自然言語処理 134－5
（1999．11．25）

n-gram分布を用いた近代日本語小説文の著者推定

松浦 司* 金田 康正**

*東京大学理学系研究科情報科学専攻 **東京大学情報基盤センター

概要

本論文では、文章中のn-gram分布状況を著者の特徴量として、文章の著者を推定する手法を提案する。また、著者を特徴づけるn-gramを自動的に抽出する手法をも提案する。本手法は、テキストの構造や品詞、単語間の区切れなど文章に関する付加的な情報を必要とせず、かつ、叙述に際しての規則・習慣や文法といった特定の言語の性質を利用しないため、多くの言語・テキストにそのまま適用可能であることが期待できる。また、分析に利用するテキストは短いテキストをいくつか集めたものでも構わないので、ひとまとまりの長い文章を残していない書き手の判別にも本手法は有効である。最後に、3-gram用い、最短2万2336文字の近代日本語小説文の著者が判別できた著者推定実験結果について報告する。

Identifying Authors of Sentences in Japanese Modern Novels via Distribution of N-grams

Tsukasa Matsuura* and Yasumasa Kanada**

*Department of Information Science, Division of Science, University of Tokyo
**Information Technology Center, Computer Centre Division, University of Tokyo

Abstract

We propose a method to discover characteristics of authors represented by N-gram distribution in sentences automatically. The characteristics which our method educes can identify the authorship. In our experiments, the author of sentences with 22,336 characters in Japanese modern novels at least can be identified via 3-grams. Our method does not require any additional information on sentences and words – for example, contextures, word classes, or separators between words. Furthermore, no characteristics of certain languages – for example, rules or habitudes of expression, or grammars, contribute to the analysis. These two points make our method applicable to sentences in various languages. Our method is effective not only on congruous compositions, but also on text data which consist of some fragments of compositions. Therefore, authors who have not written long compositions can be identified by our method.

1 はじめに

従来の著者推定に関する研究では特定品詞の使用状況を手がかりにするなど、予めどのような特徴に注目すれば著者を判別できるかを仮定し、観点を固定して分析を進めるものが多かった。本研究ではn-gram分布を使用することで、直感や経験に頼った仮定をおくことなく著者の推定を行う。更に、著者を特徴づけるn-gramを自動的に取り出し、著者の特徴を人間が理解するため、またより高精度の特定著者判別手法を構成するための手がかりとして提供する。本論文で示す手法では文章に対して形態素解析などの前処理を施す必要が全くないため、日本語のように単語区切り・形態素解析を行うことが困難な言語にも適用可能である。また解析に当たり各言語特有の性質を利用しないので、使用言語にあわせてチューニングを行う必要がないことも大きな利点である。

2 文章の著者推定

筆跡が人によって異なるように、文章の書き方には人によって一定の傾向、即ち「癖」があるとして、計量的に文章の癖を把握しようという試みは、[村上 1988]によれば1851年のAugustes de Morganの書簡にまで遡ることができる。そして書き手の性格が計量的に捉えられるものであれば、文章の著者も計量的根拠に基づいて推定することが可能であろうと期待される。

Morgan以後、欧米において専らヨーロッパ言語で書かれた文章の著者判別の研究が進められた。一語あたりの平均文字数[Mendelhall 1901]や、一文あたりに含まれる語数[Yule 1939] [Yule 1944] [Wake 1957] [Kjetsaa 1984] 、単語の使用頻度 [Ellegard 1962] [Mosteller 1964]を著者の個性を表す特徴量として使用し、著者判別を試みた研究がある。これらの手法は全て、文章中の単語の切れ目が自明であることを前提としている。日本語を含め単語の区切りを明確にすることが難しい言語に対しては、以上の研究成果をそのまま活かすことは難しい。

日本語に関しては、近年、複数の形態素解析器が開発され、文章に単語の区切りや品詞、原形を示すタグを自動的に埋め込むことが可能になった。しかし、どれほど高性能に作られた解析器であっても、その解析精度には限界がある。また確立した日本語文法が未だ存在しないために、形態素解析器ごとに採用している日本語文法が異なる状況がみられる。形態素解析された文章を題材に研究を進める以上、使用する形態素解析器の精度、及び解析器が採用しているアルゴリズムの性格、文法に研究成果が左右される。そのため、異なる形態素解析法を用いた関連研究結果間での結果の比較・統合が困難になり、各研究が孤立してしまい、後続研究に研究成果が活かされにくい難点がある。

日本語を対象とする先行研究は、形態素解析済みの文章を扱うものが殆どである。[金 1993a] [金 1993b] [金 1994] では読点の直前に現れる品詞に着目したり、読点間の文節数を数え上げたりしている。後述する[金 1997]や[村上 1999]、[安本 1994][韮山 1965]でも、解析に使用する文章は形態素解析が適切に施されていることが前提となっている。

日本語の文章、更には特定の日本語の文章に特化した解析方法も研究されている。助詞・助動詞の使用状況を手がかりとする研究[金 1997][村上 1999] や、読点の使用方法に着目した研究 [金 1993a][金 1993b][金 1994]、漢字・名詞の使用度、更には特定の語（色彩語や直喩、声喩、人格語）の使用状況まで分析に利用する研究

[安本 1994][韮山 1965] が挙げられる。勿論、特定の言語や文献の特徴に注目する研究方針は欧語文献研究にもみられる。例えば、ギリシア語では二重母音が回避される事があるとして回避された二重母音の数を手がかりとした研究 [Raeder 1906] や、ヘブライ語の接頭辞の使用状況に着目した研究 [Adams 1973] が挙げられる。これらの研究は一般性に欠けることが第一の問題点として挙げられる。他言語への応用が難しいだけでなく、特定の文献など極狭い範囲の文章にしか適用できない手法もみられる。また、経験・直感に基づく仮定はその信憑性を検証することが難しい。

3 文字を分析の単位とする著者推定

区切りを施すことが難しい日本語であっても、少なくとも現代においては[1]文字の区切りは自明である。そこで、単語や文節ではなく、文字を単位として解析を行うことを提案したい。

文字を単位とする解析は、言葉の意味や品詞を全く考慮出来ないという短所を有する。特定の品詞や、特定の意味カテゴリに属する言葉の使用状況が顕著に書き手の個性を表す場合でも、そうした特徴量を検知できない可能性は否定できない。しかし、文字列の出現頻度に偏りを生じることなしに特定カテゴリに属する言葉を多用することは寧ろ難しいと思われる。類似の文字列の出現頻度に偏りがあれば、文字を単位とする解析でも検知することができ、それを手がかりに特定カテゴリ語の使用状況の特性を把握することも可能であると考える。

[1]日本語には以前、頻出する文字の組み合わせをひとつの文字に合成して使用する習慣が存在した。

3.1 n-gram 分布

本手法では文字を単位に分析を行うが、文字同士の隣接状況を分析に反映させるために文章の n-gram 分布を使用する。n-gram 分布とは、n 個の文字が隣接して生じる文字の共起関係、即ち n-gram の出現確率を記録したものである。文字同士の隣接状況を無視し、一文字一文字の出現確率を記録したい場合には、n の値を 1 にすればよい。

3.2 文章間の類似度計算

文章の著者推定を行うためには、文章の「癖」に関する類似度を文章間で計算する必要がある。本手法では、まず各文章について n-gram 分布を求め、求めた各 n-gram 分布に対して確率分布間の距離を測る尺度を適用することで各文章間の類似度を求める。確率分布間の類似度計算方法に関しては多くの先行研究が存在しそれらの成果を利用することが期待できるが、今回は以下の類似度計算関数 sim を用いた実験結果について報告する。

長さ n の文字列 (n-gram) を x で表す。或る言語のアルファベットを χ とすると x は以下の様に表される。

$$\mathrm{x}= \{x_1 x_2 \cdots x_n | \forall i. x_i \in \chi\}$$

文章 P 及び Q の n-gram 分布がそれぞれ確率分布関数 $P(\mathrm{x})$ 及び $Q(\mathrm{x})$ で表されるとき、文章 P, Q 双方に出現する n-gram(共通 n-gram) の集合を C とし以下の様に定義する。

$$C = \{\mathrm{x} | P(\mathrm{x})Q(\mathrm{x}) \neq 0\}$$

文章 P, Q 間の類似度を以下の $sim(P, Q)$ によって表す。但し $\mathrm{card}(C)$ は集合 C の要素数を表す。

$$sim(P,Q) = \frac{1}{\mathrm{card}(C)} \sum_{\mathrm{x} \in C} |\log(P(\mathrm{x})/Q(\mathrm{x}))|$$

文章間の共通 n-gram の出現確率の比がそれぞれ 1 に近い場合に文章が類似しているとみなすと、この $sim(P,Q)$ は、文章 P と Q とが類似しているほど小さな値を示す。P と Q とが全く同じであるときに限って sim の値は 0 をとる。また、sim は対称性を有する。

3.3 実験の詳細

本稿では近代日本語小説である、芥川龍之介の短編小説 10 作品 (9 万 892 文字) と菊池寛の短編小説 8 作品 (11 万 1705 文字) とを使用して、類似度 sim を用いた著者判別の可能性を検討した実験について報告する。注目する共起文字列の長さ n は 3 に固定した。長い共起文字列に注目すれば文字の隣接情報を活用でき分析精度の向上が期待できるが、比較する文章の長さに対して n が大きすぎると共通 n-gram の数が小さくなり過ぎて却って精度の低下を招くであろうことを考慮して、この n の値を決定した。しかし妥当な n の値の選択に関しては今後検討の必要がある。

使用した作品とその文字数を表 1 に示す。全作品に関して、実験には青空文庫[2]で公開されているテキストを使用した。

作品の文章には作品名や著者名、章名は含めないが、改行や空白はそれぞれ一文字とみなす。但し、空行はデジタル化の際に整形のために挿入されたものとして予め削除してある。また改行後の空白は段落を改める意味を持つが、直前の改行によって段落の区切りは示されており冗長であるのでこれも削除した。

[2]http://www.aozora.gr.jp/main.html
ボランティアの手による WWW 上の電子図書館。著作権が消滅した文学作品を中心に無償公開を行っている。

各作品の文字数はまちまちであり、かなり短いものが含まれるため、作品間の類似度を作品毎に著者判別を試みるのは難しい。理由としてはまず、余りに短い文章では十分に著者の特徴を表しきれないことが挙げられる。次に、比較するテキスト間で文字数が極端に異なると値が大きくなるという sim の性質がある。つまり、両テキストに同じ回数出現する n-gram が sim の値を大幅にひきあげる働きをするが、限られた長さのテキスト中における特定 n-gram の低頻度の出現が書き手の文章中のその n-gram の分布をよく表していることは期待できない。そのため信憑性の低い要因によって sim の値が大きく引き上げられ、類似度としての精度が低下する可能性がある。

そこで、十分な文字数のテキストを確保し、また各テキストの文字数を平均化する必要ために、無作為な順番で作家別に全作品をつなぎ合わせ、それをほぼ同じ大きさのテキスト群に分割して使用することとする。但し、一つの作品が複数テキストに分割されることは許すが、段落が分割されることは禁止する。また作品の切れ目を含む n-gram に対しては作品の切れ目以降の文字を無効にし、複数作品にまたがる n-gram を抽出することを防ぐ。

3.4 評価

作家毎にランダムな順番 (表 1 の順) で作品をつなぎ合わせたテキストを 4 分割した場合の各テキスト比較結果をそれぞれ図 1 に示す。同一作家のテキスト間の sim は、異作家のテキスト間の sim よりも常に小さくなっていることから、著者判別が行えている様子が見て取れる。

また、テキスト分割数と失敗率との関係をまとめた

ものが表2である。3分割、4分割の場合には芥川龍之介の作品と菊池寛の作品とが完全に判別出来ている。5分割を行った場合には、テキスト「芥川1」を基準にとった場合だけが失敗している。また表2では、6分割以上に細かくテキストを分割した場合には、判別に失敗する例が多くなることが観察される。よって、4分割時のテキストの平均長である2万5000文字程度の文章であれば、この手法で著者判別を行えるであろうことが予想される。猶、異なるつなぎ合わせ順のテキスト5例に関しても同様に、4分割後のテキストは100%の精度で著者判別できた。

著者判別のために分割テキスト間の類似度simを計算したが、各3-gramがsimに及ぼす影響の大きさを図2にまとめた。このグラフを見ると、どのテキストにも類似した頻度で出現する3-gramと、テキストを特徴づける3-gram、即ちテキスト間で著しく出現頻度が異なり$|\log(P(\mathrm{x})/Q(\mathrm{x}))|$の値が大きい3-gramとはほぼはっきりと分かれていることが分かる。また、simに対する影響の小さい3-gramの$|\log(P(\mathrm{x})/Q(\mathrm{x}))|$の値は類似しているものが多く、小数点以下2位までの精度で計算するとその値は殆ど、モード[3]以下になる。

そこで、$|\log(P(\mathrm{x})/Q(\mathrm{x}))|$の値がモードより大きい3-gramをテキストを特徴づける3-gramと見なし、分割後の各テキストを比較したデータから芥川、菊池両者の文体を特徴づける3-gramを取り出すことを試みる。異著者のテキスト比較において常に$|\log(P(\mathrm{x})/Q(\mathrm{x}))|$が大きな値をとる3-gramが芥川、菊池両者の文体の差異を示していることは明らかであろう。しかし、それらの3-gramの中には文章の内容に起因して出現頻度が変化しているものも含まれていると考えられる。そこで、同一著者のテキスト比較でも$|\log(P(\mathrm{x})/Q(\mathrm{x}))|$の値が大きい3-gramは作家を特徴づける3-gramから除く必要がある。

いずれかの同一著者テキスト比較において$|\log(P(\mathrm{x})/Q(\mathrm{x}))|$の値がモードより大きい3-gramを除いて得た、作家の特徴を表す3-gramは「えない」「の代り」「をしな」の3つであり、いずれも芥川による使用頻度が菊池のそれよりも多かった。常に同一著者テキスト比較において$|\log(P(\mathrm{x})/Q(\mathrm{x}))|$の値がモードより大きい3-gramを除いた場合には、112種の3-gramが著者を特徴づけるものとして抽出できた。これは人間が検討可能な量であり、作家の文体を考える上で大きな手がかりになるものと考えている。

4 まとめと今後の課題

芥川龍之介及び菊池寛の短編小説合計18篇を分析した実験において、3-gram分布を使って2万5000文字以上の文章の著者を判別することが可能であることが確認できた。分析するテキストはひとまとまりの長いものである必要はない。このことは長大な文章を書き残していない書き手の判別をも可能にする。現実的応用として、訴訟事件で特定文書の書き手を特定することが必要になった場合にも、本手法は効果を発揮することが期待される。また著者を特徴づける3-gramを抽出することも試みた。

本稿では日本の近代小説家二人の文章を3-gram分布を用いて分別できることを報告し、n-gram分布を用いて文章の著者推定が行える可能性を示唆した。今後は、より多様な書き手・文章に関してデータを集め、本手法の妥当性を検証する必要がある。それとともに、1-gram、2-gram、や4-gram分布以上に基づいた分析の可能性を検討する予定である。

[3]分布において最も多い回数出現する値。最頻値。

表 1: 実験に使用したテキスト

作者	作品名	文字数	作者	作品名	文字数
芥川龍之介	あばばばば	6294	菊池寛	青木の出京	2 万 1546
芥川龍之介	アグニの神	8070	菊池寛	勲章を貰う話	1 万 3659
芥川龍之介	秋	1 万 1051	菊池寛	身投げ救助業	5941
芥川龍之介	あの頃の自分の事	1 万 5770	菊池寛	無名作家の日記	2 万 1029
芥川龍之介	或阿呆の一生	1 万 2209	菊池寛	大島が出来る話	9926
芥川龍之介	浅草公園——或シナリオ——	6944	菊池寛	恩讐の彼方に	2 万 2811
芥川龍之介	或敵打の話	9260	菊池寛	出世	8868
芥川龍之介	或旧友へ送る手記	3221	菊池寛	ゼラール中尉	7925
芥川龍之介	或日の大石内蔵助	8825			
芥川龍之介	一塊の土	9248			
	合計文字数	9 万 892		合計文字数	11 万 1705

表 2: テキスト分割数と分析に失敗する基準テキストの取り方の率との関係

テキスト分割数	テキストに含まれる文字数			失敗する基準テキストの取り方の割合 (%)
	最大	最小	平均	
3	3 万 7652	2 万 9603	3 万 3767	0
4	2 万 8136	2 万 2336	2 万 5325	0
5	2 万 2704	1 万 6959	2 万 260	10
6	1 万 8972	1 万 4657	1 万 6884	17
7	1 万 6220	1 万 2587	1 万 4472	29
8	1 万 4415	1 万 930	1 万 2663	38
9	1 万 2749	9555	1 万 1256	39

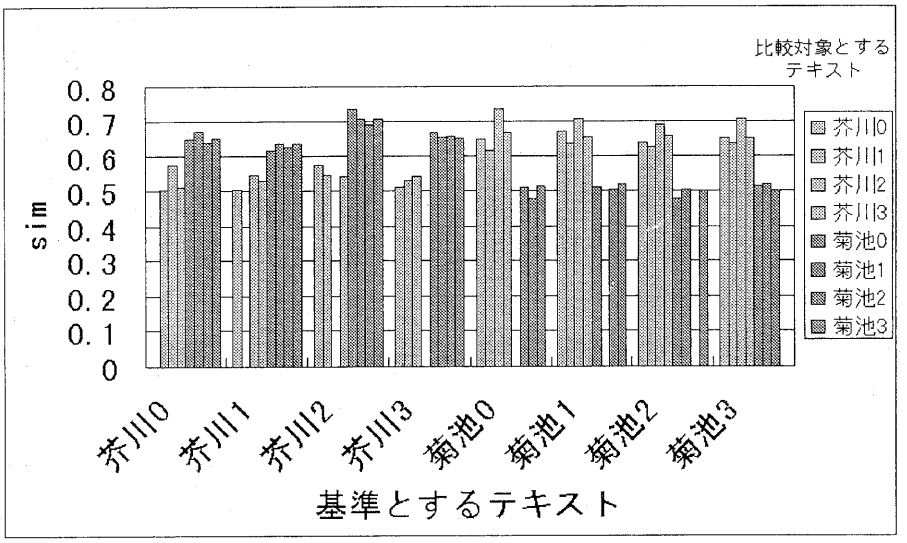

図 1: 作家別に全テキストを 4 分割して比較した結果
(各基準テキストに対する各テキストの *sim* を添え字の若い順に左から並べて表示してある。)

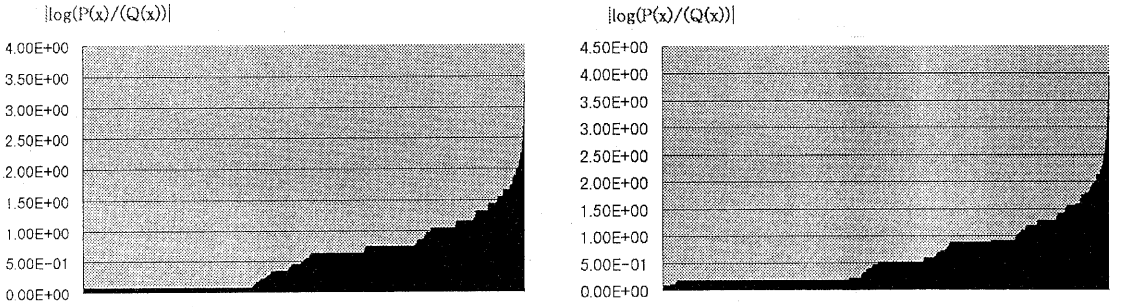

図 2: 各 3-gram の sim に及ぼす影響
(sim の値に与える影響が小さい順に 3-gram が横軸に並べてある。
左は 3 分割後の芥川著テキスト同士を比較した場合、
右は 3 分割後の芥川著テキストと菊池著テキストとの比較を行った場合である。)

参考文献

[金 1993a] 金明哲 読点と書き手の個性, 計量国語, Vol.18, No.8, pp.382-391, 1993

[金 1993b] Jin M.Z et al., Authors' Characteristic Writing Styles as Seen Through Their Use of Commas, Behaviormetrika, Vol.20, pp.63-76, 1993

[金 1994] 金明哲, 読点の打ち方と文章の分類, 計量国語, Vol.19, No.7, pp.317-329, 1994

[金 1997] 金明哲, 助詞の分布に基づいた日記の書き手の識別, 計量国語学, Vol.20, No.8, pp.357-367, 1997

[村上 1988] 村上正勝 著者はだれか?計量文献学への招待 1, 数学セミナー, 11月号, pp.55-59, 1988

[村上 1994] 村上征勝, 計量的文体研究の威力と成果, 言語, Vol.23, No.2, pp.30-37, 1994

[村上 1999] 村上征勝ら, 源氏物語の助動詞の計量分析, 情報処理学会論文誌, Vol.40, No.3, pp.774-782, 1999

[韮山 1965] 韮山正, 由良物語の著者の統計的判別, 計量国語学, No.33, 1965

[安本 1994] 安本美典, 文体を決める三つの因子, 言語, Vol.23, No.2, pp.22-29, 1994

[Adams 1973] Adams, L.L. and Rencher, A.C., The Popular Critical View of the Isaiah Problem in Light of Statistical Style Analysis, Computer Studies in the Humanities and Verbal Behavior vol.7(3-4), 1973

[Brinegar 1963] Brinegar C.S, Mark Twain and Quintus Curtius Snodgrass Letter, J.Amer.Stat.Ass., 58, pp.85-96, 1963

[Ellegard 1962] Ellegard, A., A Statistical Method for Determining Authorship: The Junius Letter 1769-1772, Gothering Studies in English, No.13 Acta Universitatis Gotheburgensis

[Kjetsaa 1984] Kjetsaa, G. el al., The Authorship of the Quiet Don, Solum Forlag A.S., Oslo; Humanities Press, New Jersey, 1984

[Mendelhall 1901] Mendelhall, T.C., A Mechanical Solution of a Literary Problem, Popular Science Monthly, Vol.60, No.2, 1901

[Mosteller 1964] Mosteller F. et. al., Inference in an Authorship Problem, J.Amer.Stat.Ass., 58, pp.275-309, 1964

[Raeder 1906] Reader, H., Über die Echtheit der Platonischen Briefe, Pheinisches Museum für Phillgie, F.F.LXI., 1906

[Wake 1957] Wake W.C., Sentence-Length Distribution of Greek Authors, J.R.Statist.Soc.A, 20, pp.331-346, 1957

[Yule 1939] Yule G.U., On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to two Cases of Disputed Authorship, Biometrika, 30, pp.363-390, 1939

[Yule 1944] Yule G.U., The Statistical Study of Literary Vocabulary, Cambridge University Press, 1944