统计与规则结合的古文对联应对模型[1]

张开旭 孙茂松

智能技术与系统国家重点实验室，清华信息科学与技术国家实验室（筹）

清华大学计算机科学与技术系 北京 100084

E-mail: zhangkx03@mails.thu.edu.cn, sms@tsinghua.edu.cn

摘要：本文将古文对联规则区分为硬规则与软规则，用软规则指导建立对联应对的有向概率图模型，使用 EM 算法估计模型参数，在解的搜索过程中加入硬规则而完全实现对联的自动应对。实验结果表明参数学习后的候选字列表由于去除了部分上下文的影响，比仅用频次统计的候选字列表更为合理。系统能够对训练语料库中工整与不工整的对联区分学习。最后的对联效果也达到了一定水平。

关键字： 对联 最大熵马尔可夫模型

An ancient Chinese couplet generating model based on statistics and rules

Kaixu Zhang, Maosong Sun

State Key Laboratory on Intelligent Technology and Systems

Tsinghua National Laboratory for Information Science and Technology

Department of Computer Science and Technology, Tsinghua University, Beijing 100084

E-mail: zhangkx03@mails.thu.edu.cn, sms@tsinghua.edu.cn

Abstract: In this article, the rules of Chinese couplet were divided into hard rules and soft rules. A probabilistic graphic model was proposed based on the soft rules. Parameters were estimated by EM algorithm. System was completed by adding hard rules into the solution finding procedure. The experiment result showed that the candidate character lists of this model are better than the one based on only counting, for the reduced effect of the context. The system can learn from the data set witch contents worse couplets. Finally the system shows an acceptable performance.

Keyword: Chinese couplet, maximum entropy Markov model.

1 引言

对联是中华传统文化的组成部分。给出上联，思考与之相对的下联，可谓最古老的语言游戏之一。它集趣味性、逻辑性、艺术性于一体。自古文人雅士乐此不疲，许多关于对联的佳话也千古流传。

能够自动对联的计算机程序，有较强的娱乐性，也可帮助不太熟悉对联的使用者得到效果较好的对联为语言润色，给文章点睛，打趣解闷，赠送亲朋，甚至可以帮助诗歌等文学作品的创作，对传承中华文化有一定的推动作用。因此，计算机对联系统不但可以加深对自然语言的理解，也有潜在的经济价值与文化价值。

在使用计算机自动对联的研究中，公开发表的成果很少。微软亚洲研究院在互联网上发布了一个对联系统，对古文与白话文都可以应对。系统较为完善，电脑生成的下联不乏绝妙奇趣之

[1]本研究得到国家 863 计划支持（项目号：2007AA01Z148）。

作，有很好的效果。但并无关于此系统的论文，只有一篇专利[5]可作参考。唯一找到的一篇关于对联的论文[6]用了较为简单的隐马尔可夫模型，其对联效果远无法与微软亚洲研究院的相比。

形而上地说，我们认为，下联是对上联的模仿。具体讲，对联的规则大致可分为软规则与硬规则两类。软规则是指在应对对联时，不但要求上下联对应字词类相同语义相关，而且上下文的句法结构也应一致，更高要求是上下联的意思、意境要有联系。这些规则都不太容易用逻辑表达式形式化的描述，而且是否满足规则是有程度区别的。除此之外，还有一些硬规则，比如上联出现的字下联一定不出现，上联不同位置用了相同的字，则下联对应位置也必须用相同的字等。与软规则不同，这些规则约束着在位置上可能相差很远的字，且他们都容易写成逻辑表达式，其对下联只做满足与不满足规则两种硬性区分。

编写一个应对古文的对联机，我们试图面对以下问题：

- 下联的某个位置用什么字，不但与上联对应字有关，也与下联上下文有关，如何将这两个因素组合，即不显得鲁莽无理，又较为符合实际情况。
- 训练集的下联也是权衡了单字相对与上下文相对的因素后得到的，我们如何从中提取到只考虑字的对应关系而不考虑上下文时候，某个字比较合理的候选字列表，这样的列表对我们人加深对对联现象的理解也是有益处的。
- 本系统使用全唐诗作为训练语料库，然而即使经过筛选，语料库中也存在大量非对仗的句子，对仗的质量也不尽相同，如何在这样的训练集上进行统计学习。

软规则很容易用统计的方法量化，但很难逻辑化，所以基于规则的模型难以刻画软规则。而硬规则的出现又是相当稀疏的，用统计模型也根本无法学习到。所以，我们同时使用基于统计与规则的方法。首先我们用统计的方法对软规则建立一个产生式模型，给每个候选的下联一个概率值。我们的目标是搜索这样的候选解空间，找到似然值最大的若干解。而在搜索的过程中我们排除那些在硬规则下不被允许的解。在第一步不考虑硬规则的合理性是因为硬规则的引入只是在这样的较为平滑的概率分布上引入一些零散的崎岖点，使这些地方的概率为 0，并不会太影响整个概率分布，且不会影响任意两个解似然值的大小关系。

本文以后的结构为：第二节用概率图模型对对联进行建模；第三节介绍如何用 EM 算法学习这样的概率图模型的参数；第四节介绍将硬规则引入后形成的组合优化模型及其解法；第五节介绍实验以及结果；第六节是对本文的总结。

2 对联的概率模型

模型一般分为判别式模型与产生式模型。对于对联问题，除非找到按“所有非对联的下句的分布”的负例样本集合，否则无法将其与对联一起学习有效的判别模型，用以评价下联的好坏。故最好是使用产生式模型。

在此我们试图将对联较之与标注模型，即与目前流行的基于标注的中文分词问题是一个框架。该类问题常用模型有隐马尔可夫模型、最大熵马尔可夫模型和条件随机场模型，能力由弱到强。故最好的概率模型是条件随机场(CRF)[3]。然而遗憾的是与分词相比，其标注并非是四五个，而是上千个，这使得特征数量庞大到普通计算机无法承受。

究其原因，是因为条件随机场的模型用概率图描述的话属于无向图模型，对于无向图模型没有很好的可分解性，故即使是无隐变量的参数估计，也要用到全局的推理。

因此我们退而求其次，在隐马尔可夫模型与条件随机场模型之间，我们选择最大熵马尔可夫模型(MEMM)[2]作为对联概率模型的蓝本。

如图 1，是对联的概率图模型。其中u_i表示上联的第i个字，d_i表示下联的第i个字，c表示下联是否要对仗。汉字的集合为$W = \{\omega_1, \omega_2, \ldots, \omega_{|W|}\}$，我们将$u_0$与$d_0$定义为$\omega_0$，则当$c = 1$时（由于以后大部分情况下我们只是关心$c = 1$时的概率分布，故下文在没有歧义的情况下不特别将$c = 1$写在概率表达式中）整个概率图模型的概率可写成如下联乘积：

$$P(D|U, c = 1) = \prod_{i=1}^{n} P(d_i | u_{i-1}, u_i, d_{i-1})$$

即对于下联位置d_i选哪个字，只与u_{i-1}、u_i、d_{i-1}有关，而与其它位置的字以及其具体的位置i无关。

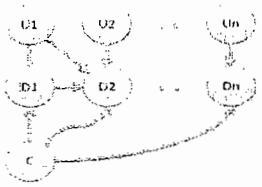

图 1：对联的概率图模型

下面我们进一步将上面联乘积的每一项分解，直观上我们对联选字时有两个需要考虑的因素，一是字与上联对应的字是否相对，这个函数我们用$q(d_i, u_i)$表示，还有就是其在下联上下文中的地位是否与上联对应字在上联上下文中的地位相仿，这个函数我们用$g(u_{i-1}, u_i, d_{i-1}, d_i)$表示。那么将这两种因素结合起来，朴素的方法则为：

$$P(d_i | u_{i-1}, u_i, d_{i-1}) \propto q^{\alpha}(d_i, u_i) g^{\beta}(u_{i-1}, u_i, d_{i-1}, d_i)$$

注意到$q(d_i, u_i)$并不与上下句对应字相对的概率$P(d_i|u_i)$相等，后者是综合以上所述两个因素后的结果。故q不能用语料库的统计量近似，每个u_i的$q(\omega, u_i)$必须有$|W|$个参数需要估计。

我们再将上式写作最大熵模型惯用的指数的形式，即为：

$$P(d_i | u_{i-1}, u_i, d_{i-1}) = \frac{1}{Z} \exp\left\{ \sum_{\omega} \lambda_{u_i, \omega} f_{\omega}(d_i) + \lambda_{u_i, g} g_{u_i}(u_{i-1}, u_i, d_{i-1}, d_i) \right\}$$

其中Z是归一化因子，f_{ω}是对应于q的特征函数，定义为：

$$f_{\omega}(d_i) = \begin{cases} 1, & d_i = \omega \\ 0, & others \end{cases}$$

$\lambda_{u_i, \omega}$即为前面讨论的$|W|$个参数。

对于g，我们只实现了一些最基本的刻画，用统计特征$P_g(u_i|u_{i-1})$与$P_g(d_i|d_{i-1})$表示上联与下列对应位置的上下文特征。其中$P_g(\omega_2|\omega_1)$表示在上下文中字ω_1之后出现字ω_2的概率，可

由语料库的 Bigram 得到。考虑到下联应与上联相仿，那么$P_B(d_i|d_{i-1})$必与$P_B(u_i|u_{i-1})$有相关性。比如对联“落花有意，流水无情”，上联的P_B(花|落)与P_B(意|有)值较大，P_B(有|花)较小，那么相应的下联P_B(水|流)与P_B(情|无)值较大，P_B(无|水)值较小。**错误！未找到引用源。**给出了他们之间的分布。

其与高斯分布十分相像，为验证猜想，用高斯混合模型（GMM）[4]，用 EM 算法，对 1814 个样本点学习，迭代 70 次，高斯分布数目与最后样本点对于模型的对数似然度关系如图 3。计算可得，用 10 个高斯分布只比用 1 个高斯模型每个样本的似然度平均增加了 1.01%，故用 1 个高斯分布拟合已经合理。

因此最后将g定义为下式的条件分布：

$$\frac{1}{2\pi|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

其中$x=[\ln P_B(u_i|u_{i-1}),\ln P_B(d_i|d_{i-1})]^T$。通过样本计算得到的高斯分布的均值向量与协方差矩阵为：

$$\mu=[-2.2776,-2.2481]^T$$

$$\Sigma=\begin{bmatrix}0.5379 & 0.2025\\ 0.2.25 & 0.5392\end{bmatrix}$$

我们这里只实现了一个很初步的关于上下文相对的特征g，如互信息量等特征的引入也可能提高效果。

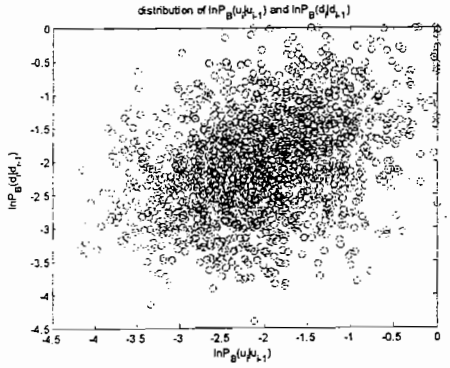

图 2：$P_B(u_i|u_{i-1})$与$P_B(d_i|d_{i-1})$的分布

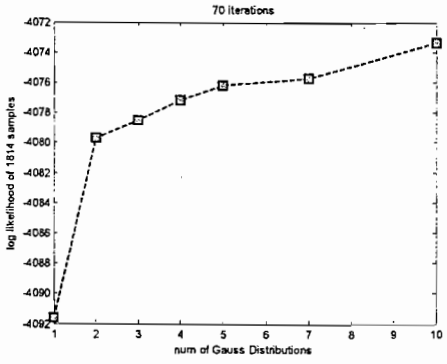

图 3：高斯分布数目与似然度的关系

3 EM 算法下的模型参数估计

对于图 1 的模型，存在一个隐变量c，因此适合使用 EM 算法[4]进行参数估计：首先在 E 步，在某组参数赋值下根据已知数据估计隐变量的概率分布，即对每个样本，用贝叶斯公式估计其为对联（$c=1$）的概率；再在 M 步，在所估计的隐变量的分布下，用使得样本似然度最大的参数取值来更新原参数的赋值。重复以上步骤直到收敛。

3.1 E 步

在 E 步我们需要由已知的UP与DP求得隐变量c的概率分布。由链式法则有：

$$P(c=1|DP)P(DP)=P(DP|c=1)P(c=1)$$

$$P(c=0|DP)P(DP)=P(DP|c=0)P(c=0)$$

联立求得：

$$P(d=1|DP)$$
$$=\frac{P(DP|c=1)}{\frac{P(c=0)}{P(c=1)}P(DP|c=0)+P(DP|c=1)}$$
$$\approx\frac{P(DP|c=1)}{d_n+P(DP|c=1)}$$

我们假设若非对联，下联任何句子都等概率出现，故等式右边分母第一项只与句子长度n有关，且$P(DP|c=1)$是我们可以求到的，为了使最后各个句子的c的分布合理，我们定义：

$d_n=\exp\{E\ln P(DP|c=1)\}$

这样我们就可以通过已知数据与给定的参数λ求得$P(DP|c=1)$进而求得c的概率分布了。

3.2 M 步

在 M 步我们要求使似然度最大的参数，我们通常优化其对数似然度：

$$\lambda^*=\arg\max_{\lambda}\ln E\prod_s l_s(\lambda)$$

其中$l_s(\lambda)$是句子s在参数为λ时出现的似然度。

由于计算复杂度的原因，直接整体优化对数的期望似然比较困难，句子多，参数有$|W|(|W|+1)$个。我们希望将其分解。由于对数函数是凸函数，利用简森不等式(Jensen's inequation)，我们转而优化其一个下界：

$$\ln E\prod_s l_s(\lambda)\geq E\ln\prod_s l_s(\lambda)$$

此时就可以利用有向概率图模型的可分解性，此下界可重写为：

$$E\ln\prod_s l_s(\lambda)$$
$$=\sum_s c_s\ln l_s(\lambda)$$
$$=\sum_{u\in W}\sum_{(u,u',d,d')}c_{u,u',d,d'}\ln P(d|d,u,u',\lambda_u)$$

其中c_s是句子s为对仗句的概率，四元组(u,u',d,d')是在语料库中出现的上联任意临近的两个字u,u'以及他们对应的下联的字(d,d')，$c_{u,u',d,d'}$是该四元组所在句子的c_s（若在多个句子中出现，则为之和）。此时可以看到外层求和号内关于u_s的每一项参数只有最多$|W|+1$，且句子大为减少，较之前面计算复杂度大大降低，如此表示之后使得用迭代算法求使似然度最大的参数成为可能。

在此我们使用最简单的迭代，对给定u时的每一个λ_u，将其更改为其使似然度最大的值，循环此过程直到似然度不再提高。

4 对联的组合优化模型

组合优化模型是人工智能中一类常见的模型，其任务是，给定一组离散变量和他们分别可以取得的有限个值，给定各种变量赋值下的目标函数，求使得目标函数最大的赋值。

对于学习完成的概率模型，给定一个上联之后，与隐马尔可夫模型一样，只需要用动态规划 Viterbi 算法就可以求出最优下联，即似然度最大的解。但如果引入对联的硬规则，使其图的数的树宽度变大，则求最优解的算法复杂度可能恶化成非多项式的。

硬规则的特点是其对对联的评价只有违反与不违反之分。违反规则的对联一定不可用，而不违反规则的对联之间，硬规则不能对取舍给出更多的指导作用。

我们将在不违反硬规则前提下，找似然度最大的若干解的问题归结为组合优化模型。即给定下联每个字的若干可能取值，对每一组赋值给出一个目标函数。若不违反硬规则，则目标函数值为该组赋值在概率模型中的似然度；若违反硬规则，则似然度为 0。

在搜索中加入的硬规则有：

- 下联不能使用上联使用过的字
- 若上来某两个字相同，则下联对应位置的字也应该相同，反之应该不同。

我们使用深度优先作为统一的算法。将下联每个字看作一个变量，取值范围为最能与上联对应字对上的前若干个字，搜索中保留最佳的若干解，将这些解中目标函数最小值作用阈值用于搜索剪枝。实验表明即使每个字使用 20 个候选字，搜索的速度也是非常快的。

5 实验结果

我们使用了全唐诗作为训练数据集。根据近体诗的格律特点[7]，那些有 8 句的诗歌很可能是律诗，则其颈联与颔联很可能是对仗句，可以作为概率模型的样本。我们一共使用了 38821 对句子训练。使用上面介绍的 EM 算法迭代 5 次后，基本收敛。其中的P_e 也是使用的全唐诗的 Bigram 得到。

前文提到，若只统计训练集中对应字的频次，是无法体现某两个字可否单独相对的程度的，而我们的模型区分了字单独相对于上下文通畅两个因素，其所得的字与字相对的列表应该比只统计频次的合理，并且有助于人对对联规律的研究，下面是与“月”相对的字的优先列表：

与“月”相对的字											
按频次排序	风	云	花	山	霜	春	烟	人	天	秋	星	年
按权重排序	风	霜	云	星	潮	花	钟	霞	灯	烟	河	泉

如“人”、“天”这样的高频字容易出现在按频次排序的列表中，因为高频字较容易形成好的上下文，而经过参数学习后其权重降低。下面是与“城”相对的字的优先列表：

与“城”相对的字											
按频次排序	水	树	路	寺	月	地	苑	海	国	日	驿	江
按权重排序	驿	苑	岳	阙	寺	成	浦	市	垒	路	邑	郭

“水”“树”这样的并不太工整但常用的对字在我们的模型中排名分别为 13、17；而如“阙”、“邑”和“郭”等比较好的候选字在我们的模型中权重都有所提高，而如果只按频次统计，此 3 字分别被排在了第 17、38、30 位。可见通过参数学习后的候选字列表更接近单独考虑字而不考虑上下文的情形。

我们再考察隐变量c的作用。语料库中有一八句的五言诗："钓濑水涟漪，富春山合沓。松上夜猿鸣，谷中清响合。冲网忽见羁，故山从此辞。无由碧潭饮，争接绿萝枝"，其中间四句会被当作对联的训练语料。然而观察发现其对仗并不工整，我们用句子本身估计其为对联的概率，下表为结果：

句子	0 次迭代	5 次迭代
松上夜猿鸣 谷中清响合	0.118	0.998
冲网忽见羁 故山从此辞	0.250	2.602×10^{-9}

经过迭代后，第三联很不工整，因此权重几乎为 0，便不会影响模型的参数取值。

加上硬规则之后形成的完整的对联应对系统，虽不是对于所有上联都能找到可以接受的下联，但在很多时候已经具有一定的对联能力，如：

上联	系统所对下联
远看山有色	遥听水无声
两岸晓烟杨柳绿	千山春雨杏花红
绿水悠悠青山隐隐	黄花寂寂白石苍苍
风声雨声读书声声声入耳	月色云色暖酒色色色连心

6 结论

本文提出对联就是下联对上联的模仿的思想，区分对联中硬规则与软规则两类规则。使用软规则进行有向概率模型学习，在最优解搜索中引入软规则，所实现的对联应对系统还比较简单，只考虑了当前字以及之前一个字的信息，上下文只使用了g一个简单的函数。若将其完善性能必会再提高。模型学习后，所提取出的单字因对列表比只用频次统计给出的合理，也能对训练集中工整与不工整的对联区分学习，综合硬规则之后所对出的某些下联已经具有一定水平。

参考文献

[1] Christopher M. Bishop, *Pattern Recognition And Machine Learning*, Springer, 2006.

[2] A McCallum, D Freitag, F Pereira, *Maximum entropy Markov models for information extraction and segmentation*, Proc. 17th International Conf. on Machine Learning, 2000.

[3] J Lafferty, A McCallum, F Pereira, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, Proc. 18th International Conf. on Machine Learning, 2001.

[4] JA Bilmes, *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*, International Computer Science Institute, 1998.

[5] Ming Zhou, Heung-yeung Shum, *Generating Chinese language couplets*, 2007.

[6] 易勇, 何中市, 李良炎, 周剑勇, 瞿义玻, 张红兵, *基于语言模型的联语应对研究*, 计算机科学, 2006.

[7] 王力, *王力近体诗格律学*, 山西古籍出版社, 2003.