面向汉语建模的自适应词表生成算法

肖镜辉 [1] 刘秉权 [1] 王晓龙 [1]

摘 要 词表的质量直接影响汉语语言模型的性能, 而当前汉语词典编撰工作同语言建模工作相脱离, 一方面使得现有的汉语语言模型受词表规模所限, 性能不能发挥到最优, 另一方面因为缺乏专业领域的词表, 难以建立面向特定领域的语言模型. 本文旨在通过建立优化词表的方式来提高现有汉语语言模型的性能, 并使其自动适应训练语料的领域. 本文首先将词表自动生成工作同汉语语言建模工作相结合, 构建一体化迭代算法框架, 在自动生成优化词表的同时能够获得高性能的汉语语言模型. 在该框架下, 本文提出汉字构词强度的概念来描述汉语的词法信息, 并将其作为词法特征与统计特征相结合, 构造一种基于多特征的汉语词表自动生成算法. 最后, 本文提出两种启发式方法, 自动根据训练语料的特点调整系统中的各项参数, 使系统能够自动适应训练语料的领域. 实验表明, 本文的方法能够在生成高质量词表的同时获得高性能的语言模型, 并且能够有效自动适应训练语料的领域.

关键词 词表自动生成, 语言建模, 汉字构词强度, 自适应
中图分类号 TP391.12

A Self-adaptive Lexicon Construction Algorithm for Chinese Language Modeling

XIAO Jing-Hui[1] LIU Bing-Quan[1] WANG Xiao-Long[1]

Abstract The lexicon quality affects the performance of Chinese language model directly. However, the lexicon compilation is separated from Chinese language modeling, resulting in two severe problems: firstly, the current language models can not achieve the optimal performance due to the limitation of lexicon scale; secondly, it is hard to apply the current language models to special areas due to the absence of lexicon. This paper aims to improve the performance of Chinese language model by constructing optimal lexicon. Meanwhile, it can self-adapt the area of training corpus automatically. Firstly, this paper combines the lexicon compilation with Chinese language modeling and proposes an iterative algorithm framework. Under this framework, it proposes the concept of character lexical significance (CLS) to describe Chinese lexical principle. Together with the statistical features, a multi-feature based algorithm is proposed for Chinese lexicon construction. Finally, it proposes two heuristic rules to adjust the parameters so as to self-adapt the area of training corpus. From the experimental results, it is found that the system can obtain the optimal Chinese lexicon as well as the high-performance Chinese language model. Moreover, the proposed techniques can self-adapt the area of training corpus successfully.

Key words Chinese lexicon construction, language modeling, Chinese lexicon significance, self-adaptive

在中文信息处理领域, 汉语语言模型发挥了重要作用, 已经被广泛地应用在汉语语音识别[1]、光学字符识别[2]、联机手写体汉字识别[3]、机器翻译[4]、中文智能输入[5] 等众多应用领域当中. 词表是汉语语言建模工作的基本资源, 词表的质量直接影响汉语语言模型的性能: 高质量的词表能够有效地覆盖目标领域的词汇, 全面地反映目标领域的知识, 在其基础上可以获得高性能的汉语语言模型; 而劣质词表包含过多的噪声信息, 难以在其上建立高性能的语言模型.

当前用于汉语语言建模的词表主要来自两个方面: 一方面来自于各种语言学资源, 如: 北大词表[6]、同义词词林[7]、Hownet[8] 等; 另一方面则是根据具体的应用领域手工构建的词表. 然而, 语言学资源在构建时, 通常同时兼顾多种中文信息处理任务, 在词条的选择上没有针对汉语建模任务进行优化, 因而在其基础上建立的语言模型的性能受词表规模所限, 并不是最优的; 而且, 语言学资源中收录的绝大部分是通用领域的词汇, 专业领域的词汇很少, 无法利用它们有效地建立面向特定领域的汉语语言模型. 而手工构建词表的方法, 不仅费时费力, 而且可移植性差, 当应用领域改变时, 往往需要重新构建词表.

本文旨在根据具体的应用领域为汉语语言模型自动构建优化词表, 并通过词表的优化来提高现有语言模型的性能. 首先, 本文将词表自动生成工作同汉语语言建模工作相结合, 提出一种词表生成同语

收稿日期 2006-09-12 收修改稿日期 2007-04-26
Received September 12, 2006; in revised form April 26, 2007
国家自然科学基金 (60435020, 60673037), 国家高科技研究发展计划 (863 计划)(2006AA01Z197, 2007AA01Z172), 教育部 – 微软语言语音重点实验室基金 (01307620) 资助
Supported by National Natural Science Foundation of China (60435020, 60673037), National High Technology Research and Development Program of China (863 Program) (2006AA01Z197, 2007AA01Z172), the Fund of MOE – MS Key Laboratory of Natural Language Processing and Speech (01307620)
1. 哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001
1. School of Computer Science and Techniques, Harbin Institute of Technology, Harbin 150001
DOI: 10.3724/SP.J.1004.2008.00040

言建模相结合的一体化迭代算法框架, 在自动生成词表的同时能够建立高性能的汉语语言模型. 然后, 在该框架下, 本文提出汉字构词强度的概念来描述汉语词法知识, 并将其作为汉字的构词特征, 与汉字的统计特征相结合, 构建了一种基于多特征的汉语词表自动生成算法. 最后, 本文提出两种启发式方法自动调整系统中的各项参数, 使系统能够自动适应训练语料的领域, 从而具有较强的自适应性.

本文后续内容按照如下方式进行组织: 第 1 节介绍相关领域的研究工作; 第 2 节介绍词表生成同汉语语言建模相结合的一体化迭代算法框架; 第 3 节介绍在该框架下的具体工作, 包括基于多特征的词表自动生成算法和基于启发式的系统参数调整方法; 第 4 节给出实验和讨论; 第 5 节为本文的结论.

1　相关领域研究工作

Yang[9] 提出通过扩充词表的方法来提高语言模型的性能. Yang 在其研究中完全从统计角度出发, 认为在语料库中出现的任意汉字组合都可以作为词表中的词, 词语抽取的唯一目的是为了提高语言模型的性能. 然而, Yang 的工作没有考虑汉语的构词规律和词语的语义属性, 将很多噪声信息引入到词表当中, 虽然在一定程度上提高了语言模型的性能, 但却严重降低了当前词表的质量. 本文提出汉字构词强度的概念描述汉字的构词信息, 并作为词法特征用于词表自动生成算法中, 能够同时得到高质量的词表和优化的语言模型. 高剑锋[10−11] 将词表优化工作同语言模型的性能相联系, 提出一种迭代的方法自动从语料中抽取新词、优化词表. 然而, 高剑锋采用人工方法调整系统参数, 不仅费时费力, 而且很难移植到新的领域. 本文提出启发式方法自动调整系统参数, 使系统具有自适应性和可移植性.

聂建云[12] 提出汉字构词力的概念来描述汉字的组词能力, 付国宏[13] 把构词力应用到汉语未登录词识别当中. 本文提出的汉字构词强度与构词力相比有两点重要的区别: 首先, 定义的方式不同. 构词力是根据汉字在训练语料中的字频和词频信息定义的, 反映的是语料的信息; 而本文提出的汉字构词强度是根据现有的词典资源统计的, 反映的是汉语的构词信息. 其次, 应用的方式不同. 在计算若干个汉字的组词概率时, 付国宏的方法是计算每个汉字构词力的乘积, 这种方法具有潜在的 “短词优先” 原则, 从而容易忽略合理的长词. 本文采用各个汉字构词强度的几何平均值, 能够避免上述问题.

2　一体化迭代算法框架

当前, 汉语词典编纂工作同语言建模工作相脱离, 使得在现有汉语词表上建立的语言模型, 其性能受到词表规模和领域所限, 不能得到最优化. 本文面向汉语语言建模这一特定任务, 根据训练语料的特点, 自动构建适合汉语建模的词表, 从而在其基础上获得高性能的汉语语言模型. 首先, 本文将词表自动生成工作同汉语建模工作相结合, 提出一种一体化的迭代算法框架, 使得在生成优化词表的同时能够获得优化的语言模型. 该框架如图 1 所示.

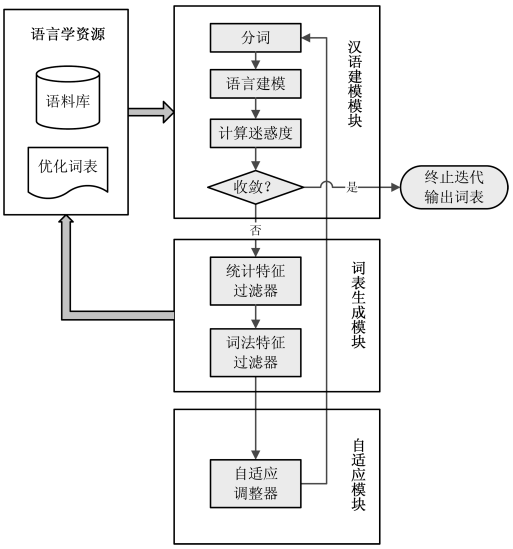


图 1　词表自动生成同语言建模相结合的一体化迭代算法框架

Fig. 1　The unified framework combining lexicon construction and language modeling

该一体化迭代算法框架包含了三个主要模块: 汉语建模模块、词表自动生成模块和系统自适应模块. 汉语建模模块的主要任务是在当前优化词表的基础上建立汉语语言模型, 其中又细分为训练语料分词、语言建模和模型评价 (用迷惑度评价) 三个具体的步骤. 鉴于分词和语言建模领域已经有成熟的技术和方法, 该模块可以根据实际需求采用不同的分词算法和语言建模技术. 词表自动生成模块的主要任务是根据汉语建模模块得到的统计信息, 分别提取汉字串的统计特征和词法特征, 并利用这些特征在候选词空间中对候选词条进行筛选, 将筛选的结果作为新一次迭代的优化词表. 系统自适应模块利用前面两个模块产生的统计信息, 根据启发式规则自动调整系统中的各项参数, 使其更加符合训练语料的特点, 从而使系统能够自动适应训练语料的领域.

具体地说, 该一体化迭代算法框架包含了以下几个步骤:

1) 选择初始词表. 如果目标领域的某些词汇已知, 则可以作为先验知识加入到初始词表中; 否则, 可以仅仅选用常用汉字作为初始词表, 由系统自动

生成优化词表.

2) 在当前优化词表的基础上, 根据实际需求, 选择合适的分词算法对训练语料进行分词.

3) 根据分词结果, 选择合适的语言建模技术建立汉语语言模型.

4) 计算语言模型的迷惑度, 评价语言模型的性能. 同上次迭代相比, 如果迷惑度变化不大, 说明词表的调整对当前语言模型的性能已经影响不大, 则跳出迭代并输出优化词表和当前优化的语言模型; 否则, 说明仍然可以通过词表的优化来提高当前语言模型的性能, 则继续迭代.

5) 根据语言建模时得到的统计信息构建统计量, 作为统计特征在候选词空间中对候选词条进行过滤.

6) 根据词法特征对候选词空间作进一步过滤, 并将过滤结果作为下一次迭代的优化词表.

7) 根据启发式规则修改系统中的各项参数, 并转到 2).

3 自适应的词表生成方法

本节将详细介绍一体化迭代算法框架中的三个主要模块. 鉴于分词和语言建模领域都有成熟的技术和方法, 而且如上文所述, 在一体化迭代算法框架中可以根据实际需求采用不同的分词和语言建模技术, 因此本文不对语言建模模块作介绍.

3.1 基于多特征的词表自动生成方法

前人在工作中通常根据各种统计特征来进行词条抽取[9−14]. 然而, 统计特征仅仅反映汉字串在语料库中的概率分布情况, 无法从汉语构词的角度来描述汉字本身的组词能力. 仅使用统计特征的词语抽取算法, 往往抽取出很多在统计意义上稳定出现, 但不符合词法、语义规律的词条. 例如, 当处理新闻领域语料时, “本报记者李” 通常作为一个词条被算法自动抽取出来. 这是因为 “本报记者” 在新闻语料中非常频繁地出现, 而 “李” 是人名常见姓, 李姓记者很多, “本报记者李” 在概率分布意义上是一个稳定出现的汉字串, 因此被算法自动抽取出来. 但是, 该汉字串显然不符合汉语的构词规则, 也不具有完整的语义属性, 通常不能够认为是一个汉语词条. 综上所述, 仅采用统计特征的算法在词表生成过程中会产生很多噪声信息, 不能够得到高质量的汉语词表. 本文提出汉字构词强度的概念作为汉字的词法特征, 与统计特征相结合, 共同抽取汉语词条. 下面分别对这两类特征作详细介绍.

3.1.1 汉字串统计特征

汉字串的统计特征反映了汉字串在语料中的概率分布情况, 本文构造了两种统计特征, 分别是汉字串的相对频率 (Relevent frequence, RF) 和汉字串之间的互信息 (Mutual information, MI).

对于一串汉字 $s = c_1c_2\cdots c_n$, 其 RF 定义为

$$\mathrm{RF}(s) = \frac{C(s)}{C_{Total}} \tag{1}$$

其中, $C(s)$ 是汉字串 s 在训练语料中出现的次数, C_{Total} 是训练语料的总词次. 将 RF 作为统计特征可以使算法过滤一些低频的候选汉字串, 从而提高算法的执行效率.

MI 是信息论中的概念, 用来衡量两个概率分布共同包含的信息量的大小. 在自然语言处理领域中, MI 通常被用来判断两个语言单位具有搭配关系的可能性. MI 越大, 语言单位之间越可能具有搭配关系: 在词表自动生成任务中, 则表示两个汉字串越有可能组合成一个词; 反之, 表明语言单位之间越不可能具有搭配关系, 两个汉字串也越不可能组成词. 形式化地说, 对于两个汉字串 s_1 和 s_2, 它们之间的 MI 定义如下

$$\mathrm{MI}\left(s_1, s_2\right) = \log \frac{p\left(s_1, s_2\right)}{p\left(s_1\right) \times p\left(s_2\right)} \tag{2}$$

其中, $p(s_1)$ 和 $p(s_2)$ 分别是汉字串 s_1 和 s_2 在训练语料中出现的概率, $p(s_1, s_2)$ 是 s_1 和 s_2 的联合分布概率. 在词表自动生成过程中, 算法分别计算每两个汉字串之间的 MI, 只有 MI 大于某个阈值, 这两个汉字串才被认为有可能组合成词, 其组合被收录到当前优化词表中.

3.1.2 汉字构词特征

本文提出汉字构词强度的概念作为汉字的构词特征. 汉字构词强度是描述一个汉字组成词语的能力的通用性度量, 是汉语词法信息的量化表示, 反映了汉语构词法的一般性规律. 采用汉字构词强度, 能够从汉语构词法的角度过滤一些统计特征无法过滤掉的噪声词条. 因其通用性, 汉字构词强度需要在现有通用词典资源上统计得到, 而不是从训练语料中获得. 根据汉字在词语中的不同位置, 可以将词语分为三个组成部分: 词首、词中和词尾. 词首是词语的第一个字, 词尾是最后一个字, 词首和词尾之间的字统称为词中. 相应地, 一个汉字的构词强度也分为词首、词中和词尾三种构词强度, 分别表示汉字在词语中不同位置的组词能力.

对于汉字 c, 其词首构词强度定义为

$$f_{head}(c) = \frac{\sum\limits_{s} Head\left(c, s\right)}{L} \tag{3}$$

其中, L 是现有词典的词条总数, $Head\left(c, s\right)$ 是二值

函数, 定义为

$$Head(c,s)=\begin{cases}1 & \text{如果 } c \text{ 是 } s \text{ 的词首}\\ 0 & \text{否则}\end{cases} \tag{4}$$

汉字 c 的词首构词强度定义为词典中所有以 c 作为词首的词条数占整个词典词条总数的比重. 类似地, 可以定义 c 的词中和词尾构词强度. c 的词中构词强度定义为

$$f_{middle}(c)=\frac{\sum_s Middle(c,s)}{L} \tag{5}$$

其中

$$Middle(c,s)=\begin{cases}1 & \text{如果 } c \text{ 是 } s \text{ 的词中}\\ 0 & \text{否则}\end{cases} \tag{6}$$

c 的词尾构词强度定义为

$$f_{tail}(c)=\frac{\sum_s Tail(c,s)}{L} \tag{7}$$

其中

$$Tail(c,s)=\begin{cases}1 & \text{如果 } c \text{ 是 } s \text{ 的词尾}\\ 0 & \text{否则}\end{cases} \tag{8}$$

以上三种构词强度分别描述了汉字 c 在词语中不同位置的组词能力, 是对汉语构词法的一种量化逼近. 在对候选词条进行过滤时, 这三种构词强度可以作为独立的统计量单独使用, 也可以联合起来, 计算若干个汉字组成词语的概率, 该概率称为汉字串的构词强度.

对于长度为 n 的汉子串 $s=c_1c_2\cdots c_n$, 其构词强度可以由下式计算

$$f_{string}(s)=\sqrt[3]{f_{head}(c_1)\times f_{tail}(c_n)\times \min(f_{middle}(c_i))} \tag{9}$$

付国宏在其工作中, 定义汉字串的组词概率为各个汉字的构词力 (Word formation power, WFP) 的乘积. 但是乘积项越多, 汉字串的组词概率就越小, 不利于长词的发现, 从而造成系统在选择词条的时候采用 “短词优先” 规则, 使得很多合理的长词被忽略. 本文定义汉字串的构词强度为三种汉字构词强度的几何平均值, 而不是乘积, 避免了概率乘积所带来的 “短词优先” 规则. 并且, 本文在定义中取各项 “词中构词强度” 的最小值, 而不是取所有 “词中构词强度” 的信息, 这是因为根据算法实际运行情况来看, 这项信息对算法的执行效果影响更大, 在定义时着意突出这一项; 这种定义方式同时也提高了算法的执行效率.

当 $n=2$ 时, 两字词的构词强度定义为

$$f_{string}(s)=\sqrt{f_{head}(c_1)\times f_{tail}(c_2)} \tag{10}$$

3.2 基于启发式的参数自动寻优方法

在上文基于多特征的词表自动生成算法中, 我们采用了汉字串的统计特征和汉字的词法特征, 在每种特征中又构建了若干个统计量. 系统在运行过程中为每个统计量设定了一个阈值, 当候选词条的统计结果超过了某个阈值, 则该词条被认为在该统计量上具有统计显著性, 可能被收录到当前优化词表中; 否则, 该词条将作为噪声信息被过滤掉. 在前人的工作中, 这些阈值参数是凭个人经验设定的, 需要根据训练语料的情况手工进行多次尝试和调整. 这种方法不仅费时费力, 而且当面对新的应用领域时, 需要重新调整系统中的各项参数, 从而造成系统的可移植性差. 为了解决上述问题, 本文提出基于启发式的参数自动寻优方法. 该方法将系统参数寻优过程同一体化迭代算法框架相结合, 根据训练语料的自身特点和算法的迭代进程来自动调整系统中的各项参数. 这种方法一方面能够节省人力物力, 另一方面能够使系统自动适应训练语料的领域, 而且能够加速迭代算法收敛, 提高系统运行效率.

下面给出两条启发式规则, 系统根据这些规则对参数进行调整.

1) 规则一

在第 $i+1$ 次迭代中, 统计量 M 的阈值 θ_M^{i+1} 由下式计算得到

$$\theta_M^{i+1}=\alpha\times E^i(M) \tag{11}$$

其中 $E^i(M)$ 是统计量 M 在系统第 i 次迭代时的平均值, 反映了各个候选词条在统计量 M 上的平均表现情况; α 是阈值调整速率, 在本文中定义为常数.

本条启发式规则将系统每次迭代时统计量 M 的阈值取值同上一次迭代时的均值联系起来, 使得系统阈值与训练语料中词语的概率分布相关, 从而使系统能够根据训练语料的特点自动调整算法的选词策略, 即自动适应训练语料的领域.

在式 (11) 中, α 被定义为常数, 统计量 M 的阈值随着算法迭代匀速地进行调整. 然而, 在算法迭代初期, 初始阈值与优化阈值相差很大, 阈值调整速率应该设置为较大的数值, 以便使算法快速逼近优化解; 在算法迭代后期, 当前阈值已经接近优化阈值, 阈值调整速率应该设置为较小的数值, 避免过分调整造成的震荡现象. 因此, 更为精确的阈值调整方法应该是将阈值调整速率同算法迭代次数相联系.

2) 规则二

$$\theta_M^{i+1} = \frac{\beta}{2^i} \times (E^i(M) - \theta_M^i) + \theta_M^i \tag{12}$$

其中, 统计量 M 的阈值 θ_M^{i+1} 由其在上一次迭代中的阈值 θ_M^i 和一个增量项组成, 该增量项不仅与 M 的均值相关, 而且是迭代次数 i 的函数.

根据式 (12), M 的阈值在每次迭代时的更新幅度为 $E^i(M) - \theta_M^i$, 更新速率为 $\beta/2^i$, 而且该速率随着迭代次数 i 呈指数递减. 随着迭代的不断进行, 统计量 M 的阈值逐渐地接近 M 在语料中的均值 $E^i(M)$, 接近的速率则呈指数递减, 最终算法收敛到 $E^i(M)$. 此时优化词表中不再有新词加入, 语言模型性能也不再变化, 系统跳出迭代, 输出优化的词表和优化的语言模型.

以上两种启发式方法均是经验公式, 无法从理论角度严格证明. 然而, 在实际应用中, 这两种启发式方法均取得了良好的应用效果.

4 实验和讨论

本节介绍实验设计和结果分析. 首先, 给出各项评价指标; 然后, 验证词表生成与语言建模相结合的一体化迭代算法框架; 接下来, 本文考察了基于多特征的词表自动生成算法的性能, 验证了汉字词法特征对词表自动生成算法性能的影响; 最后, 本文在不同语料库上分别作实验, 检验算法的可移植性.

4.1 系统评价指标

词表生成算法的性能通常用词表生成的准确率和召回率来衡量. 词表生成的准确率定义为生成词表中正确的词条数目占整个生成词表的比重[14], 如下式所示

$$P = \frac{\sum_s Correct(s)}{L_{create}} \tag{13}$$

其中, L_{create} 是生成词表的词条总数, $Correct(s)$ 是二值函数, 定义为

$$Correct(s) = \begin{cases} 1 & \text{如果 } s \text{ 是正确的词条} \\ 0 & \text{否则} \end{cases} \tag{14}$$

词表生成的召回率定义为同人工生成的标准词表相比, 算法自动生成的词表与人工词表的交集占整个人工词表的比重[14], 如下式所示

$$R = \frac{|S_{create} \cap S_{golden}|}{L_{golden}} \tag{15}$$

其中, S_{create} 是算法自动生成的词语集合, S_{golden} 是人工抽取的词语集合, L_{golden} 是人工词表的词条总数.

准确率和召回率可以用统一的 F 量度来度量, F 量度定义如下

$$F_{measure} = \frac{1}{\theta\frac{1}{P} + (1-\theta)\frac{1}{R}} \tag{16}$$

其中 θ 是常数, 表示准确率 P 和召回率 R 在 F 量度中所占的比重. 本文取 θ 值为 0.5, 表明两者同等重要.

由于本文旨在通过建立优化词表的方式来提高汉语语言模型的性能, 因此, 语言模型的性能也是评价本文系统的一个重要指标. 语言模型的性能在理论上通常由迷惑度来评价, 迷惑度定义如下

$$PP_c = 2^{-\frac{1}{N_c}\sum_{i=1}^{N_w} \ln p(w_i|h_i)} \tag{17}$$

其中, N_c 是训练语料的总字数, N_w 是训练语料的总词数, $p(w_i|h_i)$ 是语言模型中当前词的条件概率. 根据不同的语言模型, 该条件概率具有不同的概率函数形式和参数估计方法. 与此同时, 本文将语言模型应用到音字转换任务中, 将音字转换系统的错误率作为评价语言模型性能的实践度量.

4.2 一体化迭代算法实验

下面验证词表生成同语言建模相结合的一体化迭代算法框架, 用以考察词表的优化对汉语语言模型性能的影响. 本文采用 500 万字人民日报语料作为训练语料, 用 Bigram 模型作为最终建立的目标汉语语言模型. 为提高算法执行效率, 本实验仅仅提取候选词条的统计特征, 并根据式 (12) 修改系统中各项统计量的阈值参数. 本文用于比较的基准语言模型建立在现有词典资源基础上, 该词典共有 25 000 个词条, 是由多名语言学工作者根据以人民日报为主体的平衡语料库编纂而成. 基准语言模型不参与算法迭代, 模型性能恒定.

首先, 本文采用迷惑度作为评价目标语言模型性能的理论度量, 给出语言模型迷惑度在迭代算法中的变化情况, 实验结果如表 1 和图 2 所示.

表 1 语言模型迷惑度随迭代次数的变化情况

Table 1 Perplexity of language model in iteration

迭代次数	1	2	3	4	5	6	7
基本语言模型	10.09	10.09	10.09	10.09	10.09	10.09	10.09
优化语言模型	21.64	10.79	9.03	8.22	8.11	8.02	8.01
迷惑度降低率 (%)	-114.47	-6.93	10.51	18.53	19.62	20.52	20.61

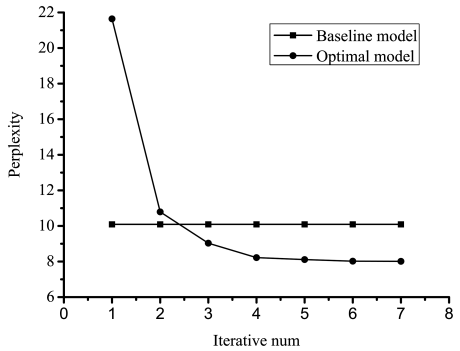


图 2　语言模型迷惑度随迭代次数的变化曲线

Fig. 2　Perplexity of language model in iteration

接下来, 本文将目标语言模型应用到音字转换任务中, 用音字转换错误率作为评价目标语言模型性能的实践度量, 给出错误率随迭代次数的变化情况, 如表 2 和图 3 所示.

表 2　音字转换错误率随迭代次数的变化情况

Table 2　Error rate on PTC system in iteration

迭代次数	1	2	3	4	5	6	7
基本语言模型	9.98	9.98	9.98	9.98	9.98	9.98	9.98
优化语言模型	11.85	10.08	9.96	9.69	9.54	9.44	9.44
错误率降低率 (%)	-18.73	-10.02	0.2	2.91	4.41	5.42	5.42

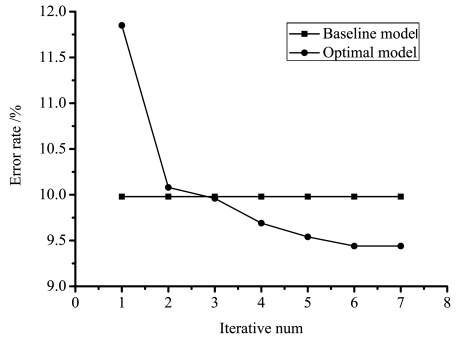


图 3　音字转换错误率随迭代次数的变化曲线

Fig. 3　Error rate on PTC system in iteration

从表 2 和图 3 中可以看到, 当算法刚开始迭代时, 优化词表尚未形成, 系统仅用常用汉字作为初始词表, 在此基础上建立的语言模型的迷惑度和音字转换错误率均高于基准语言模型, 这说明此时目标语言模型的性能远低于基准语言模型的性能. 随着算法不断迭代, 优化词条被不断抽取出来, 优化词表被逐渐建立起来, 在此基础上建立的目标语言模型的迷惑度和音字转换错误率也迅速降低, 并低于基准语言模型, 这表明目标语言模型的性能迅速提高, 超过了基准语言模型的性能. 最终, 算法收敛, 优化词表被完全建立起来, 同时在此基础上得到了高性能的目标语言模型. 同基准语言模型相比, 目标语言模型的迷惑度降幅达到了 20.61 %, 音字转换错误率下降了 5.42 %, 表明目标语言模型的性能获得了大幅提升.

以上实验结果表明, 在现有语言建模技术的基础上, 通过优化词表的方式能够明显提高汉语语言模型的性能; 本文提出的词表生成同语言建模相结合的一体化迭代算法能够有效地将词表生成工作同语言建模工作相结合, 在得到优化词表的同时能够得到优化的语言模型.

4.3　多特征词表生成算法实验

本节用实验验证基于多特征的词表自动生成算法, 主要考察汉字构词强度对词表生成算法性能的影响. 汉字构词强度是根据上节中人工构建的词典资源统计得到的, 并且该词典还被用于计算词表生成算法的召回率. 本实验采用的训练语料也与上节实验相同. 结果如表 3 所示.

表 3　汉语词法特征对词表自动生成算法性能的影响

Table 3　Influence of lexical features

序号	特征采集	准确率 (%)	召回率 (%)	F 量度	F 量度提高幅度 (%)
1	$MI + RF^{\star}$	85.29	53.05	0.6541	0
2	$1 + f_{head}$	89.74	54.58	0.6788	3.78
3	$1 + f_{middle}$	86.35	55.29	0.6741	3.07
4	$1 + f_{tail}$	88.11	54.7	0.675	3.2
5	$2 + f_{tail}$	89.76	55.59	0.6866	4.97
6	$5 + f_{middle}$	89.79	57.27	0.6993	6.91
7	$6 + f_{string}$	89.84	57.28	0.6996	6.96

在表 3 中, 算法 1 作为基准算法, 在词表自动生成过程中仅仅采用了统计特征 (MI 和 RF). 算法 2～7 分别采用了各种汉字构词强度及其组合作为词法特征, 与统计特征相结合, 进行词表自动生成工作. 从表 3 中的数据可以看到, 在加入词法特征后, 词表生成算法的准确率和召回率均有明显的提高; 并且, 随着加入的词法特征逐渐增多, 算法的性能呈逐渐上升的趋势. 查看最终生成的优化词表, 在加入词法特征后, 类似 “本报记者李” 这样不符合汉语构词法的噪声词条被有效过滤掉, 而 “本报记者” 被正确地抽取到优化词表当中. 总之, 本文提出汉字强度作为词法特征, 能够有效地过滤词表自动生成过程中产生的噪声信息, 显著地提高词表自动生成算法的性能.

4.4　系统自适应实验

下面检验启发式方法对系统性能的影响. 本节实验按如下方式进行组织: 首先, 检查启发式方法的

收敛性; 然后, 在相同语料上, 比较启发式方法同人工确定阈值方法对系统性能的影响; 最后, 将启发式方法应用到旅游领域的语料上, 检验启发式方法对系统自适应性的作用.

本文提出的启发式方法, 作为一种经验式参数调整方法, 首先必须使系统参数能够有效地收敛. 以人民日报语料为训练语料, 以统计量 MI 的阈值参数为例, 根据两种启发式方法, 随着算法的不断迭代, 其取值情况如图 4 所示.

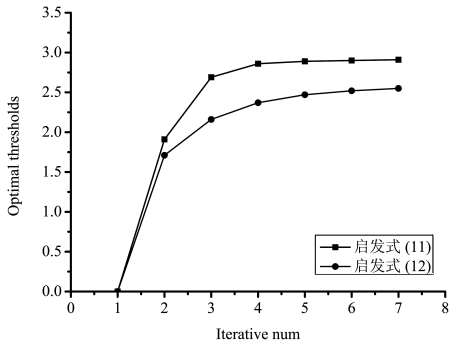


图 4 MI 阈值随算法迭代次数的变化曲线

Fig. 4 Thresholds of MI in iteration

从图 4 中的曲线可以看到, 随着算法迭代的不断进行, 统计量 MI 的阈值根据启发式规则不断地作出调整, 最终随算法收敛到某个优化阈值. 以上结果表明, 本文提出的启发式方法能够有效地收敛.

本文提出的启发式方法, 作为一种系统参数寻优策略, 其性能应该与人工确定参数方法的性能相当, 或者至少具有可比性. 否则, 对于一些要求高性能的应用领域, 人们为了追求系统的高性能, 仍会采用人工方法确定系统参数, 而不采用启发式自动寻优. 在上文实验的基础上, 本文用人工方法替换上面系统中的启发式方法来确定系统参数, 并与应用启发式方法的系统相比较, 从而比较启发式方法和人工方法在处理相同领域语料时对系统性能的影响. 实验结果如表 4 所示.

表 4 系统自适应实验 — 人民日报语料

Table 4 Experiments on *People's Daily* corpus

序号	参数确定方式	词表生成准确率 (%)	词表生成召回率 (%)	F 量度	语言模型迷惑度
1	人工确定	84.74	52.56	0.6488	8.04
2	启发式 (11)	87.97	55.34	0.6794	8.05
3	启发式 (12)	85.29	53.05	0.6541	8.01

从表 4 中的数据可以看到, 与人工确定参数的系统相比, 应用启发式方法的系统在词表自动生成方面获得了较高的准确率和召回率; 同时, 在此基础上建立的语言模型的性能也与人工方法相当. 这说明, 在处理相同领域的语料时, 同人工确定参数的系统相比, 应用启发式方法的系统能够获得与之相当甚至更高的性能. 然而, 启发式方法是自动化方法, 没有人工确定参数时的尝试和调整, 更加省时省力. 以上表明, 即使不考虑系统的可移植性问题, 在构建系统时仍然可以采用启发式方法代替人工方法来确定系统中的各项参数.

下面考察启发式方法对系统自适应性能的影响. 将上述实验系统应用在旅游领域的语料上 (训练语料是从互联网上几个著名旅游网站上下载的文章并加以整理后得到的, 共 450 万字), 对于人工确定参数的系统, 仍保持其在上文人民日报语料实验中得到的优化参数. 实验结果如表 5 所示.

表 5 系统自适应实验 — 旅游领域语料

Table 5 Experiments on tour corpus

序号	参数确定方式	词表生成准确率 (%)	词表生成召回率 (%)	F 量度	F 量度提高幅度 (%)	语言模型迷惑度	迷惑度降低率 (%)
1	人工确定	59.85	46.9	0.5259	—	19.86	—
2	启发式 (11)	71.1	60.35	0.6529	24.15	9.73	51.01
3	启发式 (12)	69.7	57.1	0.6277	19.36	9.6	51.66

从表 5 中的数据可以看到, 应用启发式方法的系统的性能要明显优于人工确定参数的系统的性能. 在词表自动生成方面, 应用启发式的系统的 F 量度最多提高了 24.15 %, 说明生成词表的质量有大幅提高; 在此基础上建立的目标语言模型迷惑度最多降低了 51.66 %, 说明获得了高性能的语言模型. 以上结果表明, 本文提出的启发式方法能够根据训练语料的特点自动调整系统参数, 从而使系统有效地适应训练语料的领域, 从而具有自适应特性.

综上所述, 与人工方法相比, 本文提出的启发式方法能够自动优化系统中的各项参数; 在处理相同领域的语料时, 启发式方法能够获得与人工方法相当、甚至更好的系统性能; 在面对陌生领域的语料时, 启发式方法能够使系统自动适应训练语料的领域, 从而使系统具有自适应性和可移植性.

5 结论

词表的质量直接影响汉语语言模型的性能, 而当前汉语词典资源建设同汉语语言建模工作相脱离, 一方面使得现有的语言模型受词表规模所限, 性能不能发挥到最优, 另一方面因为缺乏专业领域的词表, 难以建立面向特定领域的语言模型. 为了解决上述问题, 本文首先将词表自动生成工作同汉语语言建模工作相结合, 构建词表生成同语言建模相结合的一体化迭代算法框架, 在自动获得优化词表的同

时能够建立高性能的汉语语言模型. 然后, 本文提出汉字构词强度的概念用以描述汉字的构词信息, 并将其作为词法特征与统计特征相结合, 提出一种基于多特征的词表自动生成算法. 最后, 本文提出两种启发式方法自动调整系统中各项参数, 使系统能够自动适应训练语料的领域, 从而具有自适应性和可移植性. 通过实验, 得出如下结论:

1) 本文构建的一体化迭代算法框架, 能够有效地将词表自动生成工作同汉语语言建模工作结合起来, 在自动生成优化词表的同时建立高性能的汉语语言模型.

2) 本文提出的汉字构词强度的概念能够有效地提高词表自动生成算法的性能, 从而提高生成词表的质量.

3) 本文提出的启发式规则能够根据训练语料的情况自动优化系统中的各项参数, 在处理相同领域的语料时, 能够获得与人工方法相当、甚至更好的效果; 在处理陌生领域语料时, 能够使系统具有很强的自适应性和可移植性.

References

1 Wang Xuan, Wang Xiao-Long, Zhang Kai. Language model for speech recognition applications. *Acta Automatica Sinica*, 1999, **25**(3): 309−315
(王轩, 王晓龙, 张凯. 语音识别中统计与规则结合的语言模型的研究. 自动化学报, 1999, **25**(3): 309−315)

2 Wang Kai, Wang Qing-Ren. Research on Chinese/English mixed document recognition. *Journal of Software*, 2005, **16**(5): 786−798
(王恺, 王庆人. 中英文混合文章识别问题. 软件学报, 2005, **16**(5): 786−798)

3 Zhao Wei, Liu Jia-Feng, Tang Xiang-Long. An on-line free handwritten Chinese character recognition method based on component cascaded HMMs. *Journal of Harbin Institute of Technology*, 2004, **36**(5): 570−573
(赵巍, 刘家锋, 唐降龙. 基于部件 HMM 级联的联机手写体汉字识别方法. 哈尔滨工业大学学报, 2004, **36**(5): 570−573)

4 Liu Qun. Survey on statistical machine translation. *Journal of Chinese Information Processing*, 2003, **17**(4): 1−12
(刘群. 统计机器翻译综述. 中文信息学报, 2003, **17**(4): 1−12)

5 Xiao J H, Liu B Q, Wang X L. Principles of non-stationary hidden Markov model and its applications on sequence labeling task. *Lecture Notes in Computer Science*, 2005, **3651**: 827−837

6 Yu Shi-Wen, Zhu Xue-Feng, Wang Hui, Zhang Yun-Yun. *The Grammatical Knowledge-base of Contemporary Chinese (First Edition)*. Beijing: Tsinghua University Press, 1998
(俞士汶, 朱学锋, 王惠, 张芸芸. 现代汉语语法信息词典详解 (第一版). 北京: 清华大学出版社, 1998)

7 Mei Jia-Ju, Zhu Yi-Ming, Gao Yun-Qi, Yin Hong-Xiang. *Tong Yi Ci Ci Lin*. Shanghai: Shanghai Lexicographical Publishing House, 1983
(梅家驹, 竺一鸣, 高蕴琦, 殷鸿翔. 同义词词林. 上海: 上海辞书出版社, 1983)

8 Dong Z D, Dong Q. *How Net and the Computation of Meaning*. World Scientific Publishing, 2006

9 Yang K C, Ho T H, Chien L F, Lee L S. Statistics-based segment pattern lexicon – a new firection for Chinese language modeling. In: Proceedings of International Conference on Acoustic, Speech and Signal Processing. Seattle, WA: 1998. 172−172

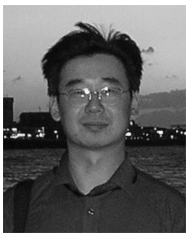
10 Zhao J, Gao J F, Chang E, Li M J. Lexicon optimization for Chinese language modeling. In: Proceedings of International Symposium Conference on Spoken Language Processing. Beijing, China: IEEE, 2000

11 Gao J F, Goodman J, Li M J, Lee K F. Toward a unified approach to statistical language modeling for Chinese. *ACM Transactions on Asian Language Information Processing*, 2002, **1**(1): 3−33

12 Nie J Y, Hannan M L, Jin W Y. Unknown word detection and segmentation of Chinese using statistical and heuristic knowledge. *Communications of COLIPS*, 1995, **5**(1-2): 47−57

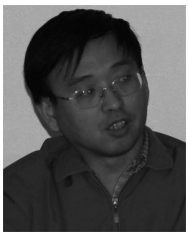
13 Fu G H, Wang X L. Unsupervised Chinese word segmentation and unknown word identification. In: Proceedings of the 5th Natural Language Processing Pacific Rim Symposium. Beijing, China: 1999. 32−37

14 Fung P. Extracting key terms from Chinese and Japanese texts. *Computer Processing of Oriental Languages*, 1998, **12**(1): 99−121



肖镜辉 哈尔滨工业大学计算机科学与技术学院博士研究生. 主要研究方向为汉语语言建模, 汉字键盘输入. 本文通信作者.
E-mail: xiaojinghui@insun.hit.edu.cn
(**XIAO Jing-Hui** Ph. D. candidate at Harbin Institute of Technology. His research interest covers language modeling and Chinese keyboard input method. Corresponding author of this paper.)

刘秉权 哈尔滨工业大学计算机科学与技术学院副教授. 主要研究方向为自然语言处理, 移动 Web 计算.
E-mail: liubq@insun.hit.edu.cn
(**LIU Bing-Quan** Associate professor at Harbin Institute of Technology. His research interest covers natural language processing and mobile web computing.)



王晓龙 哈尔滨工业大学计算机科学与技术学院教授. 主要研究方向为网络信息处理, 声图文智能计算.
E-mail: wangxl@insun.hit.edu.cn
(**WANG Xiao-Long** Professor at Harbin Institute of Technology. His research interest covers web information processing and intelligent computing.)