確率モデルに基づく言語のクラスタリング：多言語コーパスからの言語系統樹の再構築

北 研二
徳島大学 工学部

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 [7, 8]。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き、与えられた言語データを自動的に分類／クラスタリングする手法を提案する。ここでは、言語を文字列を生成する情報源であるとみなし、この情報源の確率・統計的な性質を確率モデルによりモデル化する。次に、確率モデル間に距離尺度を導入し、この距離尺度に基づき言語 (データ) のクラスタリングを行なう。また、提案した方法を用いて、ECI 多言語コーパス (European Corpus Initiative Multilingual Corpus) 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ことを試みる。

2 確率モデルに基づく言語のクラスタリング

本稿で提案する方法の概略を図 1 に示す。この方法では、まず各言語の言語データから確率的言語モデルを自動的に学習し、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。このように、本稿の方法は、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。

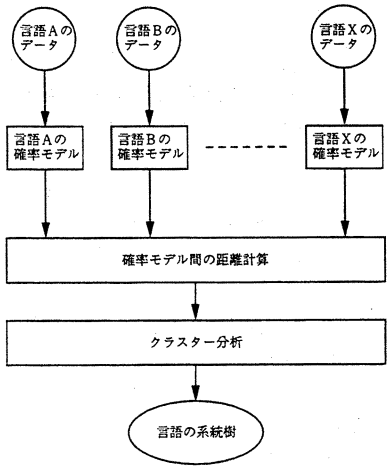

図 1: 確率モデルに基づく言語のクラスタリング

2.1 N-gram 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 [2] の特別な場合 ($N = 3$ の場合) である。

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。

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

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 $N=3$ (trigram) あるいは $N=2$ (bigram) のモデルが用いられることが多い。

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 [2] 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。

2.2 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 [6] 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、言語 L_1 および言語 L_2 の言語データとして、それぞれ 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。

$$\begin{aligned} &d_0(M_1,M_2) \\ &= \frac{1}{|D_2|}\left[\log P(D_2|M_2) - \log P(D_2|M_1)\right] \end{aligned} \quad (4)$$

式 (4) では、言語 L_1 と L_2 の間の距離を、言語 L_1 のモデル M_1 からデータ D_2 が生成される確率と、言語 L_2 のモデル 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、言語 L_1 と L_2 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 (4) 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。

$$\begin{aligned} &d(M_1,M_2) \\ &= \frac{d_0(M_1,M_2)+d_0(M_2,M_1)}{2} \end{aligned} \quad (5)$$

3 評価実験

3.1 言語データ

以上で提案した方法の有効性を実証するために、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。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: 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 (並行テキスト)
セルビア語	ser18a	
スロベニア語	slo18a	
デンマーク語	dan16a	技術文書 (並行テキスト)
オランダ語	dut16a	
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

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 の作成には、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。

3.2 実験結果および考察

上記により作成した文字 trigram モデルに対し、階層的 (凝集型) クラスター分析を行ない、言語のデンドログラム (dendrogram; 樹状図) を作成した。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。

図 2に、19 言語のクラスタリング結果を示す。言語名の左側の樹状図が実験により得られた結果である。図 2の右側に示すように、語族・語派に対応したクラスタを得ている。

次に、文献 [1] を参考に、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

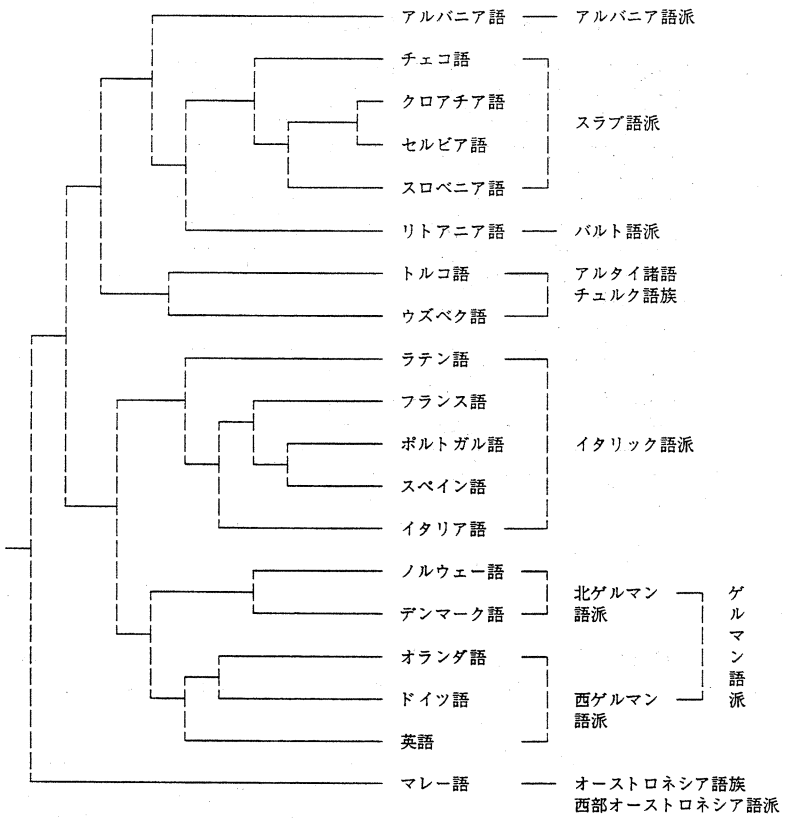

図 2: ECI 多言語コーパスより得られたクラスタリング結果

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 おわりに

本稿では、確率的言語モデルに基づいた言語のクラスタリング手法を提案した。また、提案した手法を、ECI 多言語コーパス中の 19 ヶ国語のテキスト・データを用いた実験により評価し有効性を示した。

本稿では、言語のクラスタリングを中心に扱ったが、提案した手法はテキストの分類 (Text Categorization)、文献の著者判定 (真贋分析) などにも応用可能であると考えられる。また、本稿で述べた基本的な手法は、比較言語学、方言研究、言語類型論、社会言語学などの幅広い分野で役立つものと期待される。

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) (1992).

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

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

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