram 分布中の文字パターン組みを SD_x、日本語コーパス C の 3-gram 分布の文字パターン組みを CD、日本人学生作文 J の 3-gram 分布の文字パターン組みを JD、日本人作家の作品 O の 3-gram 分布の文字パターン組みを OD とし、SD_x の大きさを N_{SDx}、OD の大きさを N_{OD} とする。SD_x と CD の間のマッチング数を M^{CD}_{SDx} とし、SD_x と CD の間の類似度を $R^{CD}_{SDx} = M^{CD}_{SDx}/N_{SDx}$ とする。SD_x と JD の間の類似度も同様にするが、SD_x と OD の間の類似度は $R^{OD}_{SDx} = M^{OD}_{SDx}/N_{SDx}$ と、$R^{OD}_{SDx} = M^{OD}_{SDx}/N_{OD}$ とする。また、テキスト間の 3-gram 分布類似度も同様に定義する。本報告では、主としてテキスト間の 3-gram 分布中の文字パターン組みの類似度に基づく分析結果を示す。

3　実験と実験データ

本報告では、中国人学生の書いた芥川龍之介の『羅生門』の感想文を対象に、EDR 日本語コーパス、及び日本人学生作文、芥川龍之介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]

中国某外国語大学日本語コース三年生 19 人が書いた芥川龍之介の『羅生門』の感想文で、19 人分の作文の総計文数 664 文、総計の文字数は 24686 文字で、一作文の平均文数は 34.94 文、平均文字数 1299 文字であり、一文の長さは平均で 37.18 文字数である。中では最も長い作文は 3018 文字、最も短いのは 788 文字である。本報告では中国人学生の作文を作文と称す。

[日本語コーパス]

約 16MB のデータ量を持つ EDR コーパスの日本語コーパス中のテキスト文を抽出したものである。テキスト文 194135 文で、総文字数は 8059226 で、作成した文字列パターンは 1469286 である。英文字だけのような文字パターンはこの中から取り除いてある。

[日本人学生の作文]

全国学校図書館協議会編「考える読書」[4] の日本人の中高校生の『羅生門』感想文 37 篇を一つのテキストファイルにしたものであり、総文字数は 68805 で、作成した文字列パターンは 27019 である。

[芥川龍之介の小説原文]

芥川龍之介作『羅生門・杜子春』[5] に掲載された『羅生門』で、総文字数は 5972 字、作成した文字列パターンは 4062 である。本報告でいう原文はこの作品を指す。

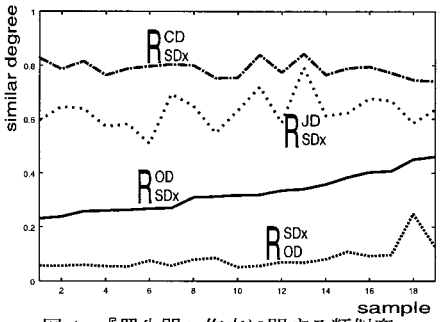

図 1: 『羅生門』作文に関する類似度

3.2　実験

図 1 は R^{CD}_{SDx}、R^{JD}_{SDx}、R^{OD}_{SDx}、R^{SDx}_{OD} 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 S と O の類似度 R^{OD}_{SDx} の大きさの順につけた。

Evaluation of Chinese Japanese learning person composition by the n-gram model
Jianbin Gao* Tomohiro Odaka ** Hisakazu Ogura**
*Graduate School of Engineering, Fukui University
**Faculty of Engineering, Fukui University

図1に示した四つの類似度のどれが作文評価指標に相応しいかについて検討した。その結果を図2と図3に示す。

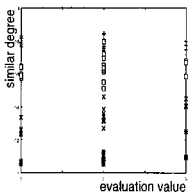

図2：日本語らしさの評価指標

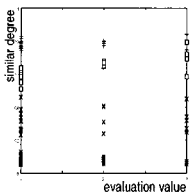

図3：感想文としての評価指標

図2と図3の縦軸は作文と比較基準との類似度を表す。横軸は作文の3段階評価の評価値を示す。3段階評価基準は表1が示す。

横軸番号	1	2	3
評価値	可	良	優

表1　3段階評価基準表

作文評価値は表2に示すような基準で採点したものである。

評価項目	優	良	可	評価項目	優	良	可
文法	3	2	1	原文理解	3	2	1
言葉遣い	3	2	1	まとめ	3	2	1
表現	3	2	1	主題把握	3	2	1
表記	3	2	1	文章構成	3	2	1
最終評価	3	2	1	最終評価	3	2	1

表2　評価値の基準表

図2の＋は R_{SDx}^{CD}、□は R_{SDx}^{JD}、×は R_{SDx}^{OD}、＊は R_{ODx}^{SDx} を表す。中では日本語らしさの評価指標として期待できるのは R_{SDx}^{OD} である。R_{SDx}^{OD} は作文の原文から引用した度合を表すもので、原文から引用したものが多ければ多いほど高くなると考えられる。

図3の印は図2と同じ類似度を表す。各類似度の中で、作文の評価指標として考えられるのは R_{SDx}^{JD} である。R_{SDx}^{JD} は中国人学生の作文と日本人学生作文との類似度である。

4　考察

4.1　評価指標の意義

R_{SDx}^{CD} は作文がコーパスのテキスト文に類似する度合を示すもので、四つの類似度の中で最も高く作文が日本語文章として成り立っていることを示している。R_{SDx}^{JD} は中国人学生の作文と日本人学生作文との類似度で、四つの類似度の中で個人間の差が著しいものの一つで、作文の日本語らしさや文章のまとめ方を評価する指標として使えると推測できる。R_{SDx}^{OD} は作文から考察した場合、作文が原文にどれほど類似するかを示し、四つの類似度の中で個人間の差が著しいもののまた一つで、作文の原文からの引用量を測定できる可能性がある。R_{OD}^{SDx} は原文から見た場合、作文がどれくらい原文から写し出したかを示すもので、作文の自作量を測定する指標になることが推定できる。

4.2　評価指標について

実験に使用したコーパスのテキスト文は殆んどジャーナル関係のもので、作文とやや異なる性質を持つため、作文とコーパスとの類似度 R_{SDx}^{CD} は作文の日本語らしさや感想文の評価指標として使いにくい。

同じテーマで書いた感想文の中国人学生の作文が、優れた感想文として認められる全国コンクール入選作品の日本人学生作文との類似度 R_{SDx}^{JD} は，高ければ高いほど感想文として高く評価できる。つまり R_{SDx}^{JD} は実験用のような文章の文章評価指標として認められる。ただし、日本人学生作文との類似度があまり高くない18番の作文は優として評価された。この文章は19人分の作文の中で最も長く、最も短い作文の四倍くらい長く、二番目に長い作文の二倍の長さである。作文が特別に長かった場合、模範文章との類似度に影響があるかどうか、もしあるとすればどんな影響があるかを検討する余地があると考えている。

作文の原文からの引用度を表す R_{SDx}^{OD} は例外があるが、日本語らしさの評価指標として考えられる。ただし R_{SDx}^{OD} 最も高い19番の作文は優として評価されなかった。この作文と18番の作文の引用量の割合がほぼ同じであるが、引用文以外の部分に間違いが多いため全体的な評価は低くなった。このような例外をどう処理すればよいかは今後の課題のもう一つである。

原文からの写し出しの度合を表す R_{OD}^{SDx} が作文の自作量を測定できると考えているが、自作量の評価指標として使えるかどうかの考察はまた今後の課題の一つである。

参考文献

[1] 高建斌他著「中国人学生作文n-gramモデルによる特徴抽出」情報処理学会第62回全国大会講演論文集(2)p2-227

[2] 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.

[3] 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.

[4] 全国学校図書館協議会編「考える読書」毎日新聞社(昭和44,50,52～55,57～61,1988～1999)

[5] 芥川龍之介作『羅生門・杜子春』岩波少年文庫(2000)