基于概率模型的语言的聚类方法：根据多语种语料库进行语言系统树的再构造

北研二　　穗志方（译）　俞士汶（校）
（日本德岛大学）　（北京大学计算语言学研究所）

摘 要：本文提出一种对语言进行自动聚类的新方法。它的基本思想为从给定的语言数据中为每种语言设计一个概率模型，通过计算语言模型之间的距离来计算语言之间的距离，并在此基础上对语言进行聚类。最后，用这种方法对 ECI 多国语言语料库中的 19 种语言的文本数据进行了实验并评价了它的有效性。

关键词：聚类　概率模型　多语种语料库

A Probabilistic-model-based Language Clustering Approach:To Reconstruct Language System Tree from the Multilingual Corpus

Kenji Kita　　Sui Zhifang （translator） Yu Shiwen （proof reader）
（Tokushima University）　（Computational Linguistics Institute of Peking University）

Abstract: This paper proposes a novel method for automatically clustering languages. The basic idea of this method involves developing a probabilistic model for each language from the given linguistic data, and then computing the distances between languages according 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 an evaluation experiment using multilingual texts of nineteen languages from the ECI/MC1 corpus .

Key words: clustering, probabilistic model, multilingual corpora

一、引言

采用基于统计的方法，对语言的比较进行计量的研究，历来被广泛使用。 Kroeber 和 Chretien 在 30 年代根据音韵、词形等语言的特征求语言之间的相关系数。用这种办法对印度、欧洲九国语言以及赫梯语言之间的类似性进行了研究【7, 8】。另外,基于聚类分析的有关语言或方言的自动分类研究也有若干个早期的例子。比较近期的研究是 Batagelj 等的研究，利用基于文字序列之间距离的语言间的相似性，标识出 65 种语言的聚类结果【5】。然而，在历来的研究中，语言之间的距离（相似性）的定义通常是不严格的，计算距离时，或是预先将

人类在进行语言分类时认为有用的语言特征如音韵或词形等抽出，或是选定用于比较的基础词汇。

本文提出基于概率的语言模型，对于给定的语言数据进行自动分类或者是聚类。这里把生成文字序列的语言看成信息源，对此信息源的概率、统计等性质用概率模型进行模型化。接着，引进概率模型间的距离尺度，基于这个距离尺度来进行语言（数据）的自动分类，并利用提出的方法，对 ECI 多国语言语料库（ European Corpus Initiative Multilingual Corpus ）中的 19 个国家的语言的文本数据进行语言的系统树的再构造实验。

二、基于概率模型的语言的聚类

本文所提方法的框架如图 1 所示。这种方法首先从各种语言的语言数据中自动学习概率的语言模型，其次通过引入概率模型间的距离来定义语言之间的距离。所以，本文的方法是自组织的（ self--organizing ），不需要人工预先抽 a 语言的语言特征，选定基础词汇。本方法的优点是可以独立选择各种语言的数据。例如，不同语言的文本类型，有差异也好，数据的规模不一致也好，这些数据的不确定性都可以在概率模型中被吸收掉。

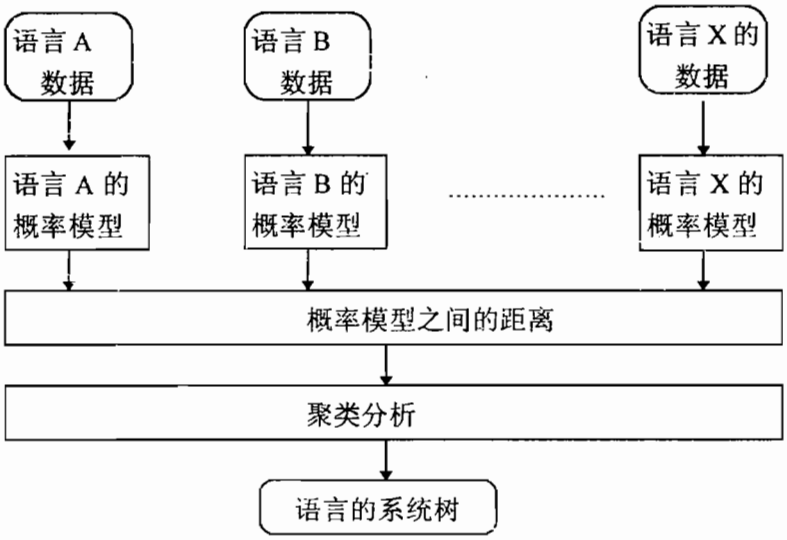

图 1:基于概率模型的语言的聚类

2.1 N 元语法模型

本文中的概率模型采用字符的三元语法模型，即当 N ＝ 3 时的 N 元语法模型。

例如，英语中，字符 q 后续为字符 u ，德语中，字符 c 后续可能是 h 或 k ，在字符序列中存在这样的概率统计性质。 N 元模型就是适合将字符序列模型化的一种概率模型。

字符的 N 元模型是用 N － 1 重马尔可夫过程来表示某个字符的发生的一个近似模型，即认为第 N 个字符的发生仅与直接在它之前出现的 N -- 1 个字符有关。即对于由 n 个字符构成

的字符序列 C1,・・・,Cn 有：$P(c_n|c_1,\ldots,c_{n-1}) \approx P(c_n|c_{n-N+1},\ldots,c_{n-1})$ （1）

在使用 N 元模型的条件下，字符序列 C1,・・・,Cn 的生成概率可由以下的公式计算出：

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

现在，字符序列 C1,・・・,Cn 在语言数据中出现的次数用 F （C1,・・・,Cn）表示。N 元语法的概率可以根据语言数据中字符的 n 元组和 n-1 元组的出现次数，由以下的公式推定。

$$P(c_n|c_{n-N+1},\ldots,c_{n-1})=\frac{F(c_{n-N+1},\ldots,c_n)}{F(c_{n-N+1},\ldots,c_{n-1})} \quad （3）$$

因为 N 的值越大，从语料库中推定可靠的概率值的难度就越大，所以通常多使用 N ＝ 3（trigram）或 N ＝ 2（bigram）的模型。

N 元语法的概率值可按公式（3）中所示，根据语言数据中字符序列的频度推定而来。然而，当给定的语言数据比较少时，就很难推定精确的概率值。为处理这个问题，我们在实验中采用线性插值法来进行 N 元语法模型的平滑。

2.2 语言模型之间的距离

以下，引入语言模型之间的距离。我们采用的距离，与文献【6】所提出的定义一致。上述文献中，定义了隐马尔可夫模型（Hidden Markov Model:HMM）之间的距离。这个定义对于一般的语言模型也同样适用。

现在，分别给定 D1，D2 作为语言 L1 和语言 L2 的语言数据。Di（i ＝ 1，2）为字符序列，其长度（字符数）用 |Di| 表示。由语言数据 Di 生成的语言模型用 Mi 表示。

首先，对于 M1 和 M2，其距离 $d_0(M_1,M_2)$ 定义如下：

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

公式（4）中，语言 L1 和 L2 之间的距离由从语言 L1 的模型 M1 生成数据 D2 的概率和从语言 L2 的模型 M2 生成同样的数据 D2 的概率的差决定。如果语言 L1 和 L2 相似，从模型生成数据的概率值也相似，则距离也小；如果不相似，数据的生成概率相差很大，则距离大。

公式（4）中，语言模型 M1 和 M2 是非对称的（即 $d_0(M_1,M_2)\neq d_0(M_2,M_1)$）。为达到对称，应取 $d_0(M_1,M_2)$ 和 $d_0(M_2,M_1)$ 的平均值。进而，语言模型 M1 和 M2 之间的距离 $d(M_1,M_2)$ 最终定义如下：

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

三、评价实验

3.1 语言数据

为验证以上提出的方法的有效性，使用ECI多国语言语料库（European Corpus Initiative Multilingual Corpus）中的语言数据，进行语言的系统树的再构造实验。ECI语料库由ELSNET（European Network in Language and Speech）用CD--ROM提供，总词数约1亿词。ECI语料库中，含有主要的欧洲各国语言及土耳其语、日语、俄语、汉语、马来语等语言数据。在本实验中使用其中19种经ISO Latin-1字符集编码的语言数据。

表1：实验中采用的语言的种类，语言数据的标识符，文本的种类。

语言	ECI语料库中的标识符	种类
阿尔巴尼亚语	alb01b	小说
捷克语	cze01a01	新闻
拉丁语	lat01a01	诗
立陶宛语	lit01a	小说
马来西亚语	mal01a01	技术文书
挪威语	nor01a01	小说
土耳其语	tur02a	新闻
克罗地亚语	cro18a	小说(并行文本)
塞尔维亚语	ser18a	
斯洛文尼亚语	slo18a	
丹麦语	dan16a	技术文书(并行文本)
荷兰语	dut16a	
英语	eng16a	
法语	fre16a	
德语	ger16a	
意大利语	ita16a	
葡萄牙语	por16a	
西班牙语	spa16a	
乌兹别克语	mul13a	小说

表1表示了本实验采用的语言的种类，语言数据的标识符以及文本的种类。表的种类栏中，“并行文本”表示同一内容多种语言的文本。

ECI语料库中的文本是按照SGML格式，进行编码化的。本实验首先除去SGML的标记，仅抽出文本部分。其次，为保持多种语言的语言数据之间的均匀性，当单词中使用大写字母时，变换为小写字母，因为不同的语言，有时加进了表示元音变音或重音等的特殊字符，英语26个字母以外的特殊字符全部变换为相应的英语字母。同时，利用表1标识符栏中所示文本的起始1000个单词作成字符的三元语法模型。

3.2 实验结果与考察

对于用以上标记作成的字符三元组，进行层次的聚类分析，而作成语言的树状图（dendrogram）。自动聚类法采用群平均法(UPGMA: Unweighted Pair-GroupMethod using

Average)【4】。群平均法是在大范围内给出较好结果的聚类分析法。

图2显示了19种语言的自动聚类结果。语言名左侧的树状图是从实验中得出的结果。图2右侧显示的是对应于语族、语系所得到的类。

其次，参考文献【1】考察语言间的细微关系。首先，根据实验结果，首先将属于斯拉夫语系的克罗地亚语和塞尔维亚语汇总为一个类。克罗地亚语和塞尔维亚语同属于南斯拉夫语族，二者的差异只是方言的差异。因而，将二者合为一类是相当合适的。另外，根据实验结果，斯拉夫语系和波罗地语系合并后，又与阿尔巴尼亚语族合并。因为斯拉夫语派和波罗地语系中的各种语言有很多类似点，也有研究者认为存在波罗地·斯拉夫祖语。阿尔巴尼亚语是属于单一语族的语言，即将一个语言看成为一个语族，但它受南斯拉夫语等语言的影响。实验结果可以反应以上所说的问题。关于西日尔曼语派，首先合并荷兰语和德语，在德语中将荷兰语看作是它的一个方言,作为低地法语处理，这两种语言极其相似。以上的实验结果中，关于语言的细分类有相当一部分与语言学的分类一致，表明了本文所提出的自动聚类方法是有效的。

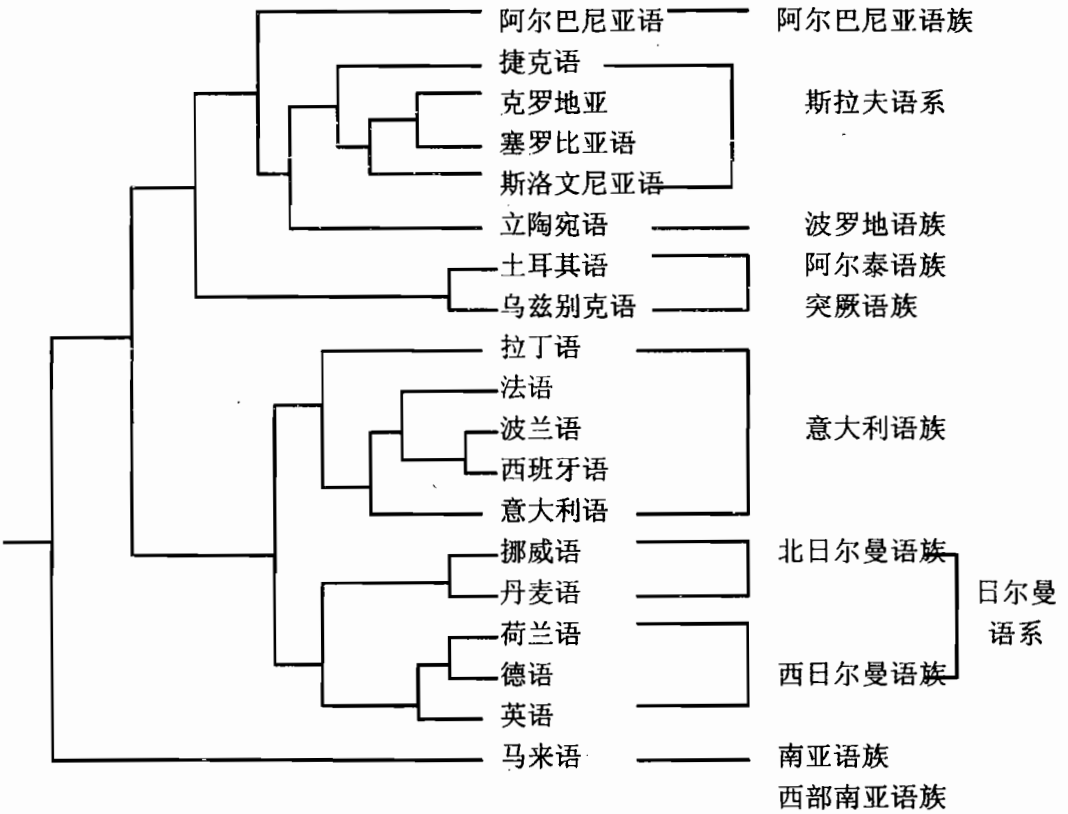

图2：从ECI多语种语料库中得到的聚类结果

四、结束语

本文提出了基于概率模型的语言的聚类方法，同时，用这种方法对 ECI 多国语言语料库中的 19 种语言的文本数据进行了实验并评价了它的有效性。

本文虽然以语言的聚类为中心，但提出的方法也可应用于文本的分类（ Text Categorization ），文献的作者判定（真伪辨别）等。同时，也可期待本文所论述的基本方法将有效地应用于比较语言学、方言研究、语言类型论、社会语言学等广泛的领域。

参考文献

【1】亀井孝，河野六郎，千野荣一（编著）：《语言学大辞典（全六卷）》，三省堂， 1988

【2】北研二，中村哲，永田昌明：《声音语言处理——基于语料库的方法》，森北出版社， 1996

【3】安本美典：《语言的科学——探求日本语的起源》，朝仓书店， 1995

【4】鹭尾泰俊，大桥靖雄：《多元数据的解析》，岩波书店, 1989

【5】Batagelj.V.., Pisanski.T..&Kerzic.D..：“Automaticclustering 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..&Chretien.C.D..：“Quantitative classification of Indo—European languages”, Language, 13 (2), 1937

【8】Kroeber.A.L..&Chretien.C.D..：“The statistical technique and Hittite”, Language. 15 (2), 1939