# 基于投票平均的最小错误率训练算法*

王志洋,姜文斌,吕雅娟,刘群
中国科学院计算技术研究所　智能信息处理重点实验室　北京　100190
E-mail: { wangzhiyang, jiangwenbin, lvyajuan, liuqun }@ict.ac.cn

**摘　要:** 最小错误率训练是统计机器翻译中标准的调参方法，但由于搜索过程中的贪婪特性，往往会导致结果不稳定或陷入局部最优。本文提出投票平均方法来增强标准调参方法——通过翻译验证集，对训练过程的中间结果进行投票平均，从而获得更稳定和准确的参数。在新闻和口语两个领域上的中文到英文翻译的实验表明，本方法是有效的。

**关键词:** 最小错误率训练；投票平均；验证集

# Minimum Error Rate Training with Voted Average Algorithm

Wang Zhiyang, Jiang Wenbin, Lü Yajuan, Liu Qun
Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190
E-mail: { wangzhiyang, jiangwenbin, lvyajuan, liuqun }@ict.ac.cn

**Abstract:** *Minimum error rate training (MERT) is a standard tuning parameter procedure in statistical machine translation. However, the greedy nature of its search process makes it unstable and apt to terminate at local optimum. In this paper we propose a novel strategy, voted average, to enhance the original MERT. By weightedly averaging the intermediate results of MERT according to their performance on another validate set; we obtain more stable and precise parameters. Experiments show that such a simple strategy is effective and efficient, obvious improvements are achieved in newswire domain and travel dialogue domain in Chinese-to-English translation tasks, with a little additional efforts.*

**Keywords:** *Minimum error rate training; Voted average; Validate set*

## 1. 引言

近年来,随着对数线性模型(log-linear model)被引入到统计机器翻译(statistical machine translation, SMT)，翻译模型中可以添加更多的特征；这直接导致统计机器翻译领域的繁荣。相应的，如何计算每个特征对系统的贡献的问题也随之而生。解决这一问题的标准方法是 Och 在 03 年提出的最小错误率训练方法(minimum error rate training, MERT)(Och, 2003)。

*本文承自然科学基金项目(项目号 60736014)和国家 863 重点项目(项目号No. 2006AA010108)的资助。

MERT 并不是通过似然估计的方法来调整参数，而是通过考虑系统翻译结果的质量来优化特征权重。在整个特征空间中，先固定其它维的权重，通过线性搜索的方法全局性的调整每一维的特征权重。尽管在每一维的调参结果是全局最优的，但是并不能保证所有维都是全局最优的。此外，在训练过程中，也可能发生过拟合的问题。

在所有基于对数线性模型的统计机器翻译系统中，MERT 是调参过程中关键的步骤；因此任何可能的改进都可以带来整个翻译质量的提高。在这方面已经有很多现成的工作。（Duh and Kirchhoff, 2008)将 MERT 作为一个弱学习器(weak learner)，然后使用 boosting 的方法对 n-best 结果重排序；(Cer et al., 2008) 比较了三种不同的搜索策略，并提出通过规则化(regularization) 和随机搜索的方法来减少陷入局部最优的问题。(Moore and Quirk, 2008)探究了在 MERT 中引入多组随机初始化点带来的翻译质量的提高。(Chiang et al., 2008) 使用 MIRA 算法来替代 MERT，从而避免了 MERT 调参过程中对参数数目的限制。

本文提出了投票平均算法(Voted Average Algorithm，VAA)来处理 MERT 过程中陷入局部最优的问题。方法很简单：使用 MERT 产生的各轮中间结果，也就是参数权重，来分别对验证集进行解码，并由此对各轮参数权重进行投票；然后通过投票数对各轮结果进行加权平均，从而得到更加稳定更加准确的参数。最终的参数结果更接近理论上的最优值，从而能够很好的减缓陷入局部最优和过拟合的问题。在不同的翻译任务的实验上，本方法都带来了翻译质量的提高。

本文的组织结构如下：第 2 节简要介绍下最小错误率训练的过程；投票平均算法在第 3 节重点介绍；然后是实验和结论。

## 2. 最小错误率训练

MERT 已广泛应用于统计机器翻译中：通过以翻译质量(一般以自动评价指标评价，例如 BLEU 值(Papineni et al., 2002) )为优化目标，有效的调整对数线性模型的参数。首先，生成 n-best 翻译候选，在获取候选翻译的同时，记录每个特征的得分；然后以这些翻译候选作为搜索空间，通过线性搜索对特征权重进行优化。由于特征权重已经更新，我们需要使用更新后的特征权重重新运行解码器，并生成 n-best 翻译候选；将新生成的翻译候选和以前的合并，在扩展的候选集上重新调参。如此循环，直到开发集中每一个句子不再有新的翻译候选生成。

最小错误率训练的思想很好，但是它并不能保证全局最优。另外，MERT 得到的特征权重很不稳定，容易陷入局部最优。这里，我们通过投票平均算法来优化 MERT 的结果，使得得到的参数权重更稳定。

## 3. 扩展

本部分提出了投票平均算法及其如何应用于最小错误率训练过程中；此外还介绍了选择验证集的方法和一个启发式约束。

### 3.1 投票平均算法

首先看看图 1。四边形的中心是理论上最佳的特征权重向量，中心附近的点是通过 MERT 训练得到的权重结果。两者之间的距离便是理论与现实的差距 $d$，我们的目标是减少甚至消除

这一差距。投票平均算可以帮助我们更接近中心：通过对靠近中心的有效点进行投票，然后再进行加权平均，我们便可以得到更稳定且接近中心的结果。

实际上，投票平均算法源自于(Krishnamurthy and Tollis, 1989)的加权平均算法(Weighted Average Algorithm)，用于计算组合网络中的信号概率。它对每一成分并不是均等对待，而是都赋予一定比例，然后再进行加权平均。如何求取每个成分的关联比例，也就是得票数目才是问题的关键。

投票方法来源于投票感知机(Freund et al.,1999; Collins,2002)的思想。在绝大部分的感知机模型中，只有当在任何训练实例上都不存在反例时，此时的结果才作为最终的参数，并应用于新的测试实例上。作为一种优化，投票感知机保存了所有的参数的中间结果。对于一个新的测试实例，分别用所有的参数进行测试，取得分最高的作为最终的结果。

投票平均算法可以这样应用于 MERT 训练中。我们可以将整个 MERT 过程看作一个函数，它的输入是开发集，输出是各轮迭代中的最好结果对应的特征权重。这里，所有有效迭代轮的结果都将保存；然后引入验证集来对这些结果进行投票。也就是说，使用这些结果权重来翻译验证集，使用某种指标，例如 BLEU 得分，作为该轮特征权重的得票数。将所有的特征权重进行投票平均便可以得到更健壮的特征权重向量。算法 1 描述了这一过程。5-7 行通过对验证集解码得到每轮权重的投票数，8 行对投票数进行了归一化；9-11 行是投票平均的过程。

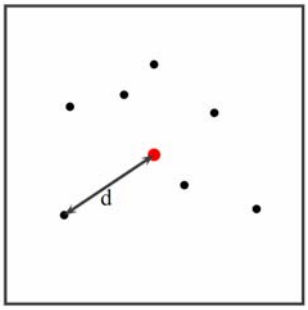


图 1 理想与现实的差距

**Algorithm 1** Applying the VAA to Validate Set.

1: **Define**: $m$ is the number of features, MERT procedure iterates $n$ times, $x_{ij}$ stores the $j$th dimension of the weight vector at the $i$th iteration.
2: **Input**:$\{x_{ij}\}$, the validate set.
3: **bleu** $\leftarrow$ **0**
4: **weight** $\leftarrow$ **0**
5: **for** $i \leftarrow 1 \,..\, n$ **do**
6: translate validate set using $(x_i.)$
7: $\textbf{bleu}_i = bleu\ score$
8: normalize(**bleu**)
9: **for** $j \leftarrow 1 \,..\, m$ **do**
10: **for** $i \leftarrow 1 \,..\, n$ **do**
11: $\textbf{weight}_j += x_{ij} \times \textbf{bleu}_i$
12: **Output: weight**

算法 1 投票平均算法

## 3.2 启发式约束

最小错误率训练并不能保证全局最优，可能会生成很差的局部最优值。为了得到更好的参数权重，我们引入了一个启发式约束：在 MERT 迭代过程中，如果某一轮的特征权重很差(可以通过 BLEU 值表现出来，实验中我们假定阈值为 0.10)，该轮参数将不参加后面的投票平均操作。

## 3.3 验证集选择

投票平均算法中，验证集的选择至关重要。为了得到更好的翻译质量，验证集应和测试集保持最大程度的相似。这里，我们使用简单的语言模型的困惑度，来确定验证集和测试集的相似程度。为了更好的利用现成的数据(NIST[†]评测和IWSLT[‡]评测数据)，以NIST数据为例说明

[†] http://www.nist.gov/speech/tests/mt
[‡] http://iwslt07.itc.it/

验证集的选择方法。假定我们以MT05 和MT08 作为测试集，我们可以据此训练一个很小的语言模型，然后将其它年的数据作为测试语料，得到困惑度最小的作为验证集。在NIST翻译任务中，我们选择MT03 的数据作为验证集；在IWSLT翻译任务中，我们选择IWSLT04 作为验证集。

## 4. 实验

为了验证效果，我们在两个不同领域的中文到英文的翻译任务上进行了实验：分别是 NIST 新闻领域评测和 IWSLT 口语领域评测。

### 4.1 实验设置

在本文中，我们使用 Bruin 系统(Xiong et al.,2006)作为基准系统。Bruin 是基于 BTG 的短语翻译模型，通过柱搜索进行 CKY 解码的解码器。Bruin 与一般的基于短语的模型的区别是可以自动的从双语训练语料中学习调序特征:使用短语的边界词作为特征，然后根据最大熵模型，将相邻的两个短语块分为顺序和逆序两种情型。其它的特征和 Pharaoh(Koehn,2004)类似。

下面的实验中，*Best* 表示传统的 MERT 训练方法，也就是在所有的迭代中选择最好的一轮对应的参数作为结果。*VAA* 表示投票平均算法。此外，我们使用大小写敏感的 BLEU 得分来评价翻译质量。

### 4.2 NIST 任务

在 NIST 任务中，使用 FBIS 语料作为训练语料，它由 7.06 M 中文词和 9.15 M 英文词组成。在数据集选择上，我们使用 MT04 作为开发集，MT03 作为验证集，MT05 和 MT08 作为测试集。此外，我们根据 Gigaword 语料新华部分的前 1/3 训练四元的语言模型。

从表 1 可以看到，投票平均方法在不同的测试语料上均表现出很好的性能。相比于传统的 MERT 方法，VAA 在测试集 MT05 上有 1.2 个点的提高，在 MT08 上有 1.7 个点的提高。由于 MT08 主要来自网络新闻和博客，而 FBIS 语料更接近新闻领域，因此整体翻译效果，MT08 没有 MT05 好。

### 4.3 IWSLT 任务

在 IWSLT 任务上，我们使用了两组不同大小的训练语料。小的语料是 BTEC，大约 40 K 句对；大的语料约有 600 K 句对。分别使用训练语料的英文端训练 3 元的语言模型。然后使用 IWSLT05 作为开发集，IWSLT04 作为验证集，IWSLT07 是测试集。表 2 是实验结果。

同样的，投票平均方法表现要好于传统方法。在 BTEC 语料上约有 0.9 个点的提高，在大语料上约有 0.6 个点的提高。

| | MT-05 | MT-08 |
|---|---|---|
| ***Best*** | 0.2232 | 0.1484 |
| ***VAA*** | 0.2357 | 0.1658 |

表 1　NIST 任务结果

| **corpus** | BTEC | 600k |
|---|---|---|
| ***Best*** | 0.3147 | 0.3204 |
| ***VAA*** | 0.3246 | 0.3269 |

表 2　IWSLT 任务结果

## 5. 讨论

此外，我们还验证了投票轮数对结果的影响。在 MT05 实验中，我们按照 BLEU 值得分来选择最好的 n 轮结果进行投票。图 2 是具体的结果。Top-n 投票方法总好于传统的方法(相当于 n=1)。

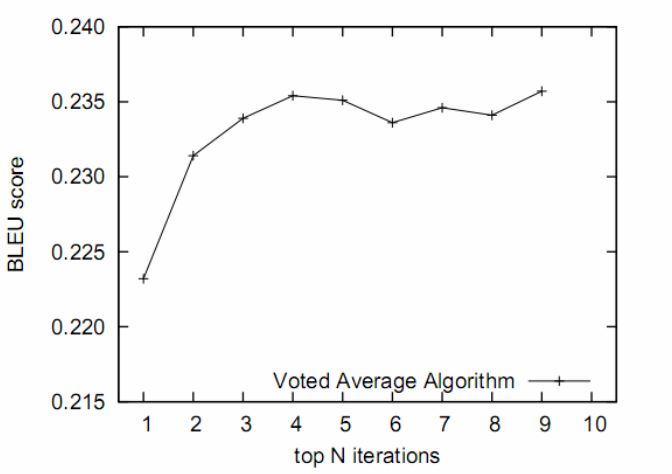


图 2 投票轮数对结果的影响

对每一轮结果进行投票是很合理的；然而，还有一种更直觉的想法：就是平等对待每一轮的结果，也就是说，每一轮都有同样的得票数。这类似于平均感知机算法，仅仅是简单的对所有轮的结果进行平均。(Chiang et al., 2008) 在 MIRA 算法中使用了这种平均操作；但是，Chiang 的文章中没有探究这一操作的效果。我们也将这种简单的平均策略应用于 MERT 中，在 MT05 上 BLEU 值是 0.2301，MT08 上是 0.1598。结果都比传统的 MERT 方法好，但不如我们的投票平均方法。

## 6. 结论与展望

本文提出投票平均算法来优化最小错误率训练的过程，并在实验上取得了不错的效果。方法很简单，但是很有效。

但是，我们还有很多问题需要解决。我们验证集的选择方法太简单，应该考虑更复杂的分类策略。此外，引入验证集后，虽然得到了更好的结果，但是训练时间也相应增加。如何平衡翻译质量和训练时间的关系也是我们需要考虑的。

## 参考文献

Daniel Cer, Daniel Jurafsky, and Christopher D. Manning. 2008. Regularization and search for minimum error rate training. In Proc. Third Workshop on Statistical Machine Translation.

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In Proc. EMNLP 2008.

Michael Collins. 2002. Ranking algorithm for named-entity extraction: boosting and the voted perceptron. In Proc. ACL 2002.

Kevin Duh, and Katrin Kirchhoff. 2008. Beyond log-linear models: boosted minimum error rate training for n-best re-ranking. In Proc. ACL 2008: HLT, Short papers.

Yoav Freund, and Robert E. Schapire. 1999. Large margin classification using the perceptron algorithm. In Machine Learning.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Proc HLT-NAACL 2003.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In Proc. Sixth Conference of the Association for Machine Translation in the Americas.

B.Krishnamurthy and I.G.Tollis. 1989. Improved Techniques for Estimating Signal Probabilities. IEEE Transactions on Computers, July 1989.

Robert C. Moore, and Chris Quirk. 2008. Random restarts in minimum error rate training for statistical machine translation. In Proc. COLING 2008.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In Proc. ACL 2003.

Kishore Papineni, Salim Roukos, ToddWard, and Weijing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proc. ACL 2002.

Ashish Venugopal, and Stphan Vogel. 2005. Considerations in maximum mutual information and minimum classification error training for staistical machine translation. In Proc. EAMT 2005.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In Proc. ACL 2006.