複数コーパスからの特徴表現抽出を利用した文体統一性の評価とその活用について

橋本喜代太[1]・安藤秀明[2]・竹内和広[2]
大阪府立大学[1]・大阪電気通信大学[2]

1. 本研究の背景と目的

作文の教育・学習支援を考える場合、用いる語の選択、文章構成などさまざまな観点からの支援が必要であるが、その一つに文体の統一性ないし一貫性への配慮がある。実際、学生の答案やレポートを見れば他の多くの欠陥とともに、答案やレポートとしては不適切な表現が目立つと同時に、全体としても文体が統一されていないことが目につくことが多い。作文作成者であれ、その採点者であれ、どこが不適切な箇所かを指摘する、さらにはその取るべき文体にとって望ましい表現候補を示すといった支援ツールがあると便利である。

文体はさまざまな定義が考えられうるが（例えば[1]や[2]を参照）、である・ですます調の統一などだけでは明らかに文体の統一性を確保したとは言えず、さらに細かな点での統一性の評価が不可欠である。

本研究は、総合的に作文の教育・学習支援を行なうツールを開発するというプロジェクトの一つとして、文体の構成要因として何を含むかを議論することは避け、ある目的のための文章、例えば学生が書くべき小論文やレポートなど、として模範となる文章群のコーパスとその目的のためには模範とすべきでない文章群のコーパスを指標として用い、ターゲットとなる文章の始めから終わりまでがどれだけ模範となるコーパスに類似した特徴を示すか、ということを測定することで文章の統一性の評価を行ない、それを分かりやすい形で視覚化する手法を提案する。

2. 指標作成のためのコーパス

2.1 特徴表現の基準コーパスの収集

評価ターゲットとなる文章がどの程度文体的に一貫しているかを測定するため、2 種類の言語資源（コーパス）の対照性を元にした指標を用いる。指標は、与えられた目的、読者対象、ジャンルなどの点から、当該作文に適した表現が出現する度合いが多いであろうコーパス（適切表現コーパス）と、そうでないコーパス（不適切表現コーパス）においてそれぞれ用いられている表現の頻度を用いて作成する。今回の実験データとしては、具体的な文書集合には、前者には、日本語版 Wikipedia[3]、後者には、２ちゃんねる[4]を利用する。

表現の抽出に当たっては、その表現レベルにおいて、表層の文字列、形態素解析を経た単語や形態素、さらには何らかの基準に基づいて抽出した文節レベルなどのチャンクなどが選択肢として考えられ、我々もプロジェクトの一環としてこれらの単位がそれぞれ文体の統一性のどのような側面を明らかにするかについて検討しているが、今回はもっとも単純に文字単位の n-gram を取り、適切・不適切な各表現を抽出することを試みた。これは、本稿の対象とする作文の教育・学習支援では、作文の対象となる表現が必ずしも自然言語処理用の電子辞書に含まれているとは限らないこと、また、学生の作文に頻出する不適切表現は日本語として標準でない表現が数多く含まれるため、辞書を前提とした処理だけでは十全な効果を期待しにくい、といったためである。

前述したように、本稿では Wikipedia を適切言語表現を多く含む集合、2 ちゃんねるを不適切な言語表現を多く含む集合と仮定する（以下、それぞれの集合に含まれる文章を単に、Wikipedia，2ch と参照する）。

Wikipedia コーパスは、2007 年 10 月 13 日時点の 808,515 ページからなる Wikipedia のダンプファイル[6]を用いた。これは、XML により記述されている

ため xml2sql を用いて XML タグを外し、その後画像のタグや表のタグ等を除外した。

2ch コーパスは、2ch のスレッドから名前、日付等を除外したものを使用した。ファイルの加工を経て、実際に作成したコーパスとその詳細を以下の表 1 に表す。

表 1:作成したコーパスのデータ量

コーパス	文字数
Wikipedia	161,223,892
2ch	108,031,243

2.2. 文字 n-gram 統計

まず、文字 n-gram とは、文字単位の n 次の連接を指す。例えば、『今日は寒いですね』では文字 2-gram は図 1 のように『今日』『日は』『は寒』『寒い』『いで』『です』『すね』の 6 連接が得られ、3-gram は『今日は』『日は寒』『は寒い』『寒いで』『いです』『ですね』の 6 連接が得られる。次に、文字 n-gram 頻度とは、あるコーパス内に各文字 n-gram がどれだけ出現するかを示す値である。例えば、『明日は晴れ』『明後日が晴れ』の 2 文からなるコーパスでは、文字 2-gram 統計は、『日は』『は晴』『晴れ』がそれぞれ頻度 2、『明日』『明後』『後日』がそれぞれ頻度 1 となる。

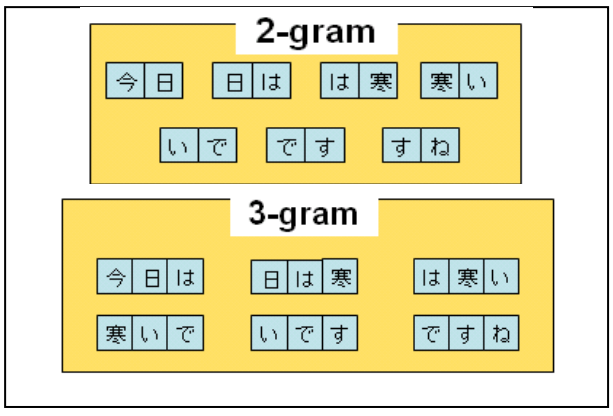

図 1：「今日は寒いですね」n-gram 分析例

以上のような方法により、コーパスの 2-gram 表現を計算した。うち、それぞれのコーパスにおいて上位に出現した表現を表 2 に示す。

表 2：2-gram での頻出表現

Wiki の表現上位	Wiki の頻度	2ch の頻度	2ch の表現上位	wiki の頻度	2ch の頻度
る。	1,207,342	81,488	って	407,570	589,544
して	799,210	323,541	ない	228,838	458,236
てい	781,615	177,572	った	595,318	371,461
た。	779,216	116,449	ww	240	356,411
ある	697,729	116,688	して	799,210	323,541

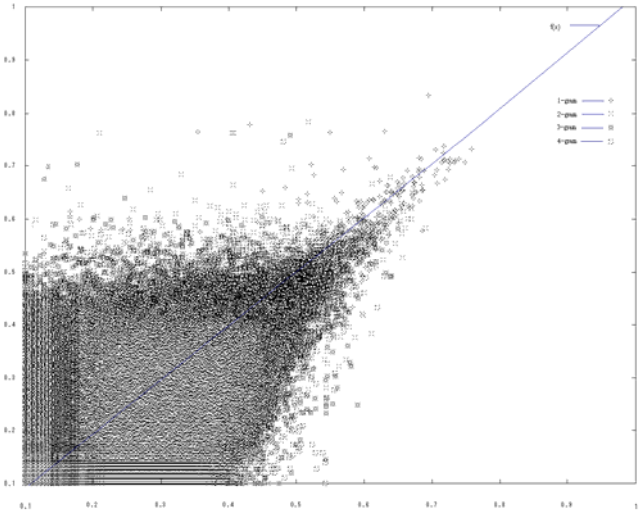

図 2: 2～4-gram をまとめた散布図

3. 文体統一性の視覚化

3.1 n-gram 表現の点数化

ターゲットとなる作文にどの程度適切・不適切表現が登場するかを計算するためには、2 節で説明した、特徴ある基準コーパスから得られた n-gram 表現の出現情報から、それぞれの n-gram 表現がどのような性質をもつかを点数化したい。そこで、本稿では、英語教育の分野で、専門用語抽出や単語のレベル分けを行った内山らの研究[5]を参考にする。 [5]では、専門用語が頻出するコーパスと一般的な用語で書かれる新聞コーパス間における、特定表現の分布の対比を、視覚的にわかりやすい形で提示する方法を提案している。

本稿の着想は内山らの有益な提示方法を表現の点数化に利用することである。内山らの方法を参考に 2.2 節で収集した 2-gram から 4-gram 表現の 2 基準コーパス間の分布を図 2 に示す。具体的には、各点は n-gram 表現の一つ一つに対応し、横軸が Wikipedia コーパスに出現する出現頻度、縦軸が 2ch コーパスに出現頻度としてプロットしてある。また、それぞれは

対数軸である。なお、[5]では、表現に対する頻度計算を単語レベルで行っているが、本稿では文字の n-gram 単位で頻度計算を行っている点が異なる。このグラフを作成するために 2 節で収集した n-gram 表現と各コーパスでの出現頻度はデータベース化を行っている。

図 2 では、特定の表現が対角線の下側に位置する場合、その表現は、より Wikipedia 側に出現する表現であり、逆に対角線より上側に位置する場合、より 2ch 側に出現する表現といえる。この点を利用し、n-gram 表現の点数化を試みる。本稿では、実験的に対象とする表現を 2-gram に絞り検討を行う。

ある 2-gram 表現 e を図 2 でプロットした時に、対角線からの距離を d(e)として、表現 e のスコアを以下のように定義する。

$$d(e) = a \times d(e)$$

$$\text{ただし } a = \begin{cases} 1 & : e \text{ が対角線より下側} \\ -1 & : e \text{ が対角線より上側} \end{cases}$$

この点数化を利用して、実際の作文事例[1] (500 文字)について、1文字目から順番に 499 文字目までの 2-gram 表現に対する点数 d(e)を縦軸プロットしたグラフを図 3 に示す。図 3 のように一つの文章は、決して不適切な表現ばかりではなく、一般に用いられる適切な表現も混在して、文体が形成されていることがわかる。本稿が対象とするのは、文章を評価する上で役立つ、文体の視覚的提示方法を工夫することである。

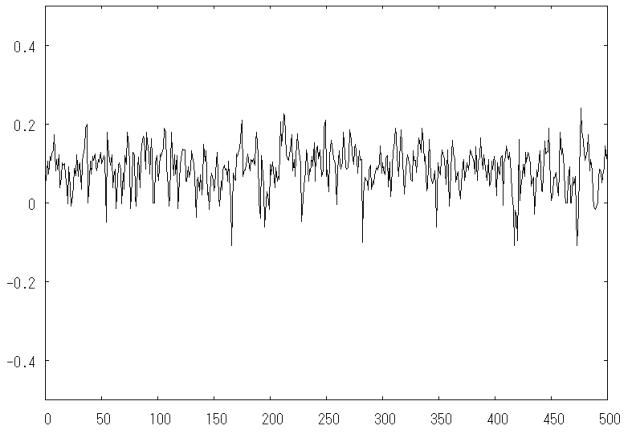

図 3　　学生の作文(500 文字)の表現点数視覚化

3.2 文体の視覚化への工夫

前述のように本稿では、文体の表示を工夫する。具体的には、通常作文には表れるべきではない不適切表現がどのように分布するかを提示することを検討したい。そこで、テスト文章として次のような文章を用意し、性質が既知のデータから、不適切表現をどのように視覚化すべきかを検討した。

1 文字目から 150 文字目を Wikipedia 文章
151 文字目から 300 文字目までを 2ch 文章
301 文字目から 450 文字目までを Wikipedia 文章
451 文字目から 600 文字目までを 2ch 文章

検討の視点としては、揺れ動きの激しい図 3 のような点数変化を句レベルの傾向を示すように滑らかにした。また、不適切表現の特徴に特化して傾向を見るため、一般表現側の変化を無視し、d(e)が負になるものだけを図 2 の対角線から離れるほど強調して表示する工夫を検討した。

検討の結果、文章の x 文字目から x+w 文字目までに含まれる 2-gram 表現の集合を返す関数を 2gram (x,x+w)とした時、以下のような関数が、文章の特徴を見る上で有効と考えた。

$$\text{score}(x,w) = -\sum_{e \in 2\text{gram}(x,x+w)} b(e)$$

ここで、 b(e)は 2-gram 表現 e の(d(e)-α)が正の時 0, 負の時$(d(e)-\alpha)^2$となる関数である。

テスト文章の一文字目から最後まで(w/2)だけずらしながら score(x,w)を計算し、プロットしたグラフを図 4 に示す。図 4 には w をそれぞれ 5, 10, 15, 20, 25 とした場合のグラフを示してある。

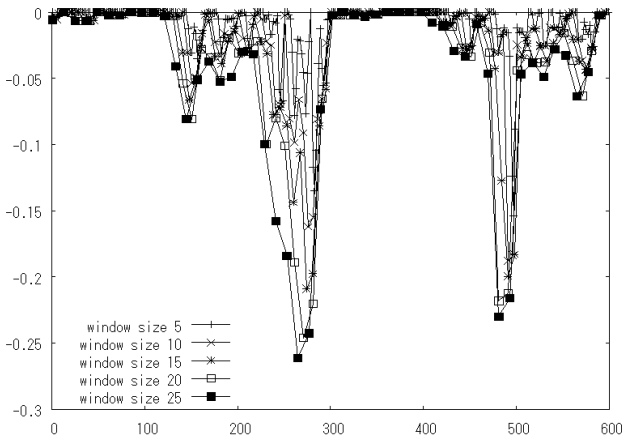

図 4　テスト文章に対する score(x,w)のグラフ

[1] 著者らの授業において大学新入生が実際に書いた作文のうち、典型的な誤りが多く見られるものをサンプルとして選択した。

この方法により実際の文章である図2で示した学生の作文と、2節で統計をとったものとは別のWikipediaの文章とをグラフ(w=10)として文体の対照表示をしたものが、図5となる。両文章とも文字数は 500 文字である(Wikipedia 文章は 500 文字程度の記事の文章末を切り、500 文字とした)。

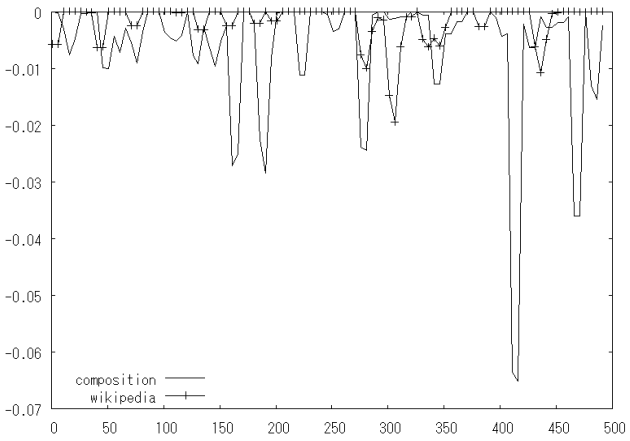

図 5. 学生の作文と新規の Wikipedia の文章に対する score(x, w)のグラフ

図 5 に示される通り、新規の Wikipedia 文章であっても、投稿された文章は一定の一般的な表現が用いられ、対照的に、図 2 で既に示した学生の作文は、不適節な表現が散在することが見受けられ、特に作文の後半部分で不適切な度合が高いことが顕著に視覚化されている。

具体的には図 5 の作文の 400 文字を超えた部分での急激な落ち込みは、「ゲームが多いということも間接的要因ではないだろうか。」という文周辺が対応し、作文前半の 150 文字目付近の落ち込みは、「今の生活がとても便利になったのも事実である。」という表現周辺である。前者の落ち込みは、「多いということも」という表現部分の不適切性が高いと同時に Wikipedia には百科事典としての文体上登場頻度が低い「ではないだろうか」という婉曲表現が、必要以上に不適切性の点数付けに影響しているとも考えられる。これは頻度情報を得た文章集合の文体の特徴が、端的に表れているとも評価できるが、Wikipedia の文章を学生が書く答案やレポートなどの模範と仮定したことがミスマッチになっていることを示しているとも言える。その意味で、より適切な模範コーパスを設定することが重要であると同時に、こうした視覚化を元に、ターゲットとなる文章とのずれを個別に見ていくことにより、適切表現コーパスに内在する文体特徴を明らかにすることも期待できる。また、適切表現コーパスを構成する文章群間においても文体の選択において揺れは当然存在し、その揺れをどのように扱うかも課題であると考える。この点については今回の試行のような 2 つだけの文章集合を仮定するのではなく、特に適切表現コーパスについて複数を仮定すると同時に n-gram 以外の頻度も合わせて考えていくことが必要であろう。

4. 終わりに

以上のように文体の統一性を診断し視覚化することを試みた。その結果、前述のように一定の結果を得ることができたと同時に、適切表現コーパスの設定についての問題点も明らかとなった。それは目的に応じた適切表現コーパスの選択という問題であると同時に、適切表現コーパスを構成する文章群にも文体選択において揺れが当然予想され、それをどのように扱うかという問題でもある。そのためには前節終わりに示した検討を行なうことが方向性として考えられるが、そのため、また、ありうべき他の方向性のためにも、作文の教育・学習支援においてニーズが高いと予想される学生の作文と模範となる文章とをいくつかのジャンルについて収集し、データベースを構築することはきわめて重要かつ有効であり、今後、そうしたデータの整備・検討をさらに行なっていきたい。また、そのような規範データに基づいて、教員がどのように文章を評価するかについても調査し、その評価基準を計量的に明らかにすることも課題であろう。

参考文献

[1] 飛田良文ら(編)『日本語学研究事典』明治書院, 2007.
[2] 国語学会(編)『国語学大辞典』東京堂出版, 1980.
[3] 日本語版 Wikipedia：http://ja.wikipedia.org/
[4] ２ちゃんねる ：http://www.2ch.net/
[5] 内山将夫,中條清美. 「単語分布と専門語彙の関連付けに関する研究」 『日本大学 生産工学部研究報告 B』, 2007 年 6 月第 40 巻 pp.13-21
[6] 日本語版 Wikipedia ダンプファイル：http://download.wikimedia.org/jawiki/20071013/jawiki-20071013-pages-articles.xml.bz2