基于统计语言模型的双向词类标注方法

刘启和　詹思瑜　杨国纬
（电子科技大学计算机学院　成都610054）

A Two-Directions Method of Chinese Corpus Tagging Based on Statistical Language Model

LIU Qi-He　ZHAN Si-Yu　YANG Guo-Wei
(Computer School, UEST, Chengdu 610054)

Abstract　In the paper, we introduce chinese corpus tagging based on statistical language model (bi-gram model)and Huang-Yu's smoothing method. Especially, we also suggest a two-directions method based on statistical language model, namely, we not only compute probability of P(C|W)(W=$w_1w_2w_3\cdots w_m$), but also compute probability of P(C|$w_nw_{n-1}\cdots\cdots w_1$). From our experience we can see it can enhance the accuracy of Chinese corpus tagging using this method of two directions computation.

Keywords　Natural language processing, Statistical language model, Smoothing method, Chinese corpus tagging

1　引言

在自然语言处理中，词类标注是一项重要的工作，它为句法分析、机器翻译、自然语言理解等提供语法知识。在进行自然语言的词类标注时，由于词的多词类现象，有许多词在不同的上下文中有不同的词类，汉语词类标注过程其实就是一个词类排歧过程[12]。当前的词类标注有两种方法：基于规则和基于统计的方法。基于规则的方法是利用系统的知识库进行词类标注，但知识库中知识的不足约束了该方法的使用效率。基于统计的方法则是利用语料库计算概率来标注词类，其正确率可以达到94%左右。文[12]介绍的使用规则和统计相结合的方法，将词类标注的正确率提高到96%左右。本文的目标是对文[12]中的统计方法进行改进，提出基于统计的双向标注的方法(第二节中公式(9))，提高了统计词类标注方法的正确率。同时，为解决数据稀疏问题，使用文[2]提出的 Huang-Yu 平滑方法。实验显示，这种双向的方法也可以单独把分词的正确率提高到96%左右。

2　词类标注

设 W 是一词串，$W=w_1w_2w_3\cdots w_m$，$C=c_1c_2c_3\cdots c_m$ 是 W 中相应词的可能词类标注，集合 A={C|C 是 W 的可能词类标注}，即 A 是 W 的所有可能词类标注的集合，对任意的 C∈A，计算条件概率 P(C|W)。在所有的这些 P(C|W)中，取得最大值的 P(C|W)中的 C 被认为是 W 的词类标注。对于 P(C|W)由乘积公式有[12]：

$$P(C|W)=P(W|C)P(C)/P(W) \tag{1}$$

其中，对任意的可能词类标注 C，P(W)是不变的值，所以只需计算 P(W|C)P(C)的值。在 $W=w_1w_2w_3\cdots w_m$ 中，假设每个词 w_i 的词类标注是相互独立的，则

$$P(W|C)=\prod_{i=1}^{m}P(w_i|c_i) \tag{2}$$

对于 P(C)，由乘积公式有：

$$P(C)=P(c_1)P(c_2|c_1)\cdots P(c_m|c_1,c_2,\cdots,c_{m-1}) \tag{3}$$

上式计算较复杂，利用统计语言模型中 bi-gram 模型[1]，则公式可简化为：

$$P(C)=P(c_1)\prod_{i=2}^{m}P(c_i|c_{i-1}) \tag{4}$$

由式(2)和(4)，可以得到[12]：

$$P(W|C)P(C)=\prod_{i=1}^{m}P(w_i|c_i)P(c_1)\prod_{i=2}^{m}P(c_i|c_{i-1}) \tag{5}$$

如果用 C^* 表示 W 的词类标注，argmax P(C|W)表示 P(C|W)取得最大值时的 C，根据以上的分析，可概括为如下的公式[12]：

$$C^*=\arg\max_{C\in A}P(C|W)=\arg\max_{C\in A}P(W|C)P(C)$$

$$=\arg\max_{C\in A}\prod_{i=1}^{m}P(w_i|c_i)P(c_1)\prod_{i=2}^{m}P(c_i|c_{i-1}) \tag{6}$$

其中，$P(w_i|c_i)$描述了词 w_i 被标注为 c_i 的概率大小，而 $P(c_i|c_{i-1})$描述了 w_{i-1}被标注为 c_{i-1}的条件下，w_i 被标注为 c_i 的概率。即词 w_i 的词类标注依赖于词 w_{i-1}的词类标注(称为前向依赖性)。但在上下文环境中，词 w_i 的词类标注也可能依赖于其后面词 w_{i+1}的词类标注(称为后向依赖性)。如词"通知"既可以作动词也可以是名词，看下面的句子[15]："中办通知要求各级党委组织干部群众认真学习悼念邓小平"中的"通知"是名词，是由该句中"通知"后面的词"要求"决定的。

为描述这种后向依赖性，设 $\delta(W)$为串 W 的逆向运算，即 $\delta(W)=w_mw_{m-1}\cdots w_1$，同样，对于词类标注串 C 有，$\delta(C)=c_mc_{m-1}\cdots c_1$。显然，$\delta(\delta(C))=C$。由于 P(C)被描述为 $c_1,c_2,\cdots,c_m$ 的共现概率，P(W|C)也是 $w_1,w_2,\cdots,w_m$ 与 $c_1,c_2,\cdots,c_m$ 的共现条件概率，则：

$$P(C)=P(\delta(C))$$

$$P(W|C)=P(\delta(W)|\delta(C))$$

由乘积公式有：

$$P(C)=P(\delta(C))=P(c_m)P(c_{m-1}|c_m)\cdots P(c_2|c_1)$$

利用 bi-gram 模型[1]，上式可以简化为：

$$P(\delta(C))=P(c_m)\prod_{i=m-1}^{1}P(c_i|c_{i+1})$$

刘启和　讲师，博士研究生，主要研究方向：自然语言处理、人工智能。詹思瑜　助教，硕士研究生，主要研究方向：自然语言处理、人工智能。杨国纬　教授，博士生导师，主要研究方向：自然语言处理、人工智能、计算机网络。

于是可以得到：

$$P(W|C)P(C)=P(\delta(W)|\delta(C))P(\delta(C))$$
$$=\prod_{i=1}^{\bar{m}}P(w_i|c_i)P(c_m)\prod_{i=m-1}^{1}P(C_i|c_{i+1}) \quad (7)$$

如前所述，集合 A 表示 W 所有可能的词类标注串，则集合 $\delta(A)=\{\delta(C)|C\in A\}$ 是串 $\delta(W)$ 的所有可能的词类标注串。同样我们可以对 $\delta(W)$ 串进行词类标注，得到公式：

$$\delta(C^*)=\arg\max_{\delta(C)\in\delta(A)}P(\delta(C)|\delta(W))$$
$$=\arg\max_{\delta(C)\in\delta(A)}P(\delta(W)|\delta(C))P(\delta(C))$$
$$=\arg\max_{\delta(C)\in\delta(A)}\prod_{i=1}^{\bar{m}}P(w_i|c_i)P(c_m)\prod_{i=m-1}^{1}P(c_i-c_{i+1}) \quad (8)$$

则 $\delta(\delta(C^*))=C^*$，C^* 是 W 的一个词类标注。由此，通过对 $\delta(W)$ 的词类标注，得到了一个 W 的词类标注。在式(8)中，描述了 $\delta(W)$ 中词的前向依赖关系，这种前向依赖描述就是 W 的后向依赖描述。为了把这种关系考虑在词类标注系统中，本文提出了如下的词类标注模型：

$$C^*=\begin{cases}C_1 & \text{当 } P(C_1|W)>=P(\delta(C_2)|\delta(W))\\ C_2 & \text{当 } P(C_1|W)<P(\delta(C_2)|\delta(W))\end{cases} \quad (9)$$

其中，$C_1=\arg\max_{C\in A}P(C|W)$

$$\delta(C_2)=\arg\max_{\delta(C)\in\delta(A)}P(\delta(C)|\delta(W))$$

选择 C^* 为 W 的词类标注结果。在本文中，式(6)称为正向方法，式(8)称为逆向方法，而由式(9)计算词类标注被称为双向方法。同时从式(9)可得，双向方法的时间复杂度仅是正向方法的两倍。因此，式(9)是一个有效的方法。

由式(7)，我们从理论上证明 $P(C|W)$ 与 $P(\delta(C)|\delta(W))$ 是相等的，但由于采用 bi-gram 进行了近似计算并用经验概率来计算式(6)和(8)，所以计算出的结果并不相等。因此式(9)是合理的。

对式(6)中的概率 $P(w_i|c_i)$ 和 $P(c_i|c_{i-1})$ 使用语料库中的经验概率来估计。即：

$$P(w_i|c_i)=C(w_i,c_i)/C(c_i) \quad (10)$$
$$P(c_i|c_{i-1})=C(c_{i-1},c_i)/C(c_{i-1}) \quad (11)$$

其中，$C(w_i,c_i)$ 表示在语料库中词 w_i 被标注为 c_i 出现的次数，$C(c_{i-1},c_i)$ 为 $c_{i-1}c_i$ 出现的次数，$C(c_i)$ 表示 c_i 的次数。这样估计概率的方法叫做最大似然估计[2](MLE，maximum likelihood estimation)。相应地，式(8)也类似地进行计算。

3 平滑方法

由式(11)可以看出，如果 $c_{i-1}c_i$ 在语料库出现的次数为零，则相应的式(5)为零。在 MLE 方法中，次数为零的事件其统计概率为零，这会给计算结果带来偏差。在大规模的语料库中，由于数据的稀疏问题，次数为零的元素总是存在的。因此，必须重新调整由 MLE 方法得到的概率，为新事件分配一个非零的概率。解决零次数的方法叫做平滑方法。

平滑方法在本质上就是将 MLE 得到的概率调整并重新分布，以减少零次数问题。这个概率的调整过程应该满足一定的性质[2]，如调整后的概率应在0和1之间，概率总和应为1。

文[2]对一些常用的基于折扣率[1](discount)平滑方法进行了讨论，如 Additive discount，Good-Turing 等方法，并指出了它们不足之处，提出了一种新的平滑方法(Huang-Yu 方法。实际上，文[2]描述的是平滑前的概率与平滑后的概率比较，使得平滑后零次数的既被解决，又让平滑前与平滑后的性质相似。所以，我们选用 Huang-Yu 方法来解决语料库中的数据稀疏问题。

设 N 是语料库的大小，U 为次数为零的 $c_i\,c_{i-1}$ 的个数，则 Huang-Yu 方法平滑后的概率可用如下公式表示[2]：

$$P^*(c_i|c_{i-1})=\begin{cases}\dfrac{d}{N+1} & \text{当 } c(c_i,c_{i-1})=0\\ \dfrac{c(c_i,c_{i-1})(N+1-Ud)}{N(N+1)} & \text{当 } c(c_i,c_{i-1})>0\end{cases} \quad (12)$$

其中，d 是一个常数：

$$d<\min\{\frac{N}{N+2U},\frac{N+2}{2U}\} \quad (13)$$

根据式(12)，平滑后所有次数为零的2元对的概率总和为：

$$\sum_{c(c_i,c_{i-1})=0}P^*(c_i|c_{i-1})=\frac{Ud}{N+1} \quad (14)$$

相似地，有 $P^*(w_i|c_i)$。将 $P^*(w_i|c_i)$ 和 $P^*(c_i|c_{i-1})$ 代入式(6)和(8)可计算 $P(C|W)$ 和 $P(\delta(C)|\delta(W))$。

4 实验及分析

本实验是在分词之后进行词类标注(分词算法可以采用如最大匹配算法等)，实验系统的输入为分词和初始标注结果，然后进行统计方法的词类排歧，输出结果为词类标注结果。采用词类标注错误率来对方法进行评价。即：方法1：基于统计语言模型的正向方法(式6)；方法2：基于统计语言模型的双向方法(式9)。所谓错误率为：

$$\frac{\text{标记错误的词次数}}{\text{被标记的文本中词总数}}$$

首先对测试文本进行手工标注，并认为手工标注为正确的标注，然后将系统标注的结果与手工标注进行比较，相同的为正确标注。实验所使用的语料库为北京大学计算语言所的 PFR 语料库，其大小约为8M，并使用相应的小标注集[15]。

	方法1	方法2
词总数	48568	48568
错误数	3216	1945
错误率	6.62%	4.00%

通过实验，我们得到如下的结果：

1)实验结果表明，由于考虑了标注词对下文的依赖性，因此基于双向的方法(即方法2)将词类标注的正确率提高到96%。如对于句子“中办通知要求各级党委组织干部群众认真学习悼念邓小平”两种方法的标注结果如下：

方法1：中办/j 通知/v 要求/v 各级/r 党委/n 组织/v 干部/n 群众/n 认真/a 学习/v 悼念/v 邓/nr 小平/nr

方法2：中办/j 通知/n 要求/v 各级/r 党委/n 组织/v 干部/n 群众/n 认真/a 学习/v 悼念/v 邓/nr 小平/nr

其中，v、n 分别表示动词和名词，其它的标记意义可参见文[15]。方法2得到了正确的标注结果。而方法1将句中的“通知”错误地标注为动词。

2)方法2只是从一个侧面表达了标注词的上下文依赖关系。在统计语言模型中，寻找一个能更加有效描述这种关系、计算上切实可行的方法，是进一步需要考虑的内容。

结束语 本文分析了统计语言模型中正向词类标注的不足，即不能表达后向的依赖性，提出了基于统计语言模型的双向词类标注方法，在一定程度上克服了正向方法的不足，通过其实验显示该方法能提高词类标注的正确率。并指出，要使

(下转第168页)

下，并对各子信息分别采用不同的加密方法、访问控制及用户身份鉴别机制。

结语 本文介绍了可信计算的主要内容、基于FC-SAN的远程备份、集群与系统恢复等抗毁机制，讨论了系统容错功能设计和信息网络系统安全保障技术，并分析了用于保障抗毁性和提高安全性的信息冗余分散策略和模型。网络技术的发展和应用使得可信计算有望成为商业应用的可能目标。

为适应发展需要，计算机系统正在日益复杂化，其系统结构已从20世纪80年代中期以前的本地集中计算模型发展为本地分布计算模型，再进一步发展到目前基于Intranet/Internet的广域分布计算模型，人们对于计算机及系统的性能评价也从最早的单一可靠性指标，发展到RAS(Reliability＋Availability＋Serviceability)指标，并进一步发展到RASIS(RAS＋完整性Integrity＋安全性Security)指标，再发展到如今的可信性指标。

目前正在发展的网格(Grid)计算技术将力求把Internet上包括硬件、软件、数据库和各种信息获取设备等所有资源，连接成一个整体，使整个网络如同一台资源极其丰富的超强计算机，向每个用户提供高性能的服务；网格技术时代的信息存储是基于虚拟存储架构VSA(Virtual Storage Architecture)和存储公用设施模型SUM(Storage Utility Model)的广域分布和高度共享方式，必将为可信计算提供更有力的支持和更灵活的选择。

可信性及其研究标志着计算机技术向更高层次发展和跃升。2000年12月11日，由美国CMU与NASA的AMES研究中心牵头，包括MIT(麻省理工学院)、华盛顿大学、乔治亚理工学院和IBM、COMPAQ、SUN、HP、Microsoft、Sybase、Adobe等12家公司成立了高可信计算协会(High Dependability Computing Consortium)，致力于对高可信性计算进行基础研究、实验研究和工程研究[15]，可见其受到学术界和产业界的重视程度。

参考文献

1 Avizienis A, Laprie J C, Randell B. Fundamental Concepts of Dependability[A]. In: Proc of the ISW-2000[C]. Boston, MA, Oct. 2000

2 Laprie J C. Dependable computing: concepts, limits, challenges [A]. In: Special Issue FTCS-25[C]. Pasadena CA. 1995. 42～54

3 Kyriakopoulos N, Wilikens M. Dependability of complex open systems[A]. In: Proc of the ISW-2000[C]. Boston, MA, Oct. 2000

4 IFIP WG10.4 on Dependable Computing and Fault Tolerance [EB/OL]. http://www.dependability.com, 2002

5 Mohan G, Murthy C S R. Lightpath Restoration in WDM Optical Network[J]. IEEE Network, 2000, 14(6): 24～32

6 Fumagalli A, Valcarenghi L. IP Restoration vs WDM Protection [J]. IEEE Network, 2000, 14(6): 34～41

7 Zhou Dongyun, Subramaniam S. Survivability in Optical Networks[J]. IEEE Network, 2000, 14(6): 16～23

8 Clark T. Designing Storage Area Networks[M]. Longman Inc., 1999

9 Farly M. Building Storage Networks[M]. McGraw-Hill Inc., 2000

10 Avizienis A. The N-Version Approach to Fault-tolerant Software [J]. IEEE Trans Sof Eng, 1985, 11(12)

11 Avizienis A. Toward Systematic Design of Fault-tolerant Systems [J]. IEEE Computer, April 1997

12 Wylie J J, Bigrigg M W, et al. Survivable Information Storage Systems[J]. IEEE Computer, Aug. 2000. 61～68

13 Hiltunen M A, Schlichting R D. Enhancing Survivability of Security Services Using Redundancy[A]. In: Proc of The Int'l Conf on Dependable Systems and Networks (DSN'01)[C]. Goteborg, Sweden, 2001. 173～182

14 Snow A P, Straub D, et al. The Survivability Principle: IT-Enabled Dispersal of Organizational Capital [EB/OL]. http://www.cis.gsu.edu, 2002

15 High Dependability Computing Consortium (HDCC) [EB/OL]. http://www.hdcc.cs.cmu.edu, 2002

16 王志刚. 存储系统可信性及其保障技术探析[J]. 计算机研究与发展, 2003, 40 (5. 增刊)

17 王志刚. 计算机容错技术及其发展与应用综述[J]. 计算机应用, 2002, 22 (8. 增刊)

(上接第60页)

统计语言模型提高词类标注的正确率和排歧能力，需进一步描述词的上下文依赖关系。

参考文献

1 Rosenfeld R. Two decades of statistical language modeling: where do we go from here?. In: Proc. of the IEEE, Vol. 8, 2000

2 Huang F-L, Yu M-S. Analyzing the properties of Smoothing Methods for Language models. IEEE 2001

3 Chen S F, Goodman J. An empirical study of smoothing techniques for language modeling. Computer Speech and Language, 1999, 13

4 Church K W, Gale W A. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bi-grams. Computer Speech and Language, 1991, 5

5 Nadas A. On Turing's formula for word probabilities. IEEE Trans. On Acoustic, Speech and Signal Processing, 1985, ASSP-33

6 Witten L H, Bell T C. Zero-frequency problem: estimating the probabilities of Novel events in Adaptive text compression. IEEE Transaction on information theory, 1991, 37

7 Su K Y, Chiang T H, Chang J S. A overview of corpus-Based statistical-oriented techniques for natural language processing. Computational Linguistics and Chinese language processing, 1996, 1

8 Ney H, Essen U. On smoothing techniques for bi-gram-based natural language modeling. In: IEEE intl. conf. on Acoustic, Speech and Signal processing, 1991

9 Essen U, Steinbiss. Cooccurrence smoothing for Stochastic language modeling. IEEE international conference on Acoustic. Speech and Signal processing, 1992, 1

10 Kenser R, Ney H. improved backing-off for m-gram language modeling. IEEE international conference on Acoustic, Speech and Signal processing, 1995

11 Jiang Ming, Zhu Xiaoyan 等. Braille to print translations for Chinese. Information and Software Technology, 2002, 44

12 周强. 规则和统计相结合的汉语词类标注方法. 中文信息学报, 1995, 9(3)

13 朱靖波, 等. 基于对数模型的词义自动消歧. 软件学报, 2001, 12(9)

14 赵石顽, 等. 基于统计的中文词分类. processing of the 3 World congress on intelligent controln and automation, 2000

15 http://www.icl.pku.edu.cn/Introduction/corpustagging.htm