ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.2, February 2007, pp.196−204
DOI: 10.1360/jos180196

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

语言建模中最小化样本风险算法的研究和改进*

袁 伟[1+], 高剑峰[2], 步丰林[1]

[1](上海交通大学 计算机科学与工程系,上海 200230)

[2](Natural Language Processing Group, Microsoft Research, Redmond 98052,USA)

A Study and Improvement of Minimum Sample Risk Methods for Language Modeling

YUAN Wei[1+], GAO Jian-Feng[2], BU Feng-Lin[1]

[1](Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai 200230, China)

[2](Natural Language Processing Group, Microsoft Research, Redmond 98052, USA)

+ Corresponding author: Phn: +86-21-64836410, E-mail: weiyuan1981@gmail.com

Yuan W, Gao JF, Bu FL. A study and improvement of minimum sample risk methods for language modeling. *Journal of Software*, 2007,18(2):196−204. http://www.jos.org.cn/1000-9825/18/196.htm

Abstract: Most existing discriminative training methods adopt smooth loss functions that could be optimized directly. In natural language processing (NLP), however, many applications adopt evaluation metrics taking a form as a step function, such as character error rate (CER). To address the problem, a newly-proposed discriminative training method is analyzed, which is called minimum sample risk (MSR). Unlike other discriminative methods, MSR directly takes a step function as its loss function. MSR is firstly analyzed and improved in time/space complexity. Then an improved version MSR-II is proposed, which makes the computation of interference in the step of feature selection more stable. In addition, experiments on domain adaptation are conducted to investigate the robustness of MSR-II. Evaluations on the task of Japanese text input show that: (1) MSR/MSR-II significantly outperforms a traditional trigram model, reducing CER by 20.9%; (2) MSR/MSR-II is comparable to the other two state-of-the-art discriminative algorithms, Boosting and Perceptron; (3) MSR-II outperforms MSR not only in time/space complexity but also in the stability of feature selection; (4) Experimental results of domain adaptation show the robustness of MSR-II. In all, MSR/MSR-II is a quite effective algorithm. Given its step loss function, MSR/MSR-II could be widely applied to many fields of NLP, such as spelling check and machine translation.

Key words: language modeling; discriminative training method; input method editor; minimum sample risk; domain adaptation modeling

摘 要: 目前,一些主流的判别学习算法只能优化光滑可导的损失函数,但在自然语言处理(natural language processing,简称 NLP)中,很多应用的直接评价标准(如字符转换错误数(character error rate,简称 CER))都是不可导的阶梯形函数.为解决此问题,研究了一种新提出的判别学习算法——最小化样本风险(minimum sample risk,简称 MSR)算法.与其他判别训练算法不同,MSR 算法直接使用阶梯形函数作为其损失函数.首先,对 MSR 算法的时空复杂性作了分析和提高;同时,提出了改进的算法 MSR-II,使得特征之间相关性的计算更加稳定.此外,还通过大量领域适应性建模实验来考察 MSR-II 的鲁棒性.日文汉字输入实验的评测结果表明:(1) MSR/MSR-II 显著优

* Received 2006-01-04; Accepted 2006-06-12

于传统三元模型,使错误率下降了 20.9%;(2) MSR/MSR-II 与另两类主流判别学习算法 Boosting 和 Perceptron 表现相当;(3) MSR-II 不仅在时空复杂度上优于 MSR,特征选择的稳定性也更高;(4) 领域适应性建模的结果证明了 MSR-II 的良好鲁棒性.总之,MSR/MSR-II 是一种非常有效的算法.由于其使用的是阶梯形的损失函数,因此可以广泛应用于自然语言处理的各个领域,如拼写校正和机器翻译.

关键词: 语言建模;判别训练算法;输入法编辑器;最小化样本风险;领域适应性建模
中图法分类号: TP391 文献标识码: A

统计语言模型是自然语言处理的重要组成部分,已经被广泛应用于众多相关领域中,如自动语音识别(automatic speech recognition,简称 ASR)[1]、统计机器翻译(statistical machine translation,简称 SMT)[2]和输入法编辑器(input method editor,简称 IME)[3].传统生成模型认为数据都是由某种潜在分布生成的,并试图为这种分布建模.如常用的 N 元模型就把语言潜在的分布视为多项式分布,采用最大似然估计(maximum likelihood estimation,简称 MLE)来求解模型参数,并用平滑算法来解决数据稀疏问题.这种方法仅当以下两个条件都满足时才是最优的:(1) 数据的概率分布形式是已知的;(2) 存在无穷多的训练数据.但在实际应用中,这两个条件根本无法满足.

判别学习算法是与生成模型相对应的一类建模方法.其假设条件比 MLE 弱得多,只要求训练数据和测试数据来自同一个分布即可.而且,判别学习算法的目标往往与实际应用的评价标准密切相关(如使得模型在训练数据上的错误率最小化).从这个角度来看,判别算法要优于生成模型.其问题在于:实际应用的评价标准(如一个有限训练集合上的错误率)通常是模型参数的阶梯形函数,难以使用梯度下降等算法直接优化.因此,目前主流的判别学习算法往往使用某种近似于真正错误率且易于优化的损失函数,如 Boosting[4]和 Perceptron[5].尽管这些算法有着理论上的完备性,如关于收敛性和泛化错误率上限的证明,但是,近似损失函数的引入毕竟偏离了真正的损失函数.针对这个问题,高剑峰等人[6]提出了一种新型的判别学习算法——最小化样本风险(minimum sample risk,简称 MSR)算法.与以往的判别算法不同,MSR 采用启发式的贪心算法直接优化一个阶梯形损失函数(如训练样本上的错误率).在优化过程中,MSR 从所有候选特征中依次寻找能够使样本风险下降最多的特征;每找到一个,就利用线性搜索调整其参数,使得样本风险最小化.

本文是对文献[6]的延伸及拓展.首先,本文从算法和实现角度对 MSR 算法的时空复杂性进行了分析和改进,进一步提升了 MSR 的效率;其次,本文提出的 MSR-II 通过动态计算特征相关度减轻了 MSR 中特征选择效果不够稳定的问题;最后,本文把 MSR/MSR-II 应用到日文汉字输入程序中,并通过领域适应性建模实验来研究其泛化性.实验和分析结果表明,MSR 算法是相当高效的,不仅在很大程度上优于传统的三元模型,也与两种主流的判别学习算法 Boosting 和 Perceptron 效果相当.与 MSR 相比,改进后的 MSR-II 不仅在效率上有了进一步提升,而且特征选择的稳定性也大为增强.此外,大量的领域适应性建模实验结果也在一定程度上证明了 MSR-II 具有良好的泛化性.

1 语言模型与日文汉字输入法

本文把语言模型应用到亚洲语言(日文)输入法中.该任务称为输入法编辑器,通过把输入的音频符号(日文假名)转换成相应的字符串来输入日文汉字.IME 的性能用字符转换错误率(character error rate,简称 CER)来衡量,即用转换错误的字符数除以正确转换的字符数.

与自动语音识别相类似,IME 可以被视为一个贝叶斯决策问题.设 A 为用户输入的音频符号串,$GEN(A)$为可能由 A 转换而成的候选字符串集合,$P(W)$为候选字符串 W 的概率,W^*为正确转换结果,则 IME 系统通过解以下等式来选择一个与 A 最相符的单词序列:

$$W^* = \arg\max_{W \in GEN(A)} P(W \mid A) = \arg\max_{W \in GEN(A)} \frac{P(W,A)}{P(A)} = \arg\max_{W \in GEN(A)} P(W)P(A \mid W) \quad (1)$$

与自动语音识别任务不同的是,因为这里的音频符号是由用户直接输入的,没有声学上的模糊性,且日文假

名到日文汉字的转换可以被看作一个一对多的关系,即 $P(A|W)=1$,所以,W^*的选择仅仅依赖于语言模型 $P(W)$,这使得 IME 成为一个比自动语音识别更直接的衡量语言模型性能的方法.使用 IME 的另一个优势在于:对日文来说,从 W 到 A 的转换是很容易的,因此可以从现有的文本语料中获得大量的数据.

传统 N 元模型使用 MLE 来估算等式(1)中的 $P(W)$,即最优的模型参数λ应该使训练数据的似然度 $P(W|\lambda)$最大化.但由于似然度与实际应用的评价标准(如 CER)没有直接联系,其效果不够理想.因此,我们更青睐于判别学习算法,因为其损失函数与实际评价标准相一致,可能会取得更好的效果.

2 最小化样本风险算法

2.1 问题定义

我们所遵循的线性模型框架来自于模式分类中的线性判别函数[7],下面的符号和定义来自于文献[5].

- 训练数据.训练数据**是一个**输入/输出**对的**集合.在 IME 的语言模型中,训练样本表示成$\{A_i, W_i^R\}$,$i=1,\ldots,M$.其中,A_i 是用户输入的一串音频符号,而 W_i^R 则是 A_i 对应的正确转换字符串.
- $GEN(A)$.$GEN(A)$是对应于 A 的一个候选字符串集合.本文的实验用一个基于词的三元模型为每个 A 生成若干个候选转换字符串 W,并只把其中概率最高的若干个保留在 $GEN(A)$中.
- 特征.假设共有 $D+1$ 个特征 $f_d(W)$,$d=0,\ldots,D$.这些特征可以是把 W 映射到实数的任意函数,而且我们希望能够把所有有利于从候选答案中区分出最佳答案的特征全都包括进来.定义向量 $f(W)\in R^{D+1}$,其中,$f(W)=[f_0(W),f_1(W),\ldots,f_D(W)]^T$.不失一般性,把 f_0 称为基本特征,并定义为基础三元模型关于 W 的概率的对数值,即 $f_0(W)=\log P(W)$.而其他的特征 $f_d(W)$,$d=1,\ldots,D$ 则定义为第 d 个 n 元组在 W 中出现的次数,在我们的实验中,$n=1$ 或 $n=2$.
- 模型参数.语言模型的参数是一个 $D+1$ 维的向量$\lambda=\{\lambda_0,\lambda_1,\ldots,\lambda_D\}$,其中,每个参数都对应于相应的特征.于是,某个转换字符串 W 的得分可以写成

$$Score(W,\lambda)=\lambda f(W)=\sum_{d=0}^{D}\lambda_d f_d(W) \tag{2}$$

于是,等式(1)中的决策又可以写成

$$W^*(A,\lambda)=\arg\max_{W\in GEN(A)} Score(W,\lambda) \tag{3}$$

编辑距离函数 $Er(W^R,W)$被用来评价语言模型的效果(本文实验中定义为 W 中转换错误的字符数(与 W^R 相比)).我们把所有训练样本上的 $Er()$总和称为样本风险(sample risk,简称 SR),并通过调整模型的参数使得样本风险最小化,如下式所示:

$$\lambda^*=\arg\min_{\lambda} SR(\lambda)=\arg\min_{\lambda}\sum_{i=1,\ldots,M} Er(W_i^R,W_i(A_i,\lambda)) \tag{4}$$

因此,我们的算法被命名为最小化样本风险算法.

2.2 算法概述

因为 $SR()$是一个不可导的分段阶梯形函数($Er()$的值总是整数),无法直接应用常见的优化算法(如梯度下降),所以,MSR 采用了一种类似于多维优化[8]的方法来优化 $SR()$:初始时,模型中没有任何参数;每次迭代中,在保持模型中现有特征及其参数不变的情况下,把某个最有效(能够使得样本风险 SR 下降最多)的候选特征选入模型,并使用线性搜索算法设置其参数以使样本风险最小化.

因此,MSR 由两个主要部分组成:一是线性搜索;二是特征子集的选择,即如何从海量候选特征中选出一个有效的特征集合加入模型.下面就将对这两个方面分别加以介绍.

2.3 网格线性搜索

MSR 的损失函数 $SR()$是每个训练样本上的损失函数 $Er()$值之和(公式(4)),所以,下面先考虑单个训练样本

(A,W^R)上的线性搜索.设 f_d 是当前被选中的特征,线性搜索就是要寻找一个最优的λ_d,使得把 f_d 加入当前模型之后,该样本上的 $Er()$最小化.我们首先把λ_d 在区间$(-\infty,+\infty)$上遍历,就能得到一个关于λ_d 的区间序列,形如$(-\infty,x_1)$,(x_1,x_2),…,$(x_n,+\infty)$.λ_d 落在任何一个区间内,$GEN(A)$中的某一个候选转换 W 都将得到最高分并被选中,而 $Er(W^R,W)$(即转换错误数)就与这个区间相联系.也就是说,我们知道了当λ_d 落在任何一个区间内时所引起的样本风险.然后遍历这些区间,我们就能够得到最小的 $Er()$值,而相应区间的中值就可以作为λ_d 的最优值.再通过合并所有训练样本的区间序列,就能把线性搜索拓展到整个训练数据集上.

为了使模型更稳定,我们实际使用的是经过平滑的样本风险[9].设λ是某个区间的中值,$SR(\lambda)$是相应区间的样本风险,那么该区间的平滑样本风险定义如下:

$$\int_{\lambda-b}^{\lambda+b} SR(\lambda)\mathrm{d}\lambda \tag{5}$$

其中,b 是一个平滑因子,由实验确定.

2.4 特征选择

在选择特征时,我们首先考虑的是该特征加入模型后能够使样本风险下降的程度,并优先选择那些能够使样本风险下降最多的特征.而且,为了降低计算复杂度,我们引入了候选特征之间的独立性假设,即不同特征的作用不会相互影响,并把一个特征的作用定义为把该特征加入到只含基础特征 f_0 的模型中后能够使样本风险 SR 下降的幅度.设 $SR(f_0)$是仅仅使用 f_0 时的样本风险值,而 $SR(f_0+\lambda_d f_d)$是同时使用特征 f_0 和 f_d 且把λ_d 调整到最优以后的样本风险值.这样,f_d 的作用定义如下:

$$E(f_d)=\frac{SR(f_0)-SR(f_0+\lambda_d f_d)}{\max\limits_{f_i,i=1,\ldots,D}(SR(f_0)-SR(f_0+\lambda_i f_i))} \tag{6}$$

等式中的分母是一个归一化因子.

另一方面,为了减少特征间的相互干扰,我们优先选择与已有特征最不相关的候选特征.特征之间的相关性定义如下:假设 $x_{md}(m=1,\ldots,M \text{ and } d=1,\ldots,D)$是一个布尔值.如果加入特征 f_d 后,第 m 个训练样本上的风险下降了,并保持不变,那么 $x_{md}=1$;否则,$x_{md}=0$.于是,两个特征 f_i,f_j 之间的相关系数如下:

$$C(i,j)=\frac{\sum_{m=1}^{M}x_{mi}x_{mj}}{\sqrt{\sum_{m=1}^{M}x_{mi}^2\sum_{m=1}^{M}x_{mj}^2}} \tag{7}$$

显然,C 是一个 0~1 之间的实数.

2.5 算法框架

MSR 的算法步骤如下(其中,f_i表示任意一个已被选入模型的特征,f_j表示任意一个尚未被选入模型的特征):

步骤 1. 初始时,模型中只包括基础特征 f_0,表示为$\{f_0\}$.对每一个候选特征 $f_d,d=1,\ldots,D$,根据公式(6)计算它的 $E(f_d)$值.然后,根据 E 值对所有的候选特征降序排列.在第 1 次迭代中,选择第 1 个(即 E 值最大)特征,称为 f_1,把它加入模型$\{f_0\}$中,并使用线性搜索设置其最优参数.

步骤 2. 在第 k 次迭代中($k=2,\ldots,L$),把满足下式的特征作为第 k 个特征加入模型,并用线性搜索优化其参数

$$f=\arg\max_{f_j}\left\{\alpha E(f_j)-\frac{1-\alpha}{k-1}\sum_{i=1}^{k-1}C(i,j)\right\} \tag{8}$$

这里的α是一个 0~1 之间的调节因子,用来调节候选特征的作用和它与以前所选特征相关性的相对重要程度,并在实验中确定.也就是说,在选择第 k 个特征时,我们不仅考虑该特征的降低样本风险能力,还考虑它与前 $k-1$ 个被选特征的相关性.显然,两类特征之间的相关性越小,它们降低样本风险的能力就越不会互相抵消.我们优先选择那些既能使样本风险大幅度降低,又不与以前所选特征互相干扰的特征.

2.6 算法效率分析与改进

我们首先分析线性搜索的效率.对一个训练样本(A,W^R)而言,在考虑特征 f_d 的最优参数λ_d 时,$GEN(A)$中的每

个候选转换 W 的得分(等式(2))可以被分解为两项:

$$Score(W,\lambda)=\lambda f(W)=\sum_{d'=0 \text{ or } d'\neq d}^{D}\lambda_{d'}f_{d'}(W)+\lambda_d f_d(W) \tag{9}$$

其中,第 1 项是不受λ_d影响的.如果有几个候选转换具有相同的 $f_d(W)$,那么,它们的相对排序与λ_d无关.在本文的实验中,$f_d(W)$都是整数(即特征 f_d 在 W 中出现的次数),可以根据 $f_d(W)$的值对 $GEN(A)$中的转换候选分组,其中每一组的候选都具有相同的 $f_d(W)$值.在每一组中,把第 1 项值最高的候选转换为该组中的活动候选.因为无论λ_d取什么值,该组中只有这个活动候选才会被选中(得分最高),这样就极大地减少了 $GEN(A)$中需要考虑的转换候选的数量.在本文的实验中,$GEN(A)$的大小为 20,则其中的活动候选转换一般少于 5 个,这就减少了单个样本上的优化复杂度.

此外,因为任何一种特征(Unigram 或 Bigram)都只在训练样本中很小的一个子集中出现,所以,我们对每一个候选特征建立了一个索引,记录了该特征出现过的所有样本,这又减少了在整个训练样本上优化一个特征的时间(即降低了 $E()$的计算时间).

在特征选择阶段,时间复杂性主要集中在对于 $E()$和 $C()$的计算上.回忆一下,D 代表候选特征的总数(即训练样本中出现过的不同的 Unigram 和 Bigram 的数量,可以有几十万个),L 代表最终被选入语言模型中的特征数量(通常不超过 1 万个).在步骤 1 中,需要对每个候选特征计算 $E()$.而在步骤 2 中,总共需要计算 $O(L\times D)$量级的 $C()$,计算耗费非常大.因此,我们作以下假设:即使某个特征与模型中其他特征的相关度很小,但如果它被加入基础模型$\{f_0\}$后样本风险下降很少或者几乎不下降,那么它就一定不会被加入最终的模型.因此,我们在步骤 2 中开了一个窗口 N,并在余下的候选特征中仅仅考虑效果最好的 N 个特征(N 一般小于 1 万).这样,就把步骤 2 的复杂度降到了 $O(L\times N)$量级的 $C()$.进一步地,由于开了窗口 N 以后,候选特征最多只有 $L+N$ 个,可以为每个候选特征保留一个相关性累加变量.每当在模型中新加入一个特征时,每个候选特征与这个新加入特征的相关度就累加到这个变量中去.这样就把步骤 2 的复杂度进一步降到了 $O(L+N)$量级的 $C()$.

此外,前期实验结果还显示 MSR 的特征相关度计算尚不够稳定.仅当公式(8)里的调节因子α落在一个狭小的范围时,引入特征相关度才能带来好处;在其他情况下,不考虑特征相关度(即α=1)的效果反而更好(见第 3.4 节).这可能是因为在公式(8)中计算的虽然是候选特征与当前模型中所有特征的平均相关度,但是这种相关度仅仅是相对于最原始模型$\{f_0\}$的.随着模型的不断变化,这种相关性计算的偏差也越来越大.为了动态地考虑特征之间的相关性,我们把当前模型中的所有特征视为一个总体特征,并计算该总体特征与当前考虑特征之间的相关度.即用公式(10)来代替公式(8).

$$f=\arg\max_{f_j}\{\alpha E(f_j)-(1-\alpha)C(F,j)\} \tag{10}$$

其中,F 代表当前模型中的所有特征.而且,这样还把步骤 2 里面的复杂度进一步降低到了 $O(N)$量级的 $C()$.我们把使用公式(10)的 MSR 算法称为 MSR-II.

2.7 相关工作

判别学习模型是近年来才被引入自然语言处理领域的[4,5],而且已被证明比生成模型更有效.这可能是因为:(1) 判别学习模型的假设更合理,只要求训练数据和测试数据来自同一分布,而不必像生成模型那样必须知道分布的形式;(2) 判别学习模型所使用的损失函数通常与实际应用中的评价标准相一致,而不是与实际评价标准无直接联系的似然度.

目前,判别学习算法的一大问题在于最直接的评价标准(如 CER)往往都是分段的阶梯形函数,难以直接优化.所以,现有的主流判别学习算法都把重点放在寻找一个光滑、易优化且足够逼近 CER 的损失函数并加以优化上.例如:Boosting 算法[4]优化的是指数级的排序错误;Perceptron 算法[5]优化的是最小平方误差(minimum square error,简称 MSE).尽管这些算法的有效性已被证明,但由于它们所使用的损失函数都不是用户真正关心的评价指标 CER,而仅仅是对 CER 的某种逼近,所以,即使这些损失函数被优化了,也不一定保证 CER 会降低.因此,MSR 采用了另一种截然不同的优化算法,即通过启发式的、类似多维优化的方式来直接优化训练样本上

的 CER,这是 MSR 的最大贡献.

MSR 与许多现有的方法有着相通之处.如 MSR 的算法框架(第 2.5 节)与多维优化[8]的思想一致;而 MSR 线性搜索的实现是 Och 的算法[10]的一个拓展,以适应本实验中大量的特征.在 Och 的实验中总共只有 8 个特征,所以他使用一种简单的网格线性搜索,通过固定搜索区间长度并穷举所有区间来寻找损失函数的极值.而在本实验中有几十万个候选特征,根本不可能通过穷举的方法来找到损失函数的最优值.因此,我们对每一个特征都设置了一个不等长区间的网格序列,并让每个区间都对应一个 SR 值.而 MSR 的特征选择方式(公式(8))则是从文献[11]中得到的启发.

3 MSR/MSR-II 在语言模型中的应用

3.1 评价标准

本文使用日文假名到日文汉字(kana-kanji)的转换作为评价语言模型的方法,用字符转换错误率来评价转换结果.CER 的值等于转换错误的字符数量与正确转换中字符数量的比值.除了衡量 MSR/MSR-II 的性能,本文还把它们与另两种判别学习算法 Boosting 和 Perceptron 作了对比.

3.2 实验设置与步骤

本实验使用了两套新闻类语料,Nikkei 作为训练语料,Yomiuri 作为测试语料.这些语料都作了分词预处理,使用的字典包括 167 107 个条目.其中,36 000 000 个词的 Nikkei 语料作为训练数据.此外,另有 100 000 词(约 5 000 句)的 Yomiuri 语料作为 dev 语料,用来调试算法的最佳学习步长、平滑因子、迭代次数等等.我们另取 100 000 词(约 5 000 句)的 Yomiuri 语料作为测试数据.

我们首先在一个含 400 000 个句子的 Nikkei 语料上训练出一个三元模型作为基础模型(使用的是文献[3]中的系统).然后,对 Nikkei 中的另外 80 000 个句子中的每一句拼音字符串,都用这个基础模型为其生成若干个候选转换汉字字符串.注意:这里使用的两部分 Nikkei 语料相互不重叠,这样就与真实情况相类似,因为实际应用中基础模型无法预见将要遇到什么样的语料.为了保持高效,对每个拼音字符串只保留了概率最大的 20 个候选转换作为判别学习的训练语料.其中:最正确的候选转换(即转换错误数最小的那个)作为相应拼音字符串的参考转换(reference),以后所有的编辑距离 $Er()$的计算都是以这些参考转换为依据的.最理想的模型应当输出所有的这些参考转换.

在应用判别学习算法时,我们只把每个候选转换中出现过的 Unigram 和 Bigram 作为候选特征来优化,而并没有考虑 Trigram,因为前期的实验[12]已经证明:加入 Trigram 会大大增加特征的数量,而带来的模型性能的提升却相当有限.在以上 80 000 句的样本上总共得到约 860 000 种特征.

在测试过程中,我们采用了重新打分的方法.在生成 dev 和测试数据时,我们先为每个拼音输入生成了概率最大的 100 个候选转换,且每个转换的得分就是其概率 $P(W)$的 log 值.然后,使用训练好的判别模型对每个拼音输入的所有候选转换重新打分,并输出得分最高的候选转换.最后计算所有输出所造成的字符转换错误率 CER.

3.3 实验结果与分析

对比实验的主要结果见表 1.第 1 行是使用基于词的三元模型(基础模型)的结果.请注意:这里的结果比目前市场上所能达到的最好效果[3]还要好得多,这可能是因为我们使用了大量的训练样本,而且测试数据与训练样本十分相似.第 2 行、第 3 行是应用第 3 节中所描述的 MSR 和 MSR-II 后的结果.同时,我们还列出了其他两种主流的判别学习算法的结果:第 4 行是文献[4]中描述的 Boosting 的一个改进实现;第 5 行是文献[5]中平均 Perceptron 的实现.这几种判别算法都使用基础模型的打分作为基础特征.

显然,所有的判别学习算法都大大优于 MLE(显著性水平<<0.01).其中:MSR/MSR-II 使得 MLE 的 CER 下降了超过 20%,而所使用的训练数据量却只是 MLE 的 1/5.同时,因为 MSR/MSR-II 仅仅使用了 Unigram 和 Bigram 的信息,而这些信息本身就是包含在 MLE 模型之中的,所以,CER 的降低完全归因于 MSR/MSR-II 本身的优良性能.同时,虽然看似性能接近,但 MSR/MSR-II 也要显著优于 Boosting 和 Perceptron(显著性水平<0.001).

Table 1 Comparison of CER results

表 **1** CER 结果比较

	Model	CER (%)	% over MLE
1.	MLE	3.73	–
2.	MSR	2.95	20.9
3.	MSR-II	2.95	20.9
4.	Boosting	3.06	18.0
5.	Perceptron	3.07	17.8

与 MSR 相比,MSR-II 在时空复杂性上都有所提高.当选择 2 000 个特征,N=1000 时,在 XEON(TM) MP 1.90GHz 的机器上,MSR-II 在 18 分钟内就可以完成训练过程,而 MSR 则需要 20 分钟.但这比 Boosting 和 Perceptron 都要快.

3.4 MSR的特征选择分析

为了验证特征选择的有效性,我们考察含 2 000 个特征、窗口为 1 000 的模型,并把公式(8)和公式(10)中的调节因子α在[0.7,1.0]*里调节.这样,总共有 4 种不同的特征选择设置.对每一种设置,我们分别用 MSR 和 MSR-II 各训练一个模型.这些模型在训练/测试数据上的 CER 曲线如下.

首先,即使在特征选择时完全不考虑特征之间的相关性(α=1),MSR/MSR-II 的性能依然非常优秀.在训练样本上的 CER 可以达到 2.09%(如图 1 所示),在测试数据上的 CER 也能达到 2.99%(如图 2 所示).这证明了引入特征独立假设的合理性.因为这里的特征都是 Unigram 和 Bigram,它们之间的冗余信息本来就比较小.

然后,引入特征相关性进一步提升了 MSR/MSR-II 的性能.这样,不仅能更好地拟和训练数据(如图 1、图 3 所示),而且泛化性也明显加强(如图 2、图 4 所示).此外,过拟和现象也大为减轻.

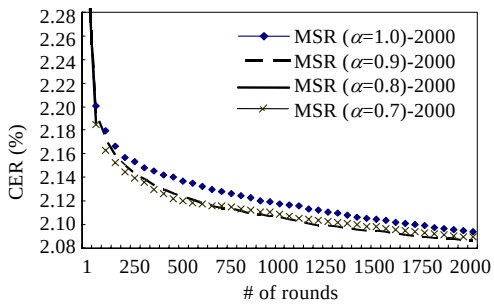

Fig.1 MSR's results on training data

图 1 MSR 在训练数据上的 CER 结果

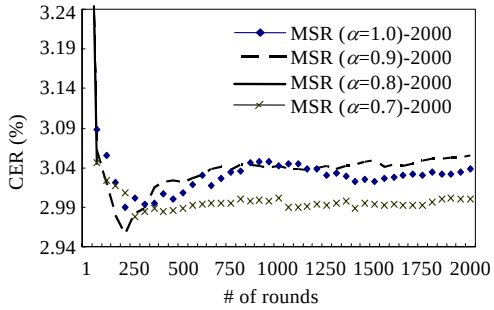

Fig.2 MSR's results on test data

图 2 MSR 在测试数据上的 CER 结果

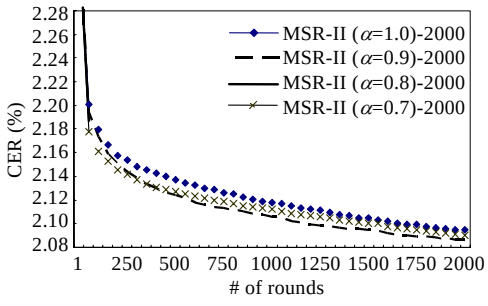

Fig.3 MSR-II's results on training data

图 3 MSR-II 在训练数据上的 CER 结果

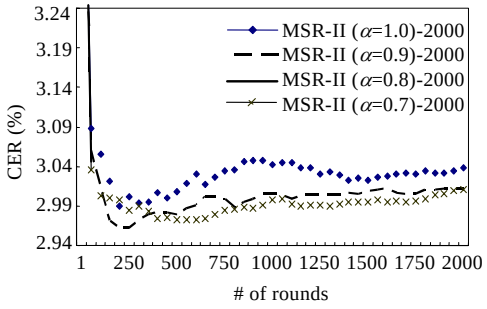

Fig.4 MSR-II's results on test data

图 4 MSR-II 在测试数据上的 CER 结果

最后,虽然 MSR 在测试数据上能取得很好的 CER 结果,但其特征选择还不够稳定,仅当α处于[0.7,0.8]这一

* 前期实验证明:当 $0\le\alpha<0.7$ 时,MSR/MSR-II 的效果都比α=1 时要差,因此,我们没有列出这些结果.

狭小范围之内时,特征相关度计算才能带来性能上的提升(如图 2 所示),这从某种程度上限制了 MSR 的应用.而在使用了动态相关度计算之后,MSR-II 的 CER 结果基本上与 MSR 相当,但是稳定性大为提高.如图 4 所示,α在一个相对较大的范围[0.7,0.9]内都能够起到比α=1 时更好的效果.这为 MSR-II 在实际应用中调节参数提供了方便.

3.5 MSR/MSR-II的鲁棒性分析

大多数关于判别学习算法鲁棒性的理论说明关注以下两个方面.一是收敛性.很多时候,训练数据是线形不可分的,比如在 IME 中,若某个拼音输入含有噪音或根本不正确,则任何一种模型都无法正确转换它.算法必须能够处理这种情况,保证在任何情况下算法都要能够在有限时间内收敛.二是泛化性,即算法不仅要尽量拟和已知数据(训练数据),还要在未知数据上同样表现良好.Boosing 和 Perceptron 都已经有这两方面的理论证明.

MSR/MSR-II 的算法设计保证了其收敛性.因为每当选中一个特征之后,MSR/MSR-II 都使用线性搜索来选择该特征的参数值,以使得样本风险最小化.一旦加入该特征会增加样本风险,就把该特征的参数设为 0.这就保证了 MSR 的样本风险一定随着迭代的增加而减小(如图 1、图 3 所示).

至于 MSR/MSR-II 的泛化性,我们迄今尚未能从理论上加以证明.尽管如此,我们也可以通过领域适应性建模实验,从一个侧面考察这一点.领域适应性建模就是指把在一个领域上训练出的模型应用到另一个未知领域上.由于不同领域的数据差别很大,所以,该实验能够比较有效地衡量算法的泛化性.

本实验总共使用了 5 种领域的数据.其中:Nikkei(新闻语料,见第 3.1 节)是背景领域,用来训练基础三元模型;另有 4 个适应领域:Yomiuri(新闻语料)、TuneUp(包含新闻语料的平衡数据集)、Encarta(百科全书)以及 Shincho(小说集).对每一个适应领域,我们都使用了 3 种规模的训练数据(800/8K/72K 句)、5K 句的 dev 数据和 5K 句的测试数据.其他建模、测试方式都与第 3.1 节相同.

本实验将 MSR-II 与另外 4 种算法作了对比.其中,Baseline 就是把背景领域上的三元模型直接应用到适应领域上的结果.LI(linear interpolation,线形插值)是一种最大化后验概率算法.该算法首先在适应领域的数据上训练一个三元模型,然后把它与背景领域上的三元模型做线形插值,以使得适应领域数据的似然度最大化.MSR-II,Boosting 和 Perceptron 都使得背景三元模型在适应领域数据上所引起的 CER 最小化(见表 2).

Table 2 CER results of domain adaptation

表 2 领域适应性建模的 CER 结果

Domain	Baseline	LI	MSR-II	Boosting	Perceptron
Yomiuri (800)	3.70	3.70	3.17	3.13	3.18
Yomiuri (8K)	3.70	3.69	2.88	2.88	2.85
Yomiuri (72K)	3.70	3.69	2.73	2.78	2.78
TuneUp (800)	5.81	5.81	5.70	5.69	5.69
TuneUp (8K)	5.81	5.70	5.47	5.47	5.47
TuneUp (72K)	5.81	5.47	5.15	5.3	5.20
Encarta (800)	10.24	9.60	9.44	9.82	9.43
Encarta (8K)	10.24	8.64	8.42	8.54	8.34
Encarta (72K)	10.24	7.98	7.40	7.53	7.44
Shincho (800)	12.18	11.86	11.89	11.91	11.90
Shincho (8K)	12.18	11.15	11.04	11.09	11.20
Shincho (72K)	12.18	10.76	10.16	10.25	10.18

从表 2 可以清晰地看出,在各种适应领域以及不同大小的训练数据上,MSR-II 都明显优于 LI 算法,且都能取得与 Boosting 和 Perceptron 相当的效果.这从一个侧面证明了 MSR-II 具有良好的泛化性.

4 结 论

最小化样本风险算法 MSR 是一种新型的判别学习算法,能够直接优化不可导的阶梯形函数.本文对 MSR 的算法实现进行了深入、细致的研究,减低了时空复杂性.在此基础上提出了 MSR-II,减轻了 MSR 特征相关度计算不够稳定的问题.此外,还通过大量领域适应性建模实验,从一个侧面验证了 MSR-II 的良好泛化性.

然而迄今为止,我们尚未能完全从理论上证明 MSR/MSR-II 的鲁棒性,这对其实际应用造成一些障碍.这还

有待于以后的工作给出相关证明.同时,鉴于 MSR/MSR-II 属于线性模型框架,具有很强的开放性,我们将考虑加入各种包含语义信息的特征来提高其性能.

另一个令人感兴趣的方向是 MSR/MSR-II 的适用性.因为它以一个不可导的阶梯形函数为损失函数,而在自然语言处理中,许多相关领域的评价标准都是这种形式,如拼写检查中的 F 值和机器翻译中的 BLEU 值.所以,可以把 MSR/MSR-II 应用到这些新的领域来研究其性能.

致谢 本文的实验由第一作者在微软亚洲研究院访问期间完成,所有数据均由微软亚洲研究院提供.感谢微软亚洲研究院的黄昌宁老师、哈尔滨工业大学的齐浩亮和孙广路同学对本文提出的宝贵意见.

References:

[1] Jelinek F. Self-Organized language modeling for speech recognition. In: Waibel A, Lee KF, eds. Readings in Speech Recognition. San Mateo: Morgan-Kaufmann Publishers, 1990. 450−506.

[2] Brown PF, Cocke J, Pietra SAD, Pietra VJD, Jelinek F, Lafferty JD, Mercer RL, Roossin PS. A statistical approach to machine translation. Computational Linguistics, 1990,16(2):79−85.

[3] Gao JF, Suzuki H, Wen Y. Exploring headword dependency and predictive clustering for language modeling. In: Hajic J, Matsumoto Y, eds. Proc. of the Empirical Methods in Natural Language Processing (EMNLP). MACL, 2002. 248−256.

[4] Collins M. Discriminative reranking for natural language parsing. In: Langley P, ed. Proc. of the 17th Int'l Conf. on Machine Learning (ICML 2000). San Francisco: Morgan Kaufmann Publishers, 2000. 175−182.

[5] Collins M. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In: Hajic J, Matsumoto Y, eds. Proc. of the Empirical Methods in Natural Language Processing (EMNLP). MACL, 2002. 1−8.

[6] Gao JF, Yu H, Yuan W, Xu P. Minimum sample risk methods for language modeling. In: Proc. of the Empirical Methods in Natural Language Processing (EMNLP). 2005. 209−216. http://research.microsoft.com/~jfgao/

[7] Duda RO, Hart PE, Stork DG. Pattern Classification. 2nd ed., Wiley-Interscience, 2000. 117−120.

[8] Press WH, Flannery BP, Teukolsky SA, Vetterling WT. Numerical Recipes in C: The Art of Scientific Computing. 2nd ed., Cambridge: Cambridge University Press, 1992. 412−419.

[9] Quirk C, Menezes A, Cherry C. Dependency tree translation: Syntactically informed phrasal SMT. In: Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL). 2005. 271−279. http://www.cs.ualberta.ca/~colinc/papers/ms_acl05.pdf

[10] Och FJ. Minimum error rate training in statistical machine translation. In: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL). 2003. 160−167. http://acl.ldc.upenn.edu/acl2003/main/pdfs/Och.pdf

[11] Theodoridis S, Koutroumbas K. Pattern Recognition. 2nd ed., Academic Press, 2003. 182−183.

[12] Yu H, Gao JF, Bu FL. One new discriminative training method for language modeling. Chinese Journal of Computers, 2005,28(10): 1708−1715 (in Chinese with English abstract).

附中文参考文献:

[12] 于浩,高剑峰,步丰林.一种新的语言模型判别训练方法.计算机学报,2005,28(10):1708−1715.

袁伟(1981－),男,硕士,主要研究领域为自然语言处理,信息检索.

步丰林(1961－),男,副教授,主要研究领域为软件工程,自然语言处理.

高剑峰(1971－),男,博士,研究员,主要研究领域为自然语言处理,信息检索,机器翻译.