计算机科学 2006Vol. 33№. 4

基于语言模型的联语应对研究*)

易 勇 何中市 李良炎 周剑勇 瞿义玻 张红兵
(重庆大学计算机学院 重庆 400044)

摘 要 基于机器学习方法和对联语料库,依据 n-元统计语言模型和隐马尔科夫模型,本文提出了联语应对的理论计算模型,并在以上方法和对联语料库的基础上,构造了计算机联语应对实验系统,取得了比较满意的结果。

关键词 自然语言处理,n-元语言模型,机器学习,隐马氏模型,对联

On Computation Medels of Chinese Couplet Responses

YI Yong HE Zhong-Shi LI Liang-Yan ZHOU Jian-Yong QU Yi-Bo ZHANG Hong-Bin
(College of Computer, Chongqing University, Chongqing 400044)

Abstract Based on machine learning methods and the corpus of Chinese Couplets, this paper proposes two computation models of Chinese Couplet responses according to HMM models and statistical language models. Furthermore, this paper built a computerized Chinese Couplet responses experiment system and achieved satisfactory results.

Keywords NLP, Language model, Machine learning, Chinese couplets

1 引言

联语是书面对联的雅称,联语应对俗称对对子、写对子、对对联。联语是汉语的独家绝唱,是中国文学的重要形式,在中国至少有一千多年的历史,它曾经发挥过其他文学样式所不曾发挥过的作用,深为群众喜闻乐见,曾经产生过浩如烟海的作品,其中不乏脍炙人口的佳作,充分体现了中国人的智慧,体现了形象思维与抽象思维的结合。好的对联同好的诗歌、小说一样能令人爱不释手,从中得到深刻的思想启迪和高度的艺术享受。

计算机理论和技术迅速地向各个领域扩展渗透,联语这一古老而又长青的语言艺术形式,如何在计算机技术高速发展的今天,用电脑智能地模仿人类联语生成的能力,具有重要意义。本文在研究机器学习的基础上,根据联语的出句,即上联,用电脑来应对生成联语的下联语句。

2 相关工作

对联是文学的一种形式,也是一种比较特殊的自然语言现象,是否可以运用相关的自然语言处理和新的语言学的方法和技术来研究,根据本文的工作,研究结果的回答是肯定的。自然语言处理技术在过去 20 年得到飞速发展,诞生了基于规则的语言处理技术,基于统计模型的语言处理技术和基于实例的语言处理技术,同时语言学理论和方法在计算机技术的辅佐下,迅速推进,产生了语料库语言学,语料库和机器学习的方法相得益彰,珠联璧合,语料库成为机器学习的原始数据源泉,机器学习为语料库提供了崭新的计算机处理和研究方法,语料库技术和机器学习方法是本文研究联语的应对生成问题的技术路线。

2.1 序列学习问题

分析对联艺术,搁置传统对联要求词性相对、语义类别相对、音韵相对等要求。从简化理论建模的角度,发现对联的上下联分别可以单纯地视为两串具有相同长度的语言单位的序列,本文所称的语言单位,可以是汉语中的字、词、字组、成语等,那么联语应对生成过程就可以视为机器学习中的序列学习问题,可表述为:

已知:给定形式为 (X_i, Y_i) 训练实例集合,其中 $X_i = hx_{i,1}, \cdots, x_{i,T_i}$; $Y_i = hy_{i,1}, \cdots, y_{i,T_i}$, T_i 是序列的长度,求出一个函数 $Y = f(X)$ 以便预测新的序列。

序列元素间有两种重要的形式关联,第一种关联:$xt's$ 和 $yt's$ 之间的关系,例如在对联中"云"常与"雨"相对,"雪"常与"风"相对,"晚照"常与"晴空"相对,"来鸿"常与"去燕"相对。第二种关联:$yt's$ 间的关联,例如在对联中,"英雄"常常跟在"千古"之后。

具体地对于联语问题而言,我们可以用已经趋于成熟的统计语言模型来进行序列的学习。

2.2 N 元统计语言模型序列学习方法

2.2.1 N 元模型(n-gram model)

如果用变量 W 代表一个一幅对联的上联中顺序排列的 n 个语言单位,即 $W = w_1 w_2 \cdots w_n$,统计语言模型的任务是给出任意语言单位序列 W 在对联语料库中出现的概率 $P(W)$。利用概率的乘积公式,$P(W)$ 可展开为:

$$P(W) = P(w_1)P(w_2 | w_1)P(w_3 | w_1 w_2)\cdots P(w_n | w_1 w_2 \cdots w_{n-1})$$

不难看出,为了预测语言单位 w_n 的出现概率,必须已知它前面所有语言单位的出现概率。从计算上来看,这太复杂了。如果任意一个语言单位 w_i 的出现概率只同它前面的 $N-1$ 个语言单位有关,问题就可以得到很大的简化。这时的语言模型叫做 N 元模型 (N-gram),即

$$P(W) = P(w_1)P(w_2 | w_1)P(w_3 | w_1 w_2)\cdots P(w_i | w_{i-N+1} \cdots w_{i-1}) \cdots \approx ?_{i=1,\cdots,n} P(w_i | w_{i-N+1} \cdots w_{i-1})$$

*)本研究受国家自然科学基金支持(项目编号 60173060)。易 勇 博士研究生,主要研究方向:自然语言处理,机器学习,数据挖掘。何中市 教授,主要研究方向:自然语言处理,机器学习,概率论,容错计算。李良炎 博士研究生,主要研究方向:自然语言处理,机器学习,心理学。

符号$?_{i=1,\cdots,n}P(\cdots)$表示概率的连乘。实际使用的通常是 $N=2$ 或 $N=3$ 的二元模型(bi-gram)或三元模型(tri-gram)。以三元模型为例,近似认为任意语言单位 w_i 的出现概率只同它紧前面的两个语言单位有关,即

$$P(W)\approx?_{i=1,\cdots,n}P(w_i|w_{i-2}\ w_{-1})$$

重要的是这些概率参数都是可以通过大规模对联语料库来估值的。比如三元概率有

$$P(w_i|w_{i-2}\ w_{i-1})\approx count(w_{i-2}\ w_{i-1}\ w_i)/\ count(w_{i-2}\ w_{i-1})$$

式中 $count(\cdots)$表示一个特定语言单位序列在整个对联语料库中出现的累计次数。

联语应对生成的过程。从序列学习的角度来分析,就可以理解为给定一个对联的上联的语言单位序列 w,推断最有可能的下联的语言单位序列 T。应用 Bayes 公式,下联的语言单位序列 T 的后验概率为:

$$P(T|W)=\frac{P(T)*P(W|T)}{P(W)} \quad (1)$$

其中,$P(T)$是对联的下联的语言单位序列 T 的先验概率(prior probability),$P(W|T)$是对联的下联的语言单位序列 T 已知情况下上联语言单位序列 W 发生的条件概率,$P(W)$是语言单位序列的非条件概率。

在语言单位序列已知时,$P(W)$对于所有可能的下联语言单位序列都是相同的,因此可以不考虑 $P(W)$,而用 $P(T)$与 $P(W)$的乘积来表示每个下联语言单位序列的后验概率。

$$P(T)*P(W|T=\prod P(t_0)P(t_1|t_0)P(t_i|t_{i-1},t_{i-2},\cdots,t_1)P(W_i|t_i,\cdots,t_1,W_{i-1},\cdots,W_1)$$

其中,$t_0,t_1,\cdots$表示下联的语言单位序列,$W_0,W_1,\cdots$表示上联的语言单位序列。

对于 n-gram model,可以假设每个下联的语言单位序列只与前面相邻的语言单位有关,(如,bigram,模型只与前面紧邻的一个语言单位序列有关),并且上联的每个语言单位只与下联的与之对齐的语言单位有关。于是,对于(1)式中分子的第一项,假设每条上联的每个语言单位的所对应的下联的语言单位只与其下联先前一个语言单位有关,则有:

$$P(T)=\prod_{i=1}^{n}P(t_i|t_1^{i-1})\approx P(t_1)\prod_{i=2}^{n}P(t_i|t_{i-1}) \quad \text{bi-gram model}$$

对于(1)式中分子的第一项,假设每条上联的每个语言单位的所对应的下联的语言单位只与其下联先前两个语言单位有关,则有:

$$P(T)\approx P(t_1)P(t_2|t_1)\prod_{i=3}^{n}P(t_i|t_{i-1}t_{i-2}) \quad \text{tri-gram model}$$

因为上联的语言单位序列 W 不变,所以 W 不影响求 $P(T|W)$的最大值。

每幅对联下联的语言单位序列的概率可以近似估计为下联语言单位序列的同现概率的联合分布率,其中最大的是下联语言单位序列的极大似然估计值。

根据 Viterbi 算法,概率最大的结果为下联产生的结果。则

$$P'(T|W)=\max P(T|W)=\max P(T)*P(W|T)=P(t_1)P(w_1|t_1)\prod_{i=2}^{n}P(t_i|t_{i-1})P(w_i|t_i) \quad \text{bi-gram model}$$

$$P'(T|W)=\max P(T|W)=\max P(T)*P(W|T)=\max P(t_1)P(t_2|t_1)P(w_1|t_1)P(w_2|t_1)\prod_{i=3}^{n}P(t_i|t_{i-1}t_{i-2})P(w_i|t_i)$$

tri-gram model

2.2.2 下联生成的算法

• 将欲处理对联的上联分割为 N 个语言单位;

• 对上述 N 个语言单位,首先查数据字典,找出语料库中所有可能的映射下联的语言单位;

• N 个相邻的上联的语言单位的每一种映射的语言单位的排列序列为一条路径;

• 根据 Viterbi 算法求出具有最大似然估计值的那条路径。

最佳路径上所对应的语言单位序列为上联所对应的下联。

2.3 HMM 模型序列学习方法

隐马尔可夫模型(HMM)是一个二重马尔科夫随机过程:一个潜在的随机过程,它是不可观察的(隐含),但是它能通过另外一个产生可观察随机序列进行观察。隐马尔可夫模型已经成功地应用于连续语音识别、手写体识别、词性标注、基因分析、浅层语法分析等领域。在统计自然语言处理中,该模型也占据着重要的地位。

联语应对的过程可以看成是从上联的语言单位序列到下联的语言单位序列的映射过程,这样的一个过程可以通过建立语言模型,用机器学习的方法来实现。这里的语言学模型,亦即语言学信息及其处理的形式化。

在联语生成中,把上联(出句)中每个语言单位及其下联对应的语言单位的出现都看作一个随机过程。上联(出句)中每个语言单位出现的序列作为可观察的,下联对应的语言单位序列作为一个隐含过程,其下联联语的生成过程,就是通过上联(出句)中每个语言单位(可观察的)去观察下联对应的语言单位序列(隐含的)。于是联语应对生成的问题,用 HMM 模型的方法描述为:

已知上联语言单位序列 $w_1w_2\cdots w_n$,求下联的语言单位序列 $t_1t_2\cdots t_n$

• HMM 模型:将下联的语言单位理解为状态,将上联的语言单位理解为输出值。

• 模型训练:统计下联语言单位序列转移矩阵[aij]和上下联对齐语言单位序列的输出矩阵[bik]。

求解:用 Viterbi 算法求出最佳的下联语言单位序列。

3 联语应对实验

在联语应对研究实验中,依据研究的需要多方收集了 9316 幅对联,建立一个对联语料库,在其中分别标识了上下联和联字的对齐,分别用 HMM 模型、Trigram 语言模型方法进行学习,在 CPU 为 intel pentiumⅣ ,主频为 2.4G,内存为 1000M 的 Windows 平台上,用高级语言编写成程序,并选取当今的春联和诗词对仗诗句来进行实验,其中三元统计语言模型方法 corpus 的训练时间为 16.68 秒,下联生成的时间为 17.66 秒。HMM 模型方法 corpus 的训练时间为 35.091 秒,下联生成的时间为 22598.665 秒,其部分实验结果如下;

上 联:	四海华人,一宵除夕,万家灯火,亿户荧屏,多少同胞同过节,
HMM 模型生成的下联:	九州江山,千门迎春,千里江山,千家家乐,不老一心共迎春。
Trigram 模型生成的下联	九州纪大,千夜安春,千处月风,州家家花,好多共祖共迎春。

(下转第 173 页)

speech and noise. In: ICASSP, 1990, 2. 845～848
27 Warren R, Riener K, J B Jr, Brubaker B. Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. Perception and Psychophysics, 1995, 57(2): 175～182
28 Zweig G. Speech Recognition with Dynamic Bayesian Networks: [Ph D Thesis]. Berkeley: University of California, 1998
29 Young S, et al. The HTK Book (for HTK Version 3.2), December 2002
30 Young S. Large Vocabulary Continuous Speech Recognition: a Review. In: Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding, Utah: Snowbird, December 1995. 3～28
31 Huang X, Acero A, Hon H. Spoken Language Processing. Prentice Hall, 2001. 375～538
32 Gong Y F. Speech Recognition in Noisy Environments: a Survey. Speech Communication, 1995, 16: 261～291
33 Juang B H. Speech Recognition in Adverse Environments. Computer Speech And Language, 1991, 5: 275～294
34 Openshaw J P, Mason J S. On the Limitations of Cepstral Features in Noise. In: Proc of ICASSP 1994, Adelaide, Australia, April 1994. 49～52
35 Hermansky H. Should Recognizers Have Ears? Speech Communication, 1998. 25: 3～27
36 Chen C P, Filali K, Bilmes J A. FrontEnd Post-Processing and BackEnd Model Enhancement on the Aurora 2.0/3.0 Databases. In: Proc. of ICSLP 2002. Denver, Colorado, September 2002. 241～244
37 Haeb-Umbach R, Ney H. Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition. Proc of IEEE on Acoustics, Speech and Signal Processing, 1992, 1: 13～16
38 Ho D K C. Speech Enhancement: Concept and MethodologyUniversity of Missouri-Columbia
39 Juang B H, Soong F K. Hands-free Telecommunication. In: HSC 2001(International Workshop on Hands-Free Speech Communication), 2001. 5～10
40 Omologo M, Svaizer P, Matassoni M. Environmental Conditions and Acoustic Transduction in Hands-Free Speech Recognition. Speech Communication, 1998, 25: 75～95
41 Elko G W. Microphone Arrays. In: HSC 2001, 2001. 11～14
42 Omologo M. Hands-Free Speech Recognition: Current Activities and Future Trends. In: HSC 2001, 2001. 23～26
43 Mangold H A. Realistic Hands-Free Applications - the Key for Really User-Friendly Applications. In: HSC 2001, 2001. 35～38
44 Gong Y. A Robust Continuous Speech Recognition System for Mobile Information Devices. In: HSC 2001, 2001. 39～42
45 Buchner H, Herbordt W, Kellermann W. An Efficient Combination of Multi-Channel Acoustic Echo Cancellation with A Beamforming Microphone Array. In: HSC 2001, 2001. 55～58
46 McCowan I A, Bourlard H. Microphone Array Post-Filter for Diffuse Noise Field. In: ICASSP 2002, 2002. 905～908
47 Fujimoto M, Ariki Y. Noise Robust Hands-Free Speech Recognition Using Microphone Array and Kalman Filter as Front-End System of Conversational TV. MMSP2002, 2002
48 Saruwatai H, Kajita S, Takeda K, et al. Speech Enhancement Using Nonlinear Microphone Array with Noise Adaptive Complementary Beamforming. In: ICASSP'00, 2000. 1049～1052
49 Acero A. Acoustical and Environmental Robustness in Automatic Speech Recognition: [Ph D thesis]. Carnegie Mellon University, 1990
50 Moreno P. Speech Recognition in Noisy Environments: [Ph D thesis]. Carnegie: Mellon University, 1996
51 Liu F H, Stern R M, Acero A, et al. Environment Normalization for Robust Speech using Direct Cepstral Comparison. In: Proc. of ICASSP 1994, Adelaide, Australia, April 1994. 61～64
52 Deng L, Acero A, Plumpe M, et al. Large-Vocabulary Speech Recognition under Adverse Acoustic Environments. In: Proc. of ICSLP 2000, Beijing, China, October 2000. 806～809
53 Deng L, Droppo J, Acero A. Recursive Estimation of Nonstationary Noise Using Iterative Stochastic Approximation for Robust Speech Recognition. IEEE Transactions on Speech and Audio Processing, 2003, 11(6)
54 Droppo J, Acero A, Deng L. Efficient On-Line Acoustic Environment Estimation for FCDCN in a Continuous Speech Recognition System. In: Proc. of ICASSP 2001, Salt Lake City, Utah, May 2001. 209～212
55 Droppo J, Acero A, Deng L. Uncertainty Decoding with SPLICE for Noise Robust Speech Recognition. In: Proc of ICASSP 2002, Orlando, Florida, May 2002. 57～60
56 Droppo J, Deng L, Acero A. Evaluation of the SPLICE Algorithm on the Aurora2 Database. In: Proc. of Eurospeech 2001, Aalborg, Denmark, September 2001. 217～220
57 Droppo J, Deng L, Acero A. Evaluation of SPLICE on the Aurora 2 and 3Tasks. In: Proc of ICSLP 2002, Denver, Colorado, September 2002. 29～32, 977～980

(上接第 163 页)

上　联：	两只黄鹂鸣翠柳；
HMM 模型生成的下联	九万紫燕舞红梅。
Trigram 模型生成的下联	三正紫燕舞晴空。

上　联：	爆竹一声除千户，
HMM 模型生成的下联	梅花千树新万里。
Trigram 模型生成的下联	梅花千色安万家。

上　联：	大江东去，浪淘尽千古英雄；
HMM 模型生成的下联：	江山春来，山河增百年盛世。

上　联：	九天揽月，华夏英豪驰宇宙；
HMM 模型生成的下联：	四海迎春，神州崛起舞天下。

上　联：	天王盖地虎；
HMM 模型生成的下联：	地焗成天牛。

从实验结果来看，用机器来应对联语的下联取得了比较满意的结果，生成的下联比较符合语义、语法，意境也大致切合。从两种模型的实验对比来看，基于 HMM 模型的对联语料机器学习和下联生成的时间都较长，Tri-gram 语言模型的机器学习时间和下联生成的时间都较短，但是 HMM 模型生成的下联明显比 Trigram 语言模型的质量好。

结论和展望　文学语言的处理与生成，是计算机艺术的重要内容，本文在对联应对生成方面所作的探索，再一次印证了机器学习在模拟人类智能方面的能力。下一阶段的研究，我们还将继续改进联语应对生成的方法，融合传统的对联创作方法，综合语义关系、语法关系、意境、主题等语言和文学的信息，并尝试用其他的机器学习方法来研究对联的应对生成问题。

参 考 文 献

1 Yi Yong, He Zhongshi, Li Liangyan. Studies on Traditional Chinese Poetry Style Identification. In: the Proc. of ICMLC04. Shanghai, 2004. 2936～2939
2 Christopher D. Manning, Foundation of Statistical Natural Language Processing. Pearson Education, 1998
3 Allen J. Natural Language Understanding, Second Esition. Pearson Education, Inc., 1995
4 Mitchell T M. Machine Learning, McGraw-Hill Companies, 1997
5 朱承平. 诗词格律教程. 暨南大学出版社, 1999