3A4-2

N-gram と決定木による筆者識別

Authorship Detection of Sentences based on Decision Tree Learning with N-gram Distribution

谷口 裕大 Yuta TANIGUCHI　殿生 剛士 Takeshi TONOO　杉村 博 Hiroshi SUGIMURA　松本 一教 Kazunori MATSUMOTO

神奈川工科大学大学院 情報工学専攻
Course of Information and Computer Sciences, Graduate School of Kanagawa Institute of Technology

This paper proposes a method for authorship detection of Japanese sentences using decision tree learning over N-gram distribution. We in many cases can observe individuality of authors in terms of measurable features. In the case of Japanese texts, we often apply morphological analysis and obtain additional grammatical information from the texts. This approach depends on the performance of the morphological tools. To avoid this problem, we in this study use distribution of N-gram that are sequences of N letters simply cut out from texts. We first experimentally investigate the best value of N that is expected to have the highest distinguish ability in authorship detection. We further discuss the feature selection method over a large set of N-gram.

1. はじめに

本研究では筆者を自動的に推定するシステムを開発する．筆者推定を行う研究にはいくつもの研究があり，成功を収めている [金 07]．岡田らは形態素解析によって特徴量を取得して推定を行っている [岡田 05]．しかし，このような高次元の辞書を作成するコストは膨大で，また，形態素解析による精度の問題も常に抱えてしまうこととなる．

また，一語あたりの文字数や一文あたりの語数を特徴量として筆者推定を行う手法もこれら辞書を必要としないが，文章中における単語間の区切れが明確であることを前提としているので，日本語や中国語などの単語分かち書きが難しい言語で書かれた文章に直接適用することはできない [松浦 00]．

本研究では，このように辞書を頼りに特徴量を計算する手法ではなく，また言語にも依存しない筆者推定システムを開発することを目的としている．

2. 筆者推定システム

本システムは二つの特徴を持つ．一つ目は N-gram によって解析文章から自動的に特徴となる単語を抽出することによって，シソーラスや専門辞書，形態素解析などに用いる高次元の辞書を必要としない点である．多くの研究はこのような辞書を用いて特徴量を取得しているが，本研究では N-gram を用いることで対象となる文章の全ての隣接文字を，機械的に抽出し自動的に辞書を作り出す．この方法によって，文章に関する付加的な情報を全く必要としないシステムとなる．

二つ目は特定の言語や文章の性質を利用しないために，多くの言語に対してそのまま適用可能な点である．欧米の言語で書かれた文章の筆者判別の研究として，一語当たりの平均文字数や使用頻度，一文あたりに含まれる語数などを筆者を表す特徴量として使用し，筆者判別を試みた研究がある [松浦 99]．しかし，日本語や中国語などは，単語分かち書きが難しく，また読点の使用方法の文法が曖昧な言語に対してはそのような手法の適用が難しい．この問題に対して，N-gram を用いて機械的に文章を切り出すことによって文章構造を理解する必要がなくなる．

また N-gram によって抽出した出現回数が高い文字列から順に変数選択を行うことで，ゼロ頻度問題に対応しつつ分類を行う．

2.1 システムの概要

分類したい筆者の作品をいくつか収集し，データベースに格納する．次にテキスト群の N-gram 分布を降順に並べ，属性候補辞書としておく．また各作品毎の N-gram の共起頻度も求める．属性候補辞書の上位から順番に一つ選択し，その属性候補から決定木を用いて実際に各作品の分類を行い，分類精度を求める．そしてすべての属性候補のうち，最も分類精度の高くなった属性候補を属性として確定し，確保リストへ記録する．確保した属性候補を属性候補リストから削除し，新たな属性候補リストの上位から順番に一つ選択した属性と，確保した属性を使い再び決定木を作成し各作品の分類精度を求める．このような変数選択を用いて分類精度が上昇する属性候補の全てを決定する．

この提案するシステムの概要を図 1 に示す．

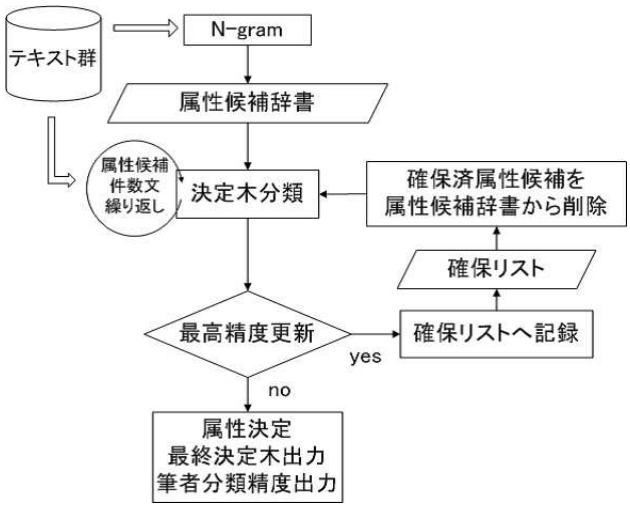

図 1: システムの概要図

連絡先: 谷口 裕大，神奈川工科大学 情報工学専攻， 神奈川県厚木市下荻野 1030，taniguchi.yuhta@gmail.com

2.2 決定木学習

本研究では決定木学習によって，筆者推定を行うためのルール抽出を行う．決定木学習はトレーニングデータを分割するための枝を生成する．生成した枝によってトレーニングデータを分割した際の評価値を計算し，その中の最大の評価値をとる分割によってトレーニングデータを分割する．評価値の計算方法として利得比基準 (gain ratio) を用いる [稲積 00].

本研究ではこの手法を用いて，筆者推定のためのルールを抽出する．属性は N-gram 分布とし，クラスは筆者とするトレーニングデータを作成する．N-gram 分布を用いたトレーニングデータの作成方法については次節で説明する．

2.3 N-gram 分布を用いたトレーニングデータ

文字同士の隣接状況を用いて解析を行う．このために本研究では N-gram 分布を用いる．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，すなわち N-gram の出現確率を記録したものである [松浦 99].

N-gram 分布を用いたトレーニングデータは図 2 のようになる．

りして	だと思	てきた	ように	class
0.009	0.003	0.006	0.003	芥川
0.012	0.012	0.021	0	芥川
0.03	0.006	0.024	0.012	芥川
0.024	0.009	0.018	0.003	豊島
0.012	0.003	0.015	0.012	豊島
0.033	0.003	0.027	0.006	豊島

図 2: N-gram 分布を用いたトレーニングデータ

2.4 変数選択

変数選択では，変数が一つも含まれていない空のモデルから出発し，変数を一つずつ増加させて変数選択を行う．本研究ではこの手法をもちいて，N-gram によって抽出した膨大な文字列集合から，分類を行う際に最適な文字列集合を自動的に発見する．さらに，頻出する文字列の上位 f 個に限定して変数選択を行うことによってゼロ頻度問題に対応する．

まず，N-gram によって全作品から部分文字列集合を抽出する．すべての部分文字列集合から，頻出順に上位 f 個を取り出し，属性候補とする．つぎに，出現上位から順に属性候補をとりだし，各作品の N-gram 分布を計算してトレーニングデータを作成する．そしてこのトレーニングデータから決定木学習を行うことによって，属性候補から得られる分類精度を計算する．

この作業を，分類精度が向上しなくなるまで属性候補を抽出する．このようにして抽出した最終的な属性候補の集合が，筆者推定を行う際の最適な属性となる．

3. 実験

本システムは類似文章の筆者を特定する目的で使用することを想定している．このため，実験では類似文章と思われる，新字新仮名で記述されたテキスト群によって行う．実験は 2 種類のテキスト群で行う．一つ目は，文学作品の中での筆者推定である．この実験ではランダムに収集した筆者 10 名，各 3 作品を扱う．二つ目は文学作品の中でも共通の思潮により類似していると思われる無頼派と呼ばれる一派の筆者 5 名，各 3 作品を扱う．

分類対象には青空文庫のテキストデータを引用した．そのままのテキストデータの中には，データ入力者によるルビや注釈，底本情報などが含まれており，そのまま本システムを適用してしまうと筆者の文章には直接かかわらない文字が抽出されてしまう．このために，前処理としてこれらの情報を取り除いている．また，作品ごとに文章量が違うこと，書き手の特徴が表れるには 2 万字以上で構成される必要がある事 [松浦 99] から，すべてのテキストの文字数を冒頭から取得した 2 万字に合わせている．

決定木作成と筆者分類には Weka を，決定木アルゴリズムには C4.5 を，精度測定には交差検定を用いている．また N-gram の N の値を変化させることで，測定精度の違いを検証する．

実験結果を表 1 に示す．

表 1: 文学作品による実験結果

	ランダム収集		無頼派	
N の値	属性数	精度 (%)	属性数	精度 (%)
1	5	73.33	3	93.33
2	6	76.67	3	86.67
3	5	80.00	4	80.00
4	5	70.00	3	73.33

4. おわりに

本研究では N-gram によって筆者推定を行うためのシステムを提案した．このシステムには事前に辞書を用意する必要性はなく，また辞書には登録されていない語の共起関係による決定木を作成することができた．これによって筆者推定のための新しい知識が発見できると期待できる．

今回のシステムでは 1gram から 4gram まで分割して分類精度を求めたが，複数の種類の gram を用いた場合についての調査も行う予定である．

参考文献

[稲積 00] 稲積 宏誠, 吉澤 有美 : 論理最小化に基づく決定木による知識発見, 人工知能学会誌, Vol. 15, No. 4, pp. 657–664 (2000)

[岡田 05] 岡田 望, 西園 敏弘, 古畑 裕介 : 文書分類における決定木アルゴリズム適用法の検討 (オフィスアプリケーション・ネットワーキング・マネジメント及び一般), 電子情報通信学会技術研究報告. OIS, オフィスインフォメーションシステム, Vol. 104, No. 568, pp. 97–102 (2005)

[金 07] 金 明哲, 村上 征勝 : テキスト分類問題を対象としたベクトル空間における距離構造の漸近解析, 統計数理, Vol. 55, No. 2, pp. 255–268 (2007)

[松浦 99] 松浦 司, 金田 康正 : n-gram 分布を用いた近代日本語小説文の著者推定, 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 99, No. 95, pp. 31–38 (1999)

[松浦 00] 松浦 司, 金田 康正 : 近代日本小説家 8 人による文章の n-gram 分布を用いた著者判別, *IPSJ SIG Notes*, Vol. 2000, No. 53, pp. 1–8 (2000)