基于字符层马尔科夫模型的多语种识别*)

冯 冲[1,2] 黄河燕[2] 陈肇雄[2] 张 亮[2,3]
(中国科大计算机系 合肥 230027)[1] (中科院计算机语言信息工程中心 北京 100083)[2]
(南京理工大学计算机系 南京 210094)[3]

摘 要 语种识别是机器翻译等多语种语言处理任务的必要预处理过程。但双字节编码语种的识别,如中文、日文等,尚未被充分研究和试验。本文采用 Markov 语言模型,提出并测试了一种有效的基于 EM 的训练算法。同时,给出了性能分析和与其他算法的比较。

关键词 字符层马尔科夫模型,语种识别,机器翻译

Multiple Language Identification Based on Character-level Markov Models

FENG Chong[1,2] HUANG He-Yan[2] CHEN Zhao-Xiong[2] ZHANG Liang[2,3]
(Dept. of CS, USTC, Hefei 230027)[1] (CCLIE, CAS, Beijing 100083)[2] (Dept. of CS, NJUST, Nanjing 430074)[3]

Abstract Language identification is a necessary pre-process in machine translation and other multi-language applications, but no experiments hase yet been reported on double-byte encoded languages, such as Chinese and Japanese. An efficient EM based training algorithm on Markov language model is proposed and evaluated. The performance analysis and comparison with other algorithms are also presented.

Keywords Character-based markov models, Language identification, Machine translation

1 引言

在多语种信息抽取、机器翻译、跨语言信息检索中,识别不同的语言(以及同一语言的不同编码规范)是必要的预处理环节之一。实际上,这个问题在多语网络文本信息处理领域普遍存在。对于网页,HTML 标记中的 codeset 属性值有助于确定语种及其编码,但并不可靠;同时,网络中大量内容信息来源于纯文本文件或文本数据库。因此,多语种识别问题具有重要研究价值。以往的研究主要有规则方法和统计方法两种不同的思路。

基于规则的方法认为,给定文本中出现的特定字符或字符串可用作判断语种的依据,编码规范的某些特殊性质也有助于识别语种。常见的具体做法之一是利用特定语种中的常见字符串。例如,英语中常出现"ery(空格)",法语中常出现"eux(空格)",德语中常出现"(空格)der"。这样做的问题在于,第一,需要人工总结出知识并将其转换为系统规则,依赖于设计人员对语言本身的充分了解和分析;第二,结果的正确性很难把握,例如,虽然某些字符串在英语中大量出现,但法语、德语等语言中并不是绝对没有,也并不是所有的英语文档中都会出现这些字符串。

基于统计的方法把句子看作由统计语言模型生成而来,语种识别的问题也就是选择对于给定输入的最大似然语言模型的问题。其最大的优点在于,识别器的构造过程不需要人工对语言和编码规范进行分析,只要提供一些语料(无需标注)就能构造语言模型。但是,以往研究中,文[1]对于中文等以字为基本单位、词间无空格分隔的语言并不适用,同时算法只基于 n-gram 的频度计数,语言模型过于简化。文[2]在处理语言模型的数据稀疏问题时采用 Laplace 公式,难以分配给大多数没有出现在训练集合中的 n-gram 以适当比例的概率空间。实际上,文[2]仅在英语和西班牙语上进行了实验,文[1]中的实验也是在英、法、德等欧洲语言上进行。在我们对国内外相关文献的调研中没有发现在汉、日等双字节编码语种上进行的语种识别实验。

本文研究基于统计方法的语种识别。在字符层马尔科夫语言模型的基础上,利用 EM 算法进行参数估计,并在英、法、德、俄、简体中文、繁体中文、日、韩等 8 种语言的测试集上进行了实验。本文工作不借助语言学知识,无需分词处理,而且与先前研究文[2]相比,能够建立混杂度(perplexity)更小、更为准确的语言模型。

后继部分结构如下:首先介绍字符层马尔科夫语言模型,然后讨论其解码问题,接下来,详细描述了如何利用 EM 算法进行语言识别器的参数估计。随后给出了实验结果和数据分析。最后对相关研究进行了对比。结论部分总结了本文工作并展望了进一步的工作方向。

2 字符层马尔科夫模型

基本思路是构造字符层的语言模型,以此为基础,估计给定的输入字符串由各个语言模型生成而来的概率。建立语言模型时唯一的假设是训练和测试数据都是字节序列,这一假设能够避免分词等针对特定语言的预处理。

2.1 马尔科夫模型

马尔科夫模型是这样的一个随机过程:其后一状态的概率分布仅依赖于前面的 k 个状态。更形式化的,马尔科夫模型定义了一个从字母表 X 取值生成的字符串随机变量。生

*)受国家自然科学基金(编号 60272088)资助。冯 冲 博士研究生,主要研究方向为统计方法的多语种信息抽取和机器翻译。黄河燕 研究员,主要研究方向为机器翻译。陈肇雄 研究员,主要研究方向为机器翻译。张 亮 博士研究生,主要研究方向为自动问答系统。

成字符串 S 的概率为

$$p(S)=p(c_1\cdots c_n)=p(c_i)\prod_{i=2}^{n}p(c_i|c_{i-1}) \quad (1)$$

马尔科夫模型的优点之一在于它在数学上易于处理。从式(1)可以看出，该模型完全由初始概率 $p(c_1)$ 和状态转移概率 $p(c_i|c_{i-1})$ 决定。

假设某语言文本中字符串的分布服从 k 阶马尔科夫模型，其字母表 X 为该语言中的所有字母集合(如 ISO-latin-1 字符集)，我们可以构造一个基本的自然语言模型。其中的 k 是一个相对较小的值。尽管高阶统计语言模型能够更详细地描述语言的内在结构，但估计模型参数所需的训练样本的数据量也会随着模型的阶数呈指数增长。

对语种识别问题而言，高阶的模型过于详尽地反映了训练样本的特点，很多片段实际上是对样本内容的逐字复制。因此，片面地提高模型阶数并不有助于解决问题。实际上，n-gram 模型阶数越高，数据稀疏问题就越严重，可靠性就越差；阶数越低，对细节的描述能力就越差，模型反映出的相关性就越少。在模型阶数选择的问题上实际存在着一个平衡，理想的识别模型应当既能够描述训练样本所独有的语种和编码特点，又不过分反映样本中特定的文本内容和语言现象。在实验基础上，本文选用 3-gram。

2.2 解码和语种选择

一般地，对于语言模型 L_x，由贝叶斯定理，观测到文本 S 的概率为 $p(L_x, S) = p(L_x|S)\ p(S) = p(S|L_x)\ p(L_x)$，其中 $p(S) = 1$。假设不同的语言模型具有相等的先验概率 $p(L_x)$，则解码问题就是选择最为可能的 $p(S|L_x)$。处理这类在多种不同现象间进行选择的问题时，根据贝叶斯决策规则，应当选择最大似然原因。

对于语种识别问题，从某一特定的 k 阶马尔科夫模型生成文本 S 的概率为

$$p(S|L_x)=p(c_1\cdots c_k|L_x)\prod_{i=k+1}^{N}p(c_i|c_{i-k}\cdots c_{i-1}|L_x)$$

可以假设在所有给定的字符串前都填充同一特殊的字符串，这样，式中的首项可以视作 1 而忽略。合并后面的乘积项并进一步改写，得到

$$p(S|L_x)=\prod_{c_1\cdots c_{k+1}\in S}p(c_{k+1}|c_1\cdots c_k|L_x)^{F(c_1\cdots c_{k+1},S)}$$

其中，$c_1\cdots c_{k+1}$ 表示一个 $(k+1)$-gram，$F(c_1\cdots c_{k+1}, S)$ 表示字符串 S 中 $c_1\cdots c_{k+1}$ 出现的次数。式中的乘积操作在字符串 S 中所有能够找出的 $(k+1)$-gram 上进行。具体计算时为了避免浮点溢出，采用对数后的值进行操作，即

$$\log p(S|L_x)=\sum_{c_1\cdots c_{k+1}\in S}F(c_1\cdots c_{k+1},S)\log p(c_{k+1}|c_1\cdots c_k|L_x) \quad (2)$$

选择语种的过程在式(2)基础上完成。先在每一可能的语言模型上计算其最大似然概率。然后，由于假设它们具有相同的先验概率，可以简单地选出可能性最大的语种。潜在的一个问题是，如果概率值最大的两个或多个模型的值非常接近，则最大似然结果并不可靠。我们引入函数 $Conf$ 解决这一问题。它描述结果的置信度，被定义为

$$Conf=\frac{\log p_{2nd}-\log p_{1st}}{\log p_{2nd}}$$

其中，p_{1st} 和 p_{2nd} 分别表示最大概率和次大概率。这样，当 $Conf$ 小于设定阈值 C_0 时，输出 $Fail$ 标志而不是给出一个极可能错误的结果，能够提高系统的健壮性，有利于后继环节的处理。

2.3 基于 EM 算法的参数估计

建立语言模型的过程也就是在给定样本上进行参数估计的过程。设语种 L_x 的训练样本由字符串 S_T 构成，则有如下最大似然估计

$$p(c_{k+1}|c_1\cdots c_k|L_x)=\frac{F(c_1\cdots c_{k+1},S_T)}{F(c_1\cdots c_k,S_T)}$$

最大似然估计简单易行并能处理很多实际任务，但在语种识别中直接使用这一参数估计方法并不能得到满意结果。这是因为训练数据中没有出现过的 $(k+1)$-gram 会被赋以 0 概率。而且如果一个特殊的字符串在各个语言中都很少出现，而偶然的出现在 L_x 的训练样本中，则输入样本很可能会被误判为 L_x，因为在其他语言上的概率为 0。

简单地扩充训练数据规模并不能解决问题。一方面，为了保证运行效率，有必要限制训练样本的大小以控制语言模型的规模；另一方面，大规模语料仍然无法避免新的 $(k+1)$-gram 的出现。因此，数据平滑手段的引入是必要的。

一种常见方法是对 $p(L_x|S)$ 做最小均方差估计。文[2]采用的贝叶斯估计是处理均匀分布的随机变量时常用的最小均方差估计。这种方法最终得到的解形式为 Laplace 公式

$$\hat{p}=\frac{F(c_1\cdots c_{k+1},S_T)+1}{F(c_1\cdots c_k,S_T)+m}$$

其中，m 为 x 上的字母表的大小。这种方法非常简洁，但根据我们的实验，对于中文、日文等双字节编码的语种，从 Laplace 公式得到的语言模型存在较大不足。原因主要在于，多数 k-gram 没有出现在训练集合中，难以分配给这些 k-gram 适当比例的概率空间。

为此，本文引入了线性插值方法。一般来说，当没有足够的数据来准确地估计高阶 k-gram 模型的参数时，由于低阶模型受数据稀疏问题的影响要小一些，可以使用更低阶的 $(k-1)$-gram 建立插值模型(interpolated models)[3]：

$$p'_\lambda(w_i|w_{i-2}w_{i-1})=\lambda_3 p_3(w_i|w_{i-2}w_{i-1})+\lambda_2 p_2(w_i|w_{i-1})+\lambda_1 p_1(w_i)+\lambda_0/|V| \quad (3)$$

其中，$\lambda_i>0$，$\sum_{i=0}^{3}\lambda_i=1$，即有 $\lambda_0=1-\sum_{i=1}^{3}\lambda_i$。插值模型中的关键问题是估计不同阶数的语言模型的插值权重 $\lambda_0=(\lambda_0, \lambda_1, \lambda_2, \lambda_3)$。

我们使用 EM 算法估计权重参数。EM 算法[4]用于从非完整数据集中对参数进行 MLE 估计，是参数估计和无督导机器学习中的重要方法。它可被直观地看作一个逐次逼近算法：事先并不知道模型的参数，可以随机选择一套参数或粗略地给定初始参数 λ_0，确定出对应于这组参数的最可能状态，计算每个训练样本的可能结果的概率，在当前的状态下再由样本对参数修正，重新估计参数 λ，并在新的参数下重新确定模型状态，这样，通过多次迭代，直至满足某个收敛条件，就可以使模型参数逐渐逼近真实值。

把语料分为训练数据 T 和支持数据 H(held-out data)两部分。先在数据 T 上估计 p_3、p_2、p_1、$|V|$；然后在数据 H 上估计 λ_i。λ_i 的估计的目标函数是使得数据的随机性最大化，也就是使得交叉熵最小化，即

$$\lambda_i=\arg\min_{\lambda_i}\left(-\frac{1}{|H|}\sum_{j=1\cdots|H|}\log_2 p'_\lambda(w_j|h_j)\right) \quad (4)$$

这是一个反复进行的迭代过程。具体算法如下：

1)为 λ_i 赋初始值，满足 $\lambda_i>0$，$\sum_{i=0}^{3}\lambda_i=1$，$i=0,1,2,3$；

2)按照式(1)，计算 $p'_\lambda(w_j|h_j)$；

3)E 步骤：计算每个 λ_i 的期望数 $c(\lambda_i)=\sum_{j=1\cdots|H|}(\lambda_i p_i(w_j|H_j)/p'_\lambda(w_j|h_j))$；

4)M步骤:根据$c(\lambda_i)$,得到新一轮的λ_i值,$\lambda_i^{new}=c(\lambda_i)/\sum_{k=0}^{s}c(\lambda_k)$;

5)判断λ_i是否已经收敛:如果$|\lambda_i-\lambda_i^{new}|<\varepsilon, i=0,1,2,3$,则参数估计过程结束;否则,转步骤2),继续迭代计算。

其中:λ_i的初始值设置为0.25,式(3)中$|V|$的值从数据H中统计得到。此外,在词汇表中引入两个特殊符号,〈BOS_1〉和〈BOS 0〉。把它们附加到每个句子前面,分别表示句子开始之前的前两个位置和前一个位置。

3 实验

多语种识别实验系统MuLid由三个模块构成:语言建模工具;解码和语种选择工具;性能测试工具。基于EM算法的参数估计在建模工具中进行。模块间处理流程如图1所示。

实验在专门构建的一组语料集上进行。该语料集含8种语言,覆盖了IHSMTS多语机器翻译系统[5]所支持的主要语种,并有针对性的包括了4种双字节编码的语言:简体中文(GB2312),繁体中文(BIG5),日文(Shift-JIS),韩文(eucKR)。其他语言包括英文(ISO88591),俄文(KOI8_R),法文(ISO88591),德文(ISO88591),皆为单字节编码。每种语言的训练语料由三个纯文本文件构成,规模分别为10k、20k和50k,内容来自于互联网新闻网站并经过简单格式处理(去除HTML标记、错误的回车换行符等)。各语种的测试语料为200个长度为100字节的句子,共计8个文件1600句。

采用常用的查准率和查全率作为系统性能指标。分别定义:查准率$p=N_3/N_2$,查全率$r=N_3/N_1$,以及综合反映二者的指标$F=(\beta^2+1)\times p\times r/(r+\beta^2\times p)=2\times p\times r/(p+r)$(取$\beta=1$)。其中$N_1$为测试语料中所含某语种样本的正确数量,$N_2$为实际识别出的某语种样本数量,即MuLid系统输出的某语种样本数量,N_3为MuLid系统正确识别出的某语种样本数量。

为考察识别性能随语料规模的变化,使用不同容量的训练语料做3轮测试。每轮测试中,训练语料又被细分为两部分:90%用作训练数据T,10%用作支持数据H(held-out data)。前者用于p_3、p_2、p_1、$|V|$的估计,后者用于λ_i的估计。识别过程在从测试语料中取出的长度不同的三个样本上进行。

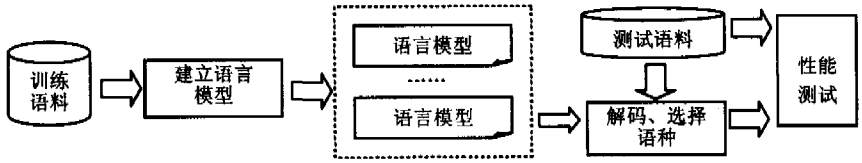

图1 Mulid系统结构和处理流程

表1给出了在长度为50字节的测试样本集上的识别实验结果,其中各语言模型在规模为50k的语料上训练得到。总体来看,识别性能是令人满意的,基本能达到实用要求。实际上,具体应用往往并不需要区分如此多的语种,例如,仅需要分辨是简体还是繁体,是英文还是俄文。这种情况下,系统的识别查准率、查全率,以及时间性能都将有进一步提升。

表1 8种主要语言上的识别结果

语种	简中	繁中	日文	韩文
编码	GB2312	BIG5	Shift_JIS	eucKR
F值	93.9%	94.3%	93.2%	95.4%
均值	94.2%			
语种	英文	俄文	法文	德文
编码	ISO88591			
F值	95.4%	96.2%	94.6%	95.8%
均值	95.5%			

表2和表3给出了简体中文和英文上的更详细的实验结果。可以看出,训练语料的规模对系统性能有较大影响。

表2 简体中文实验数据

语料规模	10字节		50字节	
	p	r	p	r
10k	92.5%	91.8%	94.6%	92.2%
50k	95.1%	94.7%	96.6%	94.3%

表3 英文实验数据

语料规模	10字节		50字节	
	P	r	p	r
10k	87.2%	84.1%	92.2%	90.8%
50k	94.6%	92.9%	94.7%	93.2%

这在双字节编码的语言上表现得更为明显。对于中文,当识别长度较短的文本(10字节)时,50k文本上得到的模型的性能要高于10k文本上的模型约7个百分点。同时,输入文本的长度也影响着识别性能,当模型本身不够充分时,输入长度对识别性能的影响更为显著。此外,通过两种语言的对比可以看出,英文对于训练规模的敏感度要小于中文:当训练规模为10k时,英文的识别能力已经和50k训练数据时比较接近。因此可以考虑对于不同的语言采用不同规模的训练数据以进一步优化语言模型、提高时间性能。

此外,对基于EM算法的参数估计方法和文[2]使用的Laplace公式方法做了定量的对比。这是通过对比两种方法在相同的10k简体中文语料上得到的语言模型的混杂度(perplexity)进行的。混杂度是衡量语言模型的常用指标,其定义为

$$PP(X,p)=[\prod_{i=1,\cdots,n}p(w_i|w_1,\cdots,w_{i-1})]^{-1/n}$$

PP表示用该语言模型描述特定文本平均的分支数,其倒数可视为每个词的平均概率。本文方法得到的语言模型的PP值和使用的Laplace公式建立的语言模型的PP值相差10%左右,显示出基于EM算法的参数估计的有效性。

4 相关研究

在语言模型的训练方面,文[1]在每个词的基础上生成n-gram并做简单统计,文[2]采用Laplace公式对马尔科夫语言模型进行数据平滑。文[1]方法简单易行,但意味着对于汉语、日语等没有空格分隔的语言,需要在分词之后的文本上才能训练语言模型,识别过程也需要分词后的文本作为输入。文[2]虽然在英语和西班牙语两种语言的区分实验中获得了成功,但对于汉语等大字符集合的语种识别问题,表现并不理

(下转第235页)

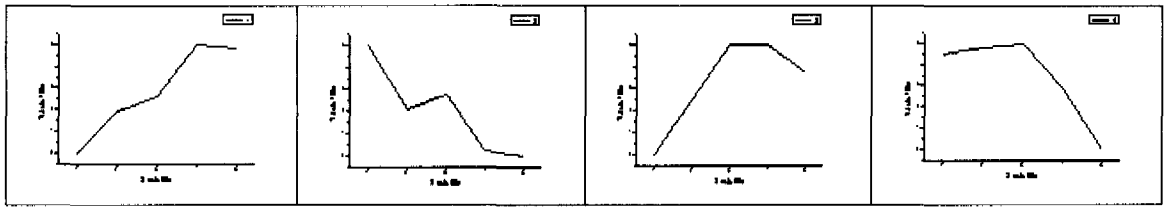

图 2　股票时间序列的部分频繁片段模式

4.4　获得的规则

我们以聚类得到的频繁片段模式作为模板，设定 IT-ARM 算法的参数：最小支持度 0.5%、最小信任度 60%、滑动时间窗口 3、规则长度 3，发现了许多关联规则。以下是部分规则。

R1：宏盛科技第 0 天是模式 11⇒青海三普第 13 天是模式 2(0.5%，83.3%)

R2：云维科技和友好集团第 0 天是模式 0⇒ST 轻骑第 1 天是模式 0(1%，90%)

R3：青岛啤酒第 0 天是模式 0，上菱电器第 0 天是模式 7⇒交运股份第 12 天是模式 0(0.5%，83.3)

结论　我们对基于片断模式的多时间序列关联分析进行了研究，提出了一种分析方法。这一方法是，首先通过聚类找出在时间序列中频繁出现的片断模式，然后将找到的片断模式作为模板，对时间序列进行跨事务关联分析。我们采用中国证券市场 1997～2001 年的数据为测试数据集，对我们提出的算法进行了测试。测试结果表明，我们的算法是有效的，它能发现多个股票之间的频繁片断模式的相互关联性，并以此预测一些股票的未来走势。

参 考 文 献

1　Lu H, Feng L, Han J. Beyond Intra-Transaction Association Analysis: Mining Multi-Dimensional Inter-Transaction association rules. ACM Transactions on Information Systems, 2000, 18(4): 423～454

2　Das G, Lin K, Mannila H, et al. Rule discovery from time series. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98). AAAI Press, 1998. http://citeseer.ist.psu.edu/das98rule.html

3　董泽坤，李辉，史忠植．多元时间序列中跨事务关联规则分析的高效处理算法．计算机科学，2004，31(3)：108～111

4　秦亮曦，史忠植．多时间序列跨事务关联分析研究(已投《软件学报》)

5　Berndt D, Clifford J. Using dynamic time warping to find patterns in time series. In: Working Notes of the Knowledge Discovery in Databases Workshop, 1994. 359～370

6　Qin L, Luo P, Shi Z. Efficiently mining frequent itemsets with compact FP-tree. In: Shi Z, He Q, eds. Proc. of Int'l Conf. on Intelligent Information Processing 2004(IIP2004), Beijing, China Springer Press, 2004. 397～406

(上接第 228 页)

想。如实验部分所述，本文方法能够建立更优的语言模型。

在研究对象和实验方面，文[1]和文[2]都只在单字节的一些西方语言字符编码方案上进行了实验。实际上，根据我们的调研，尚没有任何文献发表在中文、日文等双字节编码的语言上的语种识别实验结果。本文重点实验了针对 4 种双字节编码方案 GB2312、BIG5、Shift-JIS 和 eucKR 的语种识别效果；同时，在实验中对比研究了英、俄、法、德等 4 种单字节编码语种。

还有一些面向其他应用(如语音识别)的研究工作讨论如何进一步改善语言模型，如文[6,7]等。从实验结果看，我们认为更复杂的语言模型对于语种识别任务并不是必要的，孤立地追求训练过程的复杂化甚至可能对时间性能带来负面影响。

结论　本文研究了基于字符层马尔科夫模型的语种识别方法。有针对性的考虑了前人工作未曾很好解决的汉、日等双字节编码、词间无空格分隔的语种的识别问题，采用了一种基于 EM 算法的统计语言模型训练算法，同时也对解码算法进行了改进。本文工作的特点不仅体现在引入了基于 EM 算法的模型参数估计能够得到更为优化的语言模型，还体现在对于前人未讨论过的中、日等双字节编码、无空格分隔的语种的识别取得了良好实验结果。

基于字符层马尔科夫模型的语种识别方法具有较好的实用性，为我们的多语机器翻译系统[5]和多语言信息抽取研究提供了有益支持。但当前工作也存在些需要进一步改进的地方。后继工作将更多地面向算法的实用化问题，包括阈值参数 C_0 的自动选择、识别算法时间性能的改进等。

参 考 文 献

1　Cavnar W B, Trenkle J M. N-gram based text categorization. In 1994 Symposium on Document Analysis and Information Retrieval in Las Vegas, 1994

2　Ted D. Statistical Identification of Language: [Technical report CRL MCCS-94-273]. Computing Research Lab, New Mexico State University, 1994

3　Jelinek F, Mercer R L. Interpolated estimation of Markov source parameters from sparse data. In: Proc. of the Workshop on Pattern Recognition in Practice, Amsterdam, The Netherlands: North-Holland, 1980

4　Dempster A, Laird N, Rubin D. Maximum-likelihood from Incomplete Data via the EM algorithm. J. Royal Statist. Soc. Ser. B., 1977(39): 278～286

5　黄河燕，陈肇雄．基于多策略的交互式智能辅助翻译平台总体设计．计算机研究与发展，2004，41(7)：1266～1272

6　Goodman J T. A Bit of Progress in Language Modeling Extended Version: [Technical Report MSR-TR-2001-72]. Microsoft Research, Redmond, July 2004

7　Chelba C. Exploiting Syntactic Structure for Natural Language Modeling: [Phd Thesis]. Johns Hopkins University, 2004