青空文庫を対象とした書き手の識別とその応用

太田 貴久　　増山 繁
豊橋技術科学大学 知識情報工学系
{kikyu, masuyama}@smlab.tutkie.tut.ac.jp

1　はじめに

文章表現における特徴の集計・分析を行う学問として計量文体学がある．計量文体学では，文章には書き手の「指紋」ともいえる，その書き手独自の特徴があると考え，その特徴の解析を行う研究が行われている．これらの研究では，実際に高い精度で書き手の識別が行えるとする結果が得られている．しかしながら，これらの研究では実際に書き手がどのような仕組みでその特徴を作り出しているかを解明していない．このような問題に対して，本研究は，書き手が文章を生成する仕組みとして言語学で用いられる Harmonic Grammar [1] を仮定し，書き手の特徴をとらえる．今回，書き手の特徴をとらえる対象として，比較的書き手の特徴が現れやすく，かつ，大量の文章を入手可能な青空文庫を用いる．

2　関連研究

書き手を同定する研究は古くから数多く行われている．しかし，これらの研究では一般的に用いる素性が言語ごとに大きく異なる．以下では，日本語で書かれた文章を対象とした関連研究を中心に述べる．

日本語で書かれた文書の書き手の特徴を捉え，文書の著者を判別する研究は，古くから数多く提案されている．例えば，単語の長さ [2]，単語の使用頻度，品詞の使用率，機能語と識別語の使用頻度，文の長さ，読点の位置 [3] に着目した方法が存在する．これら以外にも，高精度に書き手の識別を行える素性として助詞の n-gram パターンを用いる方法が提案されている [4]．これを踏まえ，本研究では，助詞の n-gram パターンを基本的な書き手の特徴として用いる．さらに，本研究では，読点の打ち方や品詞の遷移などの素性も同時に利用する．

近年，書き手の同定に，SVM やランダムフォレストといった機械学習・パターン認識の技術が用いられるようになった [5]．これらの研究では，特定の素性を用いて書き手の同定を行うのではなく，多くの素性を入力し，それらを複合的に扱い，非常に高精度に書き手の同定を行う．しかし，これらの研究は，N 人の識別を N クラスの分類問題として取り扱うので，学習データの関係上，識別を行う人数を予め決めておかねばならないという問題がある．これに対して，本研究は，識別を行う書き手の間で直接素性の比較を行わないため，識別する人数を予め決める必要はない．

3　書き手の文章生成モデル

本研究では，書き手の文章生成モデルとして Harmonic Grammar を仮定している．Harmonic Grammar は，言語学で用いられる理論で，「制約」とその「重み」によって音声や文の生成を説明する理論である．本研究における例（表 1）を用いて説明する．表 1 において，出力の候補として「豊橋の南の丘の上にある学校が技科大だ」と「豊橋の南にある技科大は丘の上にある」の 2 つの文が存在する．また，出力を決定するための制約として助詞「の」の 3 回連続使用の禁止（not[の][の][の]）と助詞「は」の使用禁止（not[は]）が存在する．ここで，制約「not[の][の][の]」には 2.0 という重みが，制約「not[は]」には 1.0 という重みが与えられている．このとき，出力は Harmony（$\mathcal{H}$）と呼ばれる値が高い「豊橋の南にある技科大は丘の上にある」が選ばれる．Harmony の計算は容易で，各制約の重みとその制約の違反数（負の数）の積の和で計算される．具体的には，「豊橋の南の丘の上にある学校が技科大だ」は 2.0　1＋1.0　0＝　2.0 で，「豊橋の南にある技科大は丘の上にある」は 2.0　0＋1.0　1＝　1.0 となる．この結果，Harmony がより大きい「豊橋の南にある技科大は丘の上にある」が出力される．本研究では，各書き手は，それぞれ異なる制約の重みを持つと仮定し，これを書き手の特徴とする．

表 1: Harmonic Grammar の例

	Weight	2.0	1.0	$\mathcal{H}$
	候補	not [の][の][の]	not [は]	
	豊橋の南の丘の上にある学校が技科大だ	−1		-2.0
⇒	豊橋の南にある技科大は丘の上にある		−1	-1.0

3.1　制約

本研究で用いる制約について説明する．本研究では表 1 にあるような特定の助詞の使用禁止といった日本語に特有の表層的な制約を用いる．本研究で用いる制約を以下に示す．

助詞制約 提案手法において主となる制約で，特定の助詞の出現を禁止する制約である．例えば，制約「not[に]」では，助詞「に」の出現を禁止する．

読点制約 読点が存在することを禁止する制約である．

自立語品詞制約 文節に含まれる自立語の品詞に関する制約で，文節が特定の品詞を含むことを禁止する制約である．

本研究では以上の 3 種類の制約を基本的な制約として用いる．しかし，これらの制約では多くの文章で違反してしまい，書き手の特徴を捉えることは不可能である．そこで，各制約を拡張した制約を作る

表 2: 書き手の識別実験の結果

	上位 1 位	上位 3 位
正解率	84.1%	87.2%

ことで，書き手の特徴を捉える．拡張は大別して 2 種類存在する．1 つは助詞制約と自立語品詞制約に対する拡張で，チェックする系列の長さを伸ばす n-gram 拡張である．例えば，表 1 で用いた助詞「の」の 3 回連続での使用禁止「not[の][の][の]」が存在する．もう 1 つは，3 種類の制約に用いれられる拡張で，制約の結合 (Conjunction) による拡張である．例えば，助詞制約「not[の]」と読点制約の結合として，「助詞 [の] の直後に読点を打つことを禁止する」といった制約がある．

3.2 制約の重みの推定

本研究では Goldwater らの研究 [6] と同様に，最大エントロピー法によって各制約の重みを推定する．すなわち，入力 i から出力 o が生成される確率を以下の式で定義する．

$$P(o|i) = \frac{1}{Z(i)} \exp\left(\sum_k w_k v_k(o,i)\right) \quad (1)$$

$$Z(i) = \sum_{o \in \mathrm{Gen(i)}} \exp\left(\sum_k w_k v_k(o,i)\right) \quad (2)$$

ここで，w_k が制約 c_k の重みを表す．また，$v_k(o,i)$ は入力 i のもとで出力 o が制約 c_k に違反する数，$\mathrm{Gen}(i)$ は入力 i に対する出力候補の集合を表す．上記の $P(o|i)$ は対数をとることで，Harmonic Grammar の Harmony と同様の定義となることがわかる．

4 書き手の識別実験

青空文庫において、10 冊以上の作品が登録されている作者 50 名，1000 冊を対象に書き手の識別実験を行った．識別実験は以下の手順に従って正解を判定し，leave-one-out 法によって行った．

Step 1 識別対象以外の文章を用いて作者ごとに制約の重みを学習する．

Step 2 識別対象の文章 S の生成確率 $P(S)$ を各作者ごとに以下の式で求める．

$$P(S) = \prod_{s \in S} P(s|s') \quad (3)$$

ここで，s' は文 s の入力を表す．

Step 3 生成確率が高い上位 n 人以内に対象文章の作者が入っていた場合に正解とする．

以上の手順を全ての文章に適用し，正解率を求めた．結果を表 2 に示す．

表 2 のように，高い精度で書き手の識別が行えている．この結果は，ランダムフォレストを用いた金の研究 [5] の結果より劣っている．しかし，金の研究 [5] では，新たに識別した書き手を増やしたときに再度学習が必要な点などの問題点があるため，本研究と金の研究 [5] を直接比較することはできない．

表 3: 年代判別実験の結果

	上位 1 位	上位 3 位
精度	21.3%	50.5%

5 文章の年代判別実験

書き手の識別と同様の手法で，文章の年代判別を試みた．本実験では 1900 年〜1950 年の期間に発表された文章を実験対象とし，実験対象を 5 年単位で区切り（計 10 クラス），前節の実験と同様に leave-one-out 法によって実験を行った．結果を表 3 に示す．

表 3 より，書き手の識別に比べ，年代判別は精度が低いことが確認できる．これは，今回用いた書き手の特徴（制約）の多くは書き手に依存するもので，年代に依存するものが少なかったためと考えられる．使用する制約の選別を行えば今回用いた制約だけでも高精度に年代判別できる可能性はある．

6 おわりに

本研究では，Harmonic Grammar に基づく書き手の識別法を提案した．今回，助詞の出現パターン，読点の打ち方，品詞の出現パターンに着目し書き手の識別を行った．青空文庫を対象に書き手の識別実験を行った結果，比較的高い正解率で書き手の識別を行うことに成功した．また，同一の手法を用いて年代判別実験を試みた結果，今回用いた書き手の特徴をそのまま用いるのみでは年代判別に適していないことがわかった．今後，書き手の識別性能の向上を目指すと共に，使用する制約の分類を行う（書き手に依存する制約や年代に依存する制約を明らかにする）必要がある．

謝辞

本研究は文部科学省グローバル COE プログラム「インテリジェントセンシングのフロンティア」の支援により行われた．

参考文献

[1] S. Smolensky and G. Legendre: “the harmonic mind : from neural computation to optimality-theoretic grammar”, MIT Press (2006).

[2] 金：“動詞の長さの分布に基づいた文書の分類と和語および合成語の比率”, 自然言語処理, **2**, 1, pp. 57–75 (1995).

[3] 金：“読点の打ち方と著者の文体特徴”, 計量国語学, **19**, 7, pp. 317–330 (1994).

[4] 金：“助詞の n–gram モデルに基づいた書き手の識別”, 計量国語学, **23**, 5, pp. 225–239 (2002).

[5] 金, 村上：“ランダムフォレスト法による文章の書き手の同定”, 数理統計, **55**, 2, pp. 255–268 (2007).

[6] S. Goldwater and M. Johnson: “Learning ot constraint rankings using a maximum entropy model”, Proceedings of the Workshop on Variation within Optimality Theory, pp. 111–120 (2003).