自然言語処理 118－6
（1997．3．21）

確率的言語モデルを用いた言語データのクラスタリング

北 研二

徳島大学 工学部

本稿では、言語データから作成された確率的言語モデルの間に距離(類似度)を導入し、この距離に基づき、元の言語データをクラスタリングする手法を提案する。また、提案した手法の評価を2種類の実験により行った。まず最初に、ECI多言語コーパス中の19ヶ国語のテキスト・データから、言語の系統樹を再構築する実験を行った。得られたクラスタリング結果は、言語学で確立された言語系統樹と非常に似ており、提案した手法の有効性を示すことができた。次に、Gutenbergプロジェクトより入手した7人の作家の小説に対し、クラスタリングを試みた結果、作家ごとに分類されたクラスタを得ることができた。

Automatic Clustering of Linguistic Data Based on Probabilistic Language Models

Kenji Kita

Faculty of Engineering, Tokushima University

This paper proposes a novel method for automatically organizing linguistic data. The basic idea of this method involves developing a probabilistic model from the given linguistic data, and then computing the distances accoring to the distance measure defined on the language models. Clustering is performed based on this distance measure. The effectiveness of the proposed method has been confirmed by two kinds of experiments. First, we conducted an experiment to reconstruct the language family tree using multilingual texts of nineteen different languages from the ECI multilingual corpus. The results were very encouraging. They were very close to the family tree of languages established in linguistics. Second, we conducted a clustering experiment using 43 novels by seven differentt authors, and showed that the proposed methos is also applicable to authorship study.

1 はじめに

統計的手法に基づき、言語の比較を計量的に行なう研究は、従来から広く行なわれてきている。KroeberおよびChrétienは、1930年代に、音韻や語形等の言語的特徴から言語間の相関係数を求め、これに基づきインド・ヨーロッパ諸言語9ヶ国語およびヒッタイト語の間の類似性を求める研究を行っている[9, 10]。また、クラスター分析に基づき、自動的に言語や方言を分類する研究に関しても、いくつかの先行研究がある[3]。比較的最近の研究では、Batagelj らの研究があり、文字列間距離に基づいた言語間の類似性を用いて、65ヵ国語の言語に対するクラスタリング結果を示している[5]。しかし、従来の研究における言語間の距離(類似性)の定義は多分に恣意的である上、距離の算出において、あらかじめ人間が言語を分類する上で有用であると思われる音韻や語形等の言語的特徴を抽出したり、あるいは比較のための基礎語彙を選定するなどの作業が必要であった。

本稿では、確率的言語モデルに基づいた言語データのクラスタリング手法を提案し、これを言語系統樹の自動構築に応用する。本稿の方法では、まず各言語の言語データから確率的言語モデルを学習し、

次に確率モデル間に距離を導入することにより、言語間の距離を定義する。本稿の方法は、完全に自己組織的 (self-organizing) であり、あらかじめ人間が各言語の言語的特徴を抽出したり、基礎語彙を選定する必要はない。また、本稿の方法の利点として、各言語のデータを独立に選ぶことができるという点をあげることができる。たとえば、言語によって違うジャンルのテキストであったり、あるいはデータのサイズが異なっていても、これらのデータの揺れを確率モデルの中に吸収することができる。

また、本稿では、確率モデルに基づいたクラスタリング手法が、文献の計量分析や著者判定にも応用可能であることを示す。同種の方法は、テキストの自動分類 (Text Categorization) にも応用可能である。また、比較言語学、方言研究、言語類型論、社会言語学など、言語学の諸分野においても有用な手法を提供するものと思われる。

2 確率的言語モデル

本稿では、確率モデルとして、文字の trigram モデルを用いる。trigram モデルは、N-gram モデル [7, 2] の特別な場合 ($N=3$ の場合) であり、以下では N-gram モデルについて簡単に説明する。

2.1 N-gram モデル

たとえば、英語では文字 q には文字 u が後続するとか、ドイツ語においては文字 c に後続するのは h や k であるなど、文字の連鎖には確率・統計的な性質が存在する。N-gram モデルは、このような文字の連鎖をモデル化するために適した確率モデルである。

文字の N-gram モデルは、文字の生起を $N-1$ 重マルコフ過程により近似したモデルであり、文字の生起は直前に出現した $N-1$ 文字にのみ依存すると考える。すなわち、n 文字から成る文字列 $c_1,\cdots,c_n$ に対し、

$$P(c_n|c_1,\cdots,c_{n-1}) \approx P(c_n|c_{n-N+1},\cdots,c_{n-1}) \quad (1)$$

となる。

N-gram モデルを用いた場合、文字列 $c_1,\cdots,c_n$ の生成確率は、次のようにして計算することができる。

$$\begin{aligned} P(c_1,\cdots,c_n) &= \prod_{i=1}^{n} P(c_i|c_1,\cdots,c_{i-1}) \\ &\approx \prod_{i=1}^{n} P(c_i|c_{i-N+1},\cdots,c_{i-1}) \quad (2) \end{aligned}$$

上式において、最初の等式は、確率論の基本定理から導かれる。また、2 番目の近似式は、式 (1) による。

いま、文字列 $c_1,\cdots,c_n$ が言語データ中に出現する回数を $F(c_1\cdots c_n)$ で表すことにする。N-gram の確率は、言語データ中に出現する文字の N 個組と $(N-1)$ 個組の出現回数から、次のように推定することができる。

$$\begin{aligned} &P(c_n|c_{n-N+1},\cdots,c_{n-1}) \\ &= \frac{F(c_{n-N+1},\cdots,c_n)}{F(c_{n-N+1},\cdots,c_{n-1})} \end{aligned} \quad (3)$$

N の値が大きい場合には、統計的に信頼性のある確率値をコーパスから推定することが難しくなるため、通常は $N=3$ あるいは $N=2$ のモデルが用いられることが多い。なお、$N=3$ の場合を trigram モデル、$N=2$ の場合を bigram モデル、$N=1$ の場合を unigram モデルと呼ぶ。

2.2 N-gram モデルのスムージング

N-gram の確率値は、式 (3) に示すように、言語データ中の文字列の頻度から推定することができる。しかし、与えられた言語データが少ない場合には、精度のよい確率値を推定することが難しくなる。この問題に対処するために、我々の実験では、線形補間法と呼ばれる方法を用いて、N-gram モデルのスムージング (平滑化) を行った。

線形補間法では、N-gram の確率値を低次の M-gram ($M<N$) の確率値と線形に補間する。trigram の場合には、次のようになる。

$$\begin{aligned} P(c_n|c_{n-2}c_{n-1}) &= \lambda_1 P(c_n|c_{n-2}c_{n-1}) \\ &+\lambda_2 P(c_n|c_{n-1}) + \lambda_3 P(c_n) \end{aligned} \quad (4)$$

ここで、$\lambda_1,\lambda_2,\lambda_3$ は、それぞれ trigram, bigram, unigram に対する重み係数であり、$\sum_i \lambda_i = 1$ となるように設定される。式 (4) の補間では、学習データ中に三つ組 c_{n-2},c_{n-1},c_n が出現しない場合には、bigram と unigram から $P(c_n|c_{n-2},c_{n-1})$ の値を推定している。二つ組 c_{n-1},c_n も出現しない場合には、unigram の値によって近似している。なお、λ_i の値は、削除補間法 [6, 2] により推定した。

3 言語モデル間の距離

次に、言語モデル間に距離を導入する。我々の用いた距離は、文献[8, 11]において提案されているものと同一である。上記文献においては、隠れマルコフ・モデル (Hidden Markov Model; HMM) 間の距離として定義されているが、一般の言語モデルに対しても同様に用いることができる。

いま、2つの言語データ D_1, D_2 が与えられているとする。D_i $(i = 1, 2)$ は、文字列データであり、その長さ (文字数) を $|D_i|$ と表記する。また、言語データ D_i から作成された言語モデルを M_i で表す。

まず、言語モデル M_1 および M_2 に対し、距離尺度 $d_0(M_1, M_2)$ を次のように定義する。

$$d_0(M_1, M_2) = \frac{1}{|D_2|}\left[\log P(D_2|M_2) - \log P(D_2|M_1)\right] \quad (5)$$

式(5)では、言語モデル間の距離を、言語モデル M_1 からデータ D_2 が生成される確率と、言語モデル M_2 から同一のデータ D_2 が生成される確率の差に基づいて決めている。もし、元の言語データが類似していれば、モデルからのデータの生成確率も似た値になるので距離は小さくなるし、類似していなければ、データの生成確率が大きく違うので距離は大きくなる。

式(5)は、言語モデル M_1 および M_2 に対し、非対称である (すなわち $d_0(M_1, M_2) \neq d_0(M_2, M_1)$)。対称形にするために、$d_0(M_1, M_2)$ と $d_0(M_2, M_1)$ の平均を取る。従って、言語モデル M_1 と M_2 の間の距離 $d(M_1, M_2)$ は、最終的に次のように定義される。

$$d(M_1, M_2) = \frac{d_0(M_1, M_2) + d_0(M_2, M_1)}{2} \quad (6)$$

4 実験1：多言語コーパスからの言語系統樹の再構築

4.1 概要

言語の文字 trigram モデルと上記で定義した言語モデル間距離に基づき、階層的 (凝集型) クラスター分析を行なうことにより、言語のデンドログラム (dendrogram)・言語系統樹を作成することができる (図1参照)。我々は、ECI 多言語コーパス (European Corpus Initiative Multilingual Corpus) 中の言語データを用いて、言語の系統樹を再構築する実験を行った。

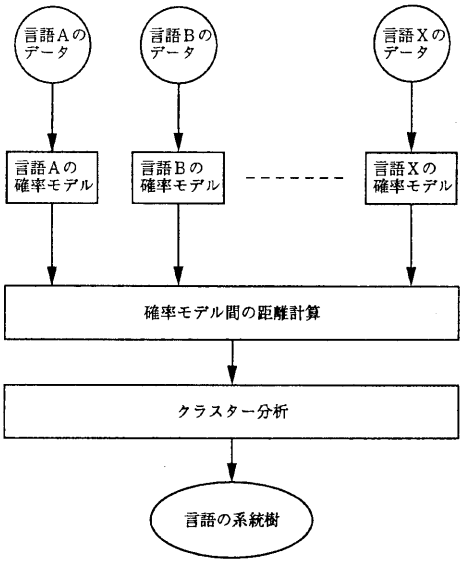

図 1: 確率モデルに基づく言語系統樹の再構築

4.2 言語データ

ECI コーパスは、ELSNET (European Network in Language and Speech) から CD-ROM により提供されているもので、総語数約1億語から成る。ECI コーパス中には、主要なヨーロッパ各国語およびトルコ語、日本語、ロシア語、中国語、マレー語等の言語データが含まれている。本実験では、このうち、ISO Latin-1 文字セットでコード化されている19言語のデータを用いた。

表1は、本実験で用いた言語の種類、各言語データの ECI コーパス中での識別子、言語データのジャンルを示している。表のジャンル欄において、「並行テキスト」と記されているのは、同一の内容を多言語で記述したものであることを示している。

ECI コーパス中のテキストは SGML によりコード化されているが、本評価実験では、まず SGML のタグを除去し、テキスト部分のみを抽出した。次に、多言語の言語データ間に均質性を持たせるために、単語表記中にアルファベット大文字が使われている

表 1: 実験で用いた言語の種類・言語データの識別子・テキストのジャンル

言語	ECI コーパス中の識別子	ジャンル
アルバニア語	alb01b	小説
チェコ語	cze01a01	新聞
ラテン語	lat01a01	詩
リトアニア語	lit01a	フィクション
マレー語	mal01a01	技術文書
ノルウェー語	nor01a01	フィクション
トルコ語	tur02a	新聞
クロアチア語	cro18a	小説
セルビア語	ser18a	(並行テキスト)
スロベニア語	slo18a	
デンマーク語	dan16a	技術文書
オランダ語	dut16a	(並行テキスト)
英語	eng16a	
フランス語	fre16a	
ドイツ語	ger16a	
イタリア語	ita16a	
ポルトガル語	por16a	
スペイン語	spa16a	
ウズベク語	mul13a	小説

場合は小文字に変換し、言語によってはウムラウトやアクセント記号等を表す特殊符号が入っていたが、英語式アルファベット 26 文字以外の特殊文字は、すべて対応するアルファベットに変換した。たとえば、ã は a に変換した。また、文字の trigram は、表 1 の識別子欄に示されているテキストの最初の 1,000 単語を用いた。

4.3 実験結果および考察

図 2に、19 言語のクラスタリング結果を示す。なお、クラスタリング・アルゴリズムには、群平均法 UPGMA(Unweighted Pair-Group Method using Average)[4] と呼ばれる方法を用いた。群平均法は、広い範囲においてよい結果を与えるクラスター分析法であるといわれている。

言語名の左側の樹状図が実験により得られた結果であり、右側には各言語のおおまかな分類を記している。以下では、言語学的な観点から、クラスタリング結果の妥当性について考察する。なお、言語の分類および諸言語間の関係に関しては、文献 [1] を参考にした。まず、評価実験で用いた言語は、以下のように大きく分類される。

(A) インド・ヨーロッパ語族

(A-1) アルバニア語派 (アルバニア語)

(A-2) スラブ語派 (チェコ語, クロアチア語, セルビア語, スロベニア語)

(A-3) バルト語派 (リトアニア語)

(A-4) イタリック語派 (ラテン語, フランス語, ポルトガル語, スペイン語, イタリア語)

(A-5) ゲルマン語派

(A-5-1) 北ゲルマン語派 (ノルウェー語, デンマーク語)

(A-5-2) 西ゲルマン語派 (オランダ語, ドイツ語, 英語)

(B) アルタイ諸語 (トルコ語, ウズベク語)

(C) オーストロネシア語族 (マレー語)

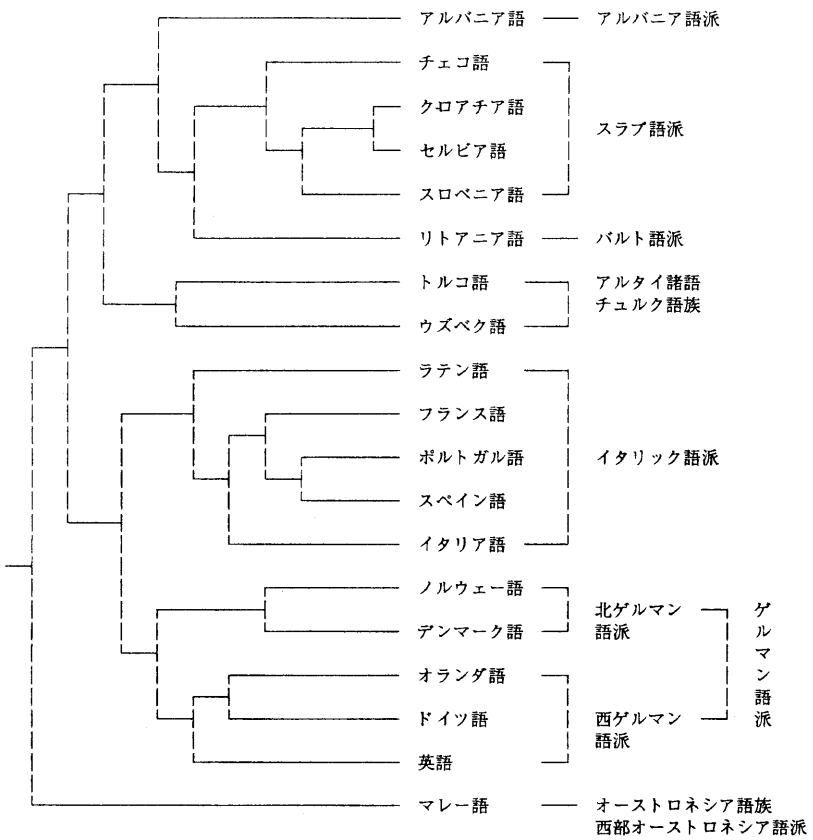

図 2: ECI 多言語コーパスより得られたクラスタリング結果

図 2の右側に示すように、実験により得られた結果は、上記の大分類を反映したものになっている。

次に、言語間のより細かな関係について調べる。まず、実験結果では、スラブ語派に属するクロアチア語とセルビア語を、最初に一つのクラスタとしてまとめている。クロアチア語とセルビア語はともに、南スラブ語群に属し、両者の差異は方言的なものであるとされている。従って、両者を一つのクラスタとすることは、きわめて妥当であるといえる。また、実験結果では、スラブ語派とバルト語派を併合した後に、これをアルバニア語派と併合している。スラブ語派とバルト語派の諸言語には、多くの類似点が見られ、バルト・スラブ祖語の存在を考えている研究者もいる。アルバニア語は、同一の語派に属する言語がなく、1 言語で 1 語派の扱いを受けているが、南スラブ語等の言語からの影響を受けている。実験結果は、以上の点を反映しているということができる。西ゲルマン語派に関しては、オランダ語とドイツ語を、まず併合しているが、ドイツ語学では、オランダ語をドイツ語の 1 方言、低地フランク語として扱っており、この 2 言語はきわめて類似している。以上のように、実験結果は、言語の細分類に関しても、かなりの部分で言語学での分類と一致しており、提案したクラスタリング手法が有効なものであることを示している。

4.4 言語識別の実験

追加実験として、確率モデルに基づく言語間距離を用いた、言語識別の実験を行った。この実験では、上記で得られた各言語の文字 trigram モデルを用いて、未知のテキストから、そのテキストの使用言語を特定することを試みた。

実験の手順は、以下のようである。各言語に対し、2つの未知テキストを評価データとして用意した。未知テキストは、上記の言語モデルの作成に用いたデータ(以下、学習データと呼ぶ)とは別のテキスト・データである。言語によっては、学習データ以外のテキスト・データがないものもあり、言語識別用の評価データとして13言語・26テキストを用いた。

つぎに、未知テキストから言語モデル(文字 trigram モデル)を作成し、未知テキストの言語モデルと各言語の言語モデルとの間の距離を計算して、最も小さな距離を与える言語を、未知テキストの使用言語と推定した。

図3に、未知テキストの最初の10単語、20単語、30単語、40単語、50単語、100単語、1000単語を用いた場合の言語識別率を示す。図から分かるように、未知テキストからの使用言語の特定には、20単語程度あれば十分であるということができる。20単語を用いたときには、26個の未知テキストのうち、25個についてその言語を正確に推定できた(識別率96.2%)。なお、識別に失敗したものは、例えばセルビア語のテキストをクロアチア語と間違うなど、近親関係の言語間での間違いが主であった。

5 実験2：7作家の作品のクラスタリング

次に、Gutenbergプロジェクトより入手した7人の作家の小説・43作品から文字 trigram モデルを作成し、クラスタリングを行った。

実験結果を、図4に示す(右端が作品名および作家名)。なお、この実験では、Neighbor-Joining 法[12]と呼ばれるクラスタリング・アルゴリズムを用いた。群平均法によってもクラスタリングを試みたが、Neighbor-Joining 法を使った場合の方がきれいな結果を得た(群平均法では鎖状のクラスターを生成した)。図4より分かるように、概ね作家ごとに分類されたクラスタを得ており、確率モデルに基づくクラスタリングは、文献の計量分析や著者判定(真贋分析)にも応用可能であることを示している。

6 おわりに

本稿では、確率的言語モデルに基づいた言語データのクラスタリング手法を提案した。提案した手法を用いて、ECI多言語コーパス中の19ヶ国語のテキスト・データから言語の系統樹を再構築する実験を行ない、実験結果を言語学での分類と比較することにより、提案した手法の有効性を示した。また、Gutenbergプロジェクトより入手した7人の作家の小説に対し、クラスタリング実験を行い、提案した手法が文献の計量分析にも応用可能であることを示した。今後は、テキストの自動分類やクラスタリングに関する応用について検討してみたい。

参考文献

[1] 亀井 孝・河野 六郎・千野 栄一(編著):「言語学大辞典(全6巻)」, 三省堂 (1988).

[2] 北 研二・中村 哲・永田 昌明:「音声言語処理 – コーパスに基づくアプローチ –」, 森北出版 (1996).

[3] 安本 美典:「言語の科学 – 日本語の起源をたずねる」, 朝倉書店 (1995).

[4] 鷲尾 泰俊・大橋 靖雄:「多次元データの解析」, 岩波書店 (1989).

[5] Batagelj, V., Pisanski, T., & Keržič, D.: "Automatic clustering of languages", *Computational Linguistics*, 18(3), pp. 339–352 (1992).

[6] Jelinek, F., & Mercer, R.: "Interpolated estimation of Markov source parameters from sparse data", In: Gelsema, E. S., & Kanal, L. N. (eds.), *Pattern Recognition in Practice*, pp. 381–397 (1980).

[7] Jelinek, F.: "Self-organized language modeling for speech recognition", In: Waibel, A., & Lee, K-F. (eds.), *Readings in Speech Recognition*, Morgan Kaufmann Publishers, pp. 450–506 (1990).

[8] Juang, B. H., & Rabiner, L. R.: "A probabilistic distance measure for hidden Markov models", *AT&T Technical Journal*, 64(2), pp. 391–408 (1985).

[9] Kroeber, A. L., & Chrétien, C. D.: "Quantitative classification of Indo-European languages", *Language*, 13(2), pp. 83–103 (1937).

[10] Kroeber, A. L., & Chrétien, C. D.: "The statistical technique and Hittite", *Language*, 15(2), pp. 69–71 (1939).

[11] Rabiner, L. and Juang, B. H.: *Fundamentals of Speech Recognition*, Prentice Hall (1993).

[12] Saitou, N. and Nei, M.: "The neighbor-joining method: a new method for reconstructing phylogenetic trees", *Molecular Biology and Evolution*, Vol. 4, No. 4, pp. 406-425 (1987).

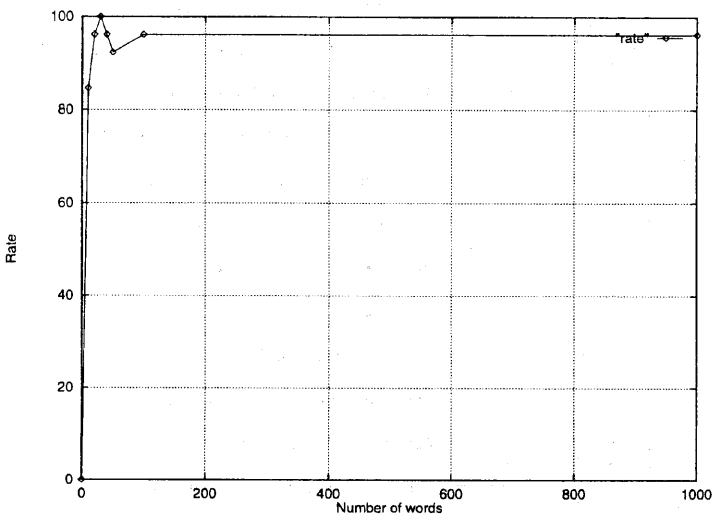

図 3: 未知テキスト中の単語数と言語識別率

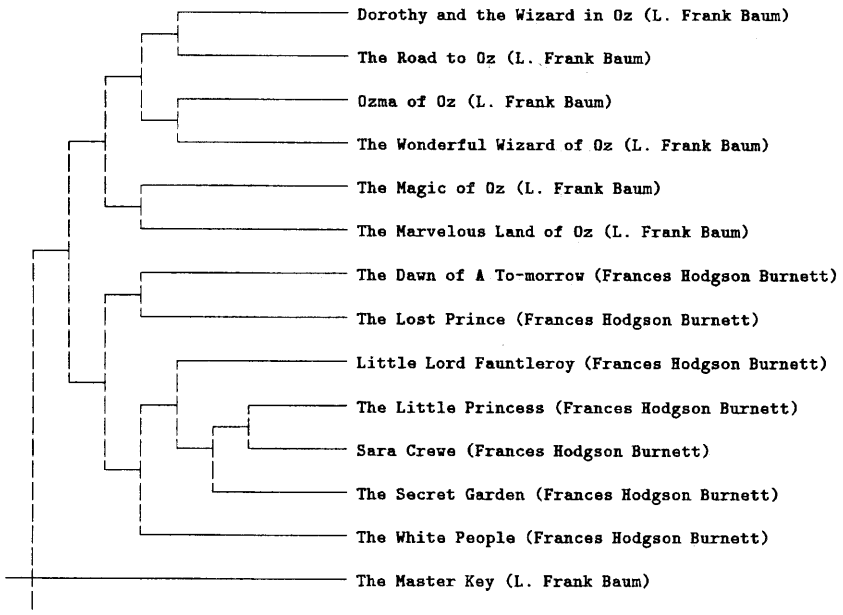

図 4: 7人の作家，43小説より得られたクラスタリング結果

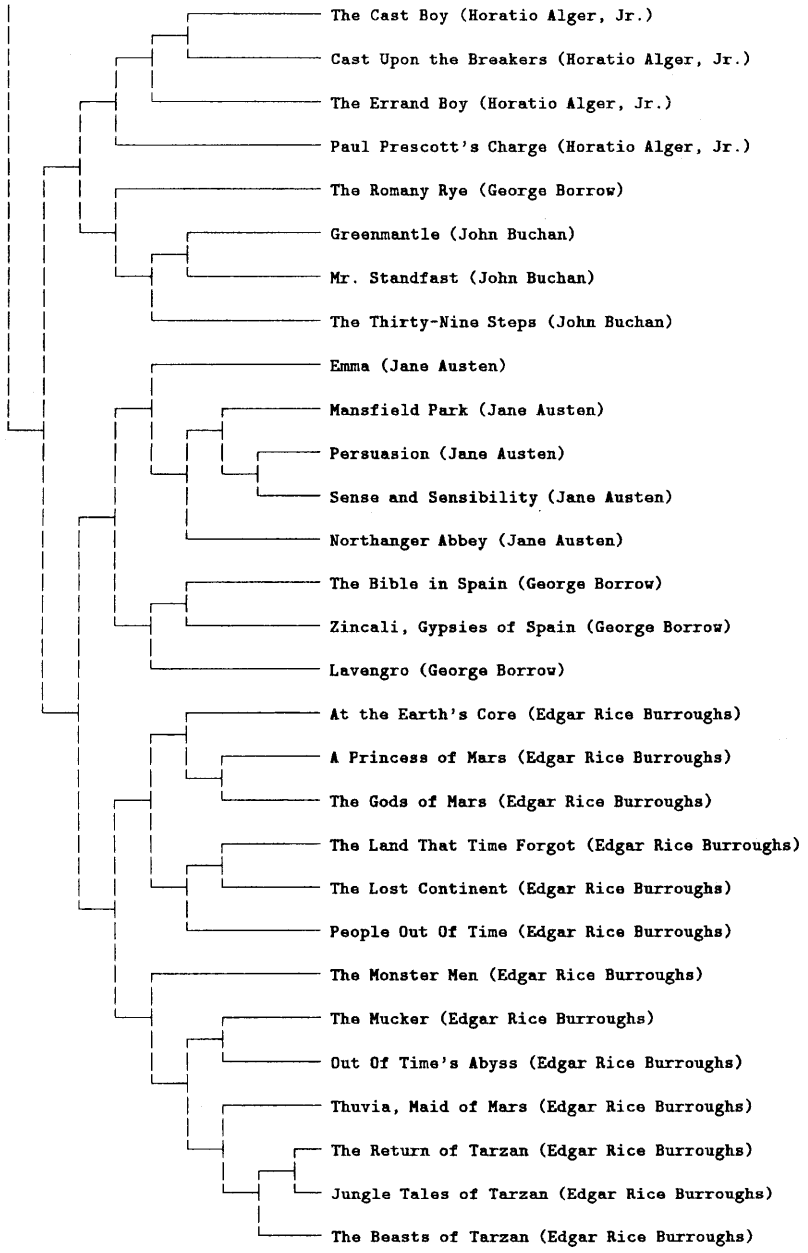

図4 (続き): 7人の作家, 43小説より得られたクラスタリング結果