

基于最大匹配和马尔科夫模型的对联系统*

郑亚斌 曹嘉伟 刘知远 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

{yabin.zheng, dover2cindy, liuliudong}@gmail.com sms@mail.tsinghua.edu.cn

摘要: 对联, 雅称“楹联”, 俗称对子, 它言简意深, 对仗工整, 平仄协调, 是一字一音的汉语语言独特的艺术形式。可以说, 对联艺术是中华民族的文化瑰宝。本文开发了一种基于前向最大匹配和一阶马尔科夫模型的对联系统。首先对用户输入的上联进行前向最大匹配的切分, 进而发现匹配结果的若干候选, 利用一阶马尔科夫模型假设和动态规划算法找到和上联最为匹配的下联, 初步的实验结果表明我们的方法具有一定效果。

关键字: 前向最大匹配; 一阶马尔科夫模型; 动态规划; 对联

Couplet System Based on Maximum Matching and Markov Model

Zheng Yabin Cao Jiawei Liu Zhiyuan Sun Maosong

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, China

Abstract: Couplet is considered as Chinese nation's cultural treasures. It is always brief and to the point, and also has neat antithesis, harmonious tone. We developed a couplet system based on forward maximum matching and first-order Markov model. First we segment the first line of a couplet on a scroll using FMM, then find matched candidates from the corpus. Using assumption of first order Markov model and dynamic programming technique, we finally get the second line of a couplet that best suits the input. Preliminary experiment shows the effectiveness of our proposed method.

Keyword: Forward Maximum Matching; First-order Markov Model; Dynamic Programming; Couplet

1 前言

对联[1]是由两个工整的对偶语句构成的独立篇章。其基本特征是字数相等, 平仄相对; 词性相近, 句法相似; 语义相关, 语势相当。对联作为一种雅俗共赏的文学体裁和文化现象, 孕育在“骈语”和“律句”之中, 形成在“骈文”和“律诗”之后, 独立在“骈文”和“律诗”之外; 又与“书法艺术”相表里, 发达在“骈文”和“律诗”之上。

对联文字长短不一, 短的仅一、两个字; 长的可达几百字。对联形式多样, 有正对、反对、流水对、联球对、集句对等。但不管何类对联, 使用何种形式, 却又必须具备以下特点:

一要字数相等, 断句一致。除有意空出某字的位置以达到某种效果外, 上下联字数必须相同, 不多不少。

*本文承国家自然科学基金(项目号 60573187)的资助。

二要平仄相合，音调和谐。传统习惯是“仄起平落”，即上联末句尾字用仄声，下联末句尾字用平声。

三要词性相对，位置相同。一般称为“虚对虚，实对实”，即名词对名词，动词对动词，形容词对形容词，数量词对数量词，副词对副词，而且相对的词必须在相同的位置上。

四要内容相关，上下衔接。上下联的含义必须相互衔接，但又不能重覆。

目前，微软亚洲研究院[2,3]已经推出了计算机自动对联系统，首先用户给定上联，然后系统自动提供若干下联供用户选择，用户可以通过交互手段优选字词来生成满意的下联；当确定一副对联后还可以生成若干四字横批供用户参考。目前可处理十字以下的对联，但是不支持嵌字联、拆字联、音韵联。

本文尝试搭建一个简单的对联系统，以期比较满意的给出候选下联。首先，我们借助中文分词中的前向最大匹配[4]算法的思想，对用户输入的上联进行切分；其次，我们找到语料库中和切分结果匹配的候选序列，结合一阶马尔科夫模型[5]假设，同时考虑了纵向的对仗工整和横向的语意连贯通顺，给出概率最大的匹配序列作为下联返回。

本文的组织结构如下：第二节主要介绍我们提出的对联产生算法；第三节介绍我们使用的数据集和初步实验结果；第四节给出了结论和未来工作展望。

2 算法设计

在这一节中我们主要介绍系统的算法设计，主要包括对用户输入上联的预处理，即利用前向最大匹配对上联进行切分，找出语料库中的若干匹配候选。结合一阶马尔科夫模型假设找到这些候选中和上联匹配概率最大的作为最终结果。在这个过程中，我们同时考虑了纵向的对仗工整和横向的语意连贯性。我们将在下文详细介绍两部分的算法。

2.1 前向最大匹配

在中文分词中，前向最大匹配算法是一种简单而有效的分词方法，其主要思想为借助预先定义的词典，每次从当前字开始，从左向右找到和词典匹配的最长词。然后从下一处匹配位置开始重复上述步骤，直到得到最终切分结果。

利用该思想，我们对用户输入的上联进行切分。直观上，如果用户输入的是语料库中某条记录的上联，那么我们直接返回对应的下联即可，这样既提高了效率，又有比较好的结果。需要指出的是，每一次匹配成功，都要把对应下联相应位置的序列作为候选。举例：上联为“朝辞白帝彩云间”，首先查找“朝辞白帝彩云间”，如果语料库中不存在该记录，查找“朝辞白帝彩云”，如果仍然没有找到，再查找“朝辞白帝彩”，如果查找成功，记录下联中和“朝辞白帝彩”对应的序列作为候选，再用“云间”进行类似的匹配；否则继续查找“朝辞白帝”，直到结束。

2.2 一阶马尔科夫模型

利用自然语言处理中的一阶马尔科夫模型，我们可以认为一个字出现的概率只与它前面一个字的情况有关。

假设某个下联表示为： $S = w_1w_2, \dots, w_n$ ，则该下联匹配上的概率为：

$$\begin{aligned}
p(S) &= p(w_1 w_2 w_3, \dots, w_n) \\
&= p(w_1) p(w_2 | w_1) p(w_3 | w_1, w_2) \cdots p(w_n | w_1, w_2, \dots, w_{n-1}) \quad (1) \\
&= p(w_1) p(w_2 | w_1) p(w_3 | w_2) \cdots p(w_n | w_{n-1})
\end{aligned}$$

其中第三个等号利用了一阶马尔科夫模型的假设，即 w_i 出现的概率只与 w_{i-1} 有关。这样候选下联匹配上的概率就能通过该表达式计算，给出概率最大的项即可以作为我们最终下联的结果。

2.3 整体流程

在这里，我们采取了两种方法得到两个候选的下联：1.利用最大匹配算法 2.直接利用单字，结合马尔科夫模型。其中方法 2 可以看作是最大匹配算法退化的情况：即最大匹配的结果是单字的匹配。对于单字有完整意思的上联，利用方法 2 能得到较好的效果；而对于上联中存在多字词的情况，则用方法 1 能得到更好的下联匹配。

假设上联的输入是 ABCDEFG，经过最大匹配的算法，得到的结果是 ABC DE FG，相应的候选集合为 $a_i b_i c_i$ ($i=1,2,3,\dots,l$) $d_j e_j$ ($j=1,2,3,\dots,m$) $f_k g_k$ ($k=1,2,3,\dots,n$)。如下图示：

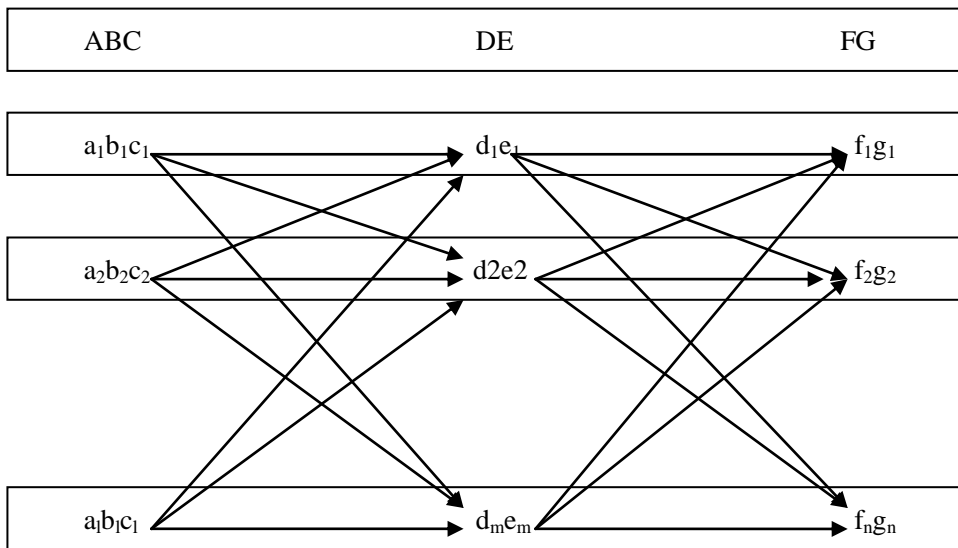


图 1：下联匹配整体流程

其中箭头表示两者之间的状态转移，我们最终的目标就是找出一条最优路径，使得沿着路径的概率值最大。需要指出的是，转移概率必须同时考虑纵向的匹配和横向的意义关联，其中纵向匹配表示与上联的对仗是否工整，横向使得下联语意比较连贯通顺。下面说明我们如何计算横向和纵向的概率：

$$V(d_1e_1 | DE) = \frac{p(d_1e_1, DE)}{p(DE)} \quad (2)$$

其中 $p(d_1e_1, DE)$ 表示对仗匹配上的概率， $p(DE)$ 表示 DE 出现的概率。

$$H(d_1e_1 | a_1b_1c_1) = \frac{p(d_1e_1, a_1b_1c_1)}{p(a_1b_1c_1)} \quad (3)$$

其中 $p(d_1e_1, a_1b_1c_1)$ 表示 d_1e_1 出现在 $a_1b_1c_1$ 后的概率， $p(a_1b_1c_1)$ 表示 $a_1b_1c_1$ 出现的概率。特别的，对于第一列的情况， $H(a_1b_1c_1) = p(a_1b_1c_1)$ ，那么，从 $a_1b_1c_1$ 到 d_1e_1 的转移概率就可以用下式计算：

$$t(d_1e_1 | a_1b_1c_1) = \frac{H(d_1e_1 | a_1b_1c_1)V(d_1e_1 | DE)}{H(d_1e_1 | a_1b_1c_1) + V(d_1e_1 | DE)} \quad (4)$$

对于第一列我们有：

$$t(a_1b_1c_1) = \frac{H(a_1b_1c_1)V(a_1b_1c_1 | ABC)}{H(a_1b_1c_1) + V(a_1b_1c_1 | ABC)} \quad (5)$$

其余的转移概率也通过类似的方法计算得到。通过动态规划算法[6]我们可以找到一条概率最大的最优路径，即为候选下联。这样，我们就能比较好的综合横向意义的连贯通顺和纵向对仗的工整，自动给出下联。目标问题转换为如下形式：

$$\underset{\substack{1 \leq i \leq l, 1 \leq j \leq m, \\ 1 \leq k \leq n}}{\text{Max}} t(a_i b_i c_i) t(d_j e_j | a_i b_i c_i) t(f_k g_k | d_j e_j) \quad (6)$$

3 实验及结果分析

3.1 对联语料库

实验中我们用到的对联语料库均从网上下载得到，分为古代的对联语料库和现代的对联语料库，前者主要是全唐诗的内容，从中抽取出现严格对仗的部分进行训练；后者主要从网上搜集的比较现代的对联，为的是对于所有的语料有一个比较全面的覆盖。古代的语料对于单字有比较完整意思的对联，例如：朝辞白帝彩云间，能找到匹配效果较好的下联；而现代的语料能比较好的匹配到多字词，例如：社会主义，新年，除夕等。语料库的大小分别为：古代语料库的对仗工整上下联 40091 条，现代语料库包含 1100 条对联记录。

3.2 实验结果示例

通过实验结果我们发现：对于单字词比较多的上联，利用退化的单字匹配寻找最优路径的效果要优于最大匹配的方法，这种情况多见于早期的古代对联；而对于多字词的情况，利用最大匹配能很好的利用语料库的知识，这种情况多见于现代对联。下表是一些对联的例子，同时也给出了微软亚洲研究院对联系统结果作为对比。

上联	下联 1(最大匹配)	下联 2(退化单字匹配)	下联 3(MSRA)
春风送暖百花开	大地回春残雪在	日月随风一叶落	瑞雪迎新四化出
神州大地传喜讯	万里山河耀春光	楚国诸天不知音	华夏摇篮报佳音
举国上下迎新春	神州大地过佳节	无人间人识旧日	五湖四海报晓福
爆竹声声辞旧岁	梅花朵朵迎新春	桃花气色迎新年	梅花点点庆新春
四面湖山收眼底	中央钟呗纳胸间	三江海月出门前	万家忧乐到心头
孤帆远影碧空尽	来雁清光红不移	远水清光寒不归	野水疏烟明月来
月缺花残人自怜	花明月在山不见	风飘叶落日不见	山长水远天不见

表 1: 对联结果示例

从上述的例子我们可以发现：对于一些多字词，例如：举国，爆竹，喜讯等，采用最大匹配的方法能找到较好的下联；而对于一些单字词，例如：床，月，花，尽等，采用退化的单字匹配能得到更好的结果。

4 结论及未来工作

本文提出了一种基于前向最大匹配和马尔科夫模型的对联匹配算法，首先利用最大匹配思想对上联进行切分，找到候选序列，继而根据一阶马尔科夫模型假设，找到和上联匹配概率最大的下联作为结果。初步的实验结果表明我们的方法具有一定的效果，通过对实验结果的观察我们发现，对于单字词较多的上联，采用退化的单字匹配有较好的效果；对于多字词较多的上联，采用最大匹配有较好的效果。

目前，我们的系统还存在许多需要改进的地方：首先，我们的系统对于平仄没有做进一步的约束和考虑，传统习惯是“仄起平落”；其次，没有对上联的格式进行更细的分析，例如是否存在叠字，回文等现象；再次，对于字形上的分析和考虑，例如：“鸿是江边鸟”等特殊形势的对联，没有这方面的处理。最后，如何客观地评价对联系统也是一个值得商榷的问题。

参考文献

- [1] 维基百科 <http://wikipedia.org/>
- [2] 微软亚洲研究院 电脑对联 <http://duilian.msra.cn/>
- [3] Ming Zhou and Heung-yeung Shum, Generating Chinese language couplets, United States Patent 20070005345.
- [4] 黄昌宁, 赵海. 中文分词十年回顾, 中文信息学报, 2007, 21(3), pp 8-19.
- [5] Christopher D. Manning, Hinrich Schütze. Foundations of Statistical Natural Language Processing, MIT Press, 1999
- [6] 徐士良. 计算机常用算法（第二版），清华大学出版社, 1995