SSD模型及其在词性标注中的应用*

邢富坤[1,2]　宋柔[1]　罗智勇[1]

1 北京语言大学 语言信息处理研究所，　100083

2 解放军外国语学院，　471003

E-mail: xingfukun@blcu.edu.cn

摘　要：本文提出了一种以符号解码与数值解码并举的 SSD（Symbol-and-Statistics Decoding Model）模型，该模型被用于汉语词性标注任务，其标注正确率在封闭测试中达到 97.08%，开放测试中达到 95.67%，较 2 阶 HMM 的 95.56%和 94.70%都有较为显著提高。SSD 模型的正确率虽然不及最大熵模型和 CRF 模型，但它的训练时间远少于后者，说明 SSD 模型在处理自然语言中的特定任务时是一种较强的实用模型。

关键词：SSD 模型；符号解码；数值解码；HMM；词性标注

Symbol-and-Statistics Decoding Model and Its Application in POS Tagging

XING Fu-Kun[12],　Song Rou[1] , Luo Zhiyong[1]

1. Center of Language Information Processing, Beijing Language and Culture University, Beijing 100083

2. Foreign Languages University of PLA, Luoyang 471003

E-mail: xingfukun@blcu.edu.cn

Abstract: A statistical language model named Symbol-and-Statistics Decoding (SSD) language model is presented in this article. The 2-gram SSD model is applied to the Chinese POS tagging task and achieves a quite good result. The precision rate in the closed test is as high as 97.08% and in the open test is 95.67%, which are both significantly higher than 95.56% and 94.70% by the HMM model. Although the performance of SSD model is not as good as the conditional models such as Maximum Entropy Model and CRF model, the training time of SSD is much less than the conditional models, which makes SSD model more applicable to certain tasks in natural language processing.

Keywords: SSD model, Symbol decoding, Statistics decoding, HMM, POS tagging

1. 引言

隐马尔可夫模型（HMM）被广泛地应用于自然语言处理任务之中，最早被用来解决语音识别问题，并取得令人满意的效果，成为主流方法，而后又被用于解决词性标注问题，也取得了相当好的成绩。HMM 在解决词性标注问题时，其主要思路是在给定模型参数的情况下，求出给定语句的最大概率的状态序列，即给定观察序列 $W=w_1w_2...w_h$，求该序列对应的概率最大的状态序列 $\hat{Q}=q_1,...,q_h$，也就是求：

$$\hat{Q}=\arg\max_{Q} P(Q \mid W)$$

该式可以进一步转化为下式：

$$\hat{Q}=\arg\max_{Q} P(Q)P(W \mid Q) \qquad \text{（公式 1）}$$

HMM 存在一个重要假设，称为输出独立性假设，其基本内容是当前可能状态到当前观察值

*本文得到国家自然科学基金项目（60572159, 60872121）的资助，在成文过程中，还得到了杨尔弘、张劲松、荀恩东和林民等老师的指导和帮助，特此致谢。

的发射概率只与当前观察值有关，而与其他观察值无关。这种假设在解决某些特定问题时是基本成立的，但是在自然语言中，这种假设与现实差别很大。例如：

例（1）领导/n 强调/v 深入/v a 细致/a 的/u 工作/vn 作风/n

例（2）领导/n 要/v 深入/v a 困难/a 的/u 群众/n 中间/f

假定在这两句中，只有“深入”是兼类词，有动词 v 和形容词 a 两个可能词性，需要进行词性排岐，而其他词只有唯一词性。当利用 1 阶 HMM 模型估计例（1）中“深入”的词性 X 时，根据上述公式有：

$$\hat{Q}_1 = \arg\max_{X\{a,v\}} \mathrm{p(n)p(v\mid n)p(X\mid v)p(a\mid X)p(u\mid a)p(vn\mid u)p(n\mid vn)p(领导\mid n)p(强调\mid v)}$$

$$\mathrm{p(深入\mid X)p(细致\mid a)p(的\mid a)p(工作\mid vn)p(作风\mid n)}$$

由于除了“深入”以外，其他词性均唯一且确定，因此可以得到

$$\hat{Q}_1 = \arg\max_{X\in\{a,v\}} \mathrm{p(X\mid v)p(a\mid X)p(深入\mid X)}$$

同理，我们也可以求出例（2）中“深入”的词性为

$$\hat{Q}_2 = \arg\max_{X\in\{a,v\}} \mathrm{p(X\mid v)p(a\mid X)p(深入\mid X)}$$

可以发现$\hat{Q}_1$与$\hat{Q}_2$二者完全相同，因此利用 1 阶 HMM 判断例 1 与例 2 中“深入”的词性时，会得出二者具有相同词性的估计，要么都判定为动词，要么都判定为形容词。显然这种判断是错误的，错误原因在于 HMM 在判断“深入”词性时并没有考虑其前后出现的词对其词性的影响，即输出独立性假设导致错误判断。

本文提出了一种模型，称为 SSD（Symbol-and-Statistics Decoding）模型，该模型以 n 元词序列为观察单元，并在相邻观察单元间具有 n-1 元搭接关系，较好地克服了 HMM 模型的不足。

本文的结构安排是：第 1 节对 HMM 进行介绍及分析；第 2 节是对 SSD 模型的形式化描述及与 HMM 的对比分析；第 3 节介绍 SSD 模型的参数估计及稀疏数据处理方法；第 4 节介绍评价方法；第 5 节介绍词性标注实验并与最大熵模型进行比较。

2. SSD 模型介绍

n 元 SSD 模型的观察单元是由 n 个词组成的序列，而不是单个词。我们这里给出 2 元 SSD 模型的形式化描述，n 大于 2 的模型可由此类推。

设有基本状态集 Q。给定观察序列 S=$w_1 \ldots w_h$，设 2 元观察序列为$w_{i-1}w_i$（2≤i≤h）对应的可能状态序列的集合是$e_{i-1,i} = \{q_{i-1}^j q_i^j\}$，其中$q_{i-1}^j$是$w_{i-1}$对应的基本状态之一，$q_i^j$是$w_i$对应的基本状态之一，j=1,2,…用以区分不同的状态序列。注意，由于$e_{i-1,i}$是在观察序列$w_{i-1}w_i$中两个观察依序共现的前提下出现的，考虑了这个前提，可能的状态序列不会太多。

利用 2 元 SSD 模型求解 S=$w_1 w_2 \ldots w_h$的最优状态序列的过程可以表示为：

$$\hat{Q} = \arg\max_{Q\in U^h} \mathrm{P(Q)P(S\mid Q)} \approx$$

$$\arg\max_{q_{i-1},q_i}(p(q_1)p(q_2\mid q_1)\prod_{i=3}^{h}p(q_{i-1}q_i\mid q_{i-2}q_{i-1})p(o_1\mid q_1)\prod_{i=2}^{h}p(o_{i-1}o_i\mid q_{i-1}q_i))$$

（公式 2）

为了便于计算，我们在序列 S 的起始位置统一加入起始标记序列“*开始*-*开始*”，其状态记为 B-B，结束标记序列“*结束*-*结束*”，其状态记为 E-E，则公式 2 可以进一步表示为：

$$\hat{Q}=\arg\max_{q_{i-1},q_i}(\prod_{i=1}^{h+2}p(q_{i-1}q_i\mid q_{i-2}q_{i-1})\prod_{i=1}^{h+2}p(o_{i-1}o_i\mid q_{i-1}q_i))$$ （公式 3）

在模型中，$w_{i-2}w_{i-1}$ 可能状态序列 $q_{i-2}^{k}q_{i-1}^{k}$ 与同其邻接的 $w_{i-1}w_i$ 可能状态序列 $q_{i-1}^{j}q_i^{j}$ 之间有搭接部分。$q_{i-2}^{k}q_{i-1}^{k}$ 中，搭接部分为 q_{i-1}^{k}，在 $q_{i-1}^{j}q_i^{j}$ 中为 q_{i-1}^{j}，由于搭接部分的观察序列完全相同，因此搭接部分的状态序列也完全相同，即 $q_{i-1}^{k}=q_{i-1}^{j}$。这样才能够形成有效转移，否则转移概率无定义。因此，我们将

$$q_{i-1}^{k}=q_{i-1}^{j}$$ （公式 4）

公式 4 称为 2 元 SSD 模型的搭接约束条件公式。

通过以上公式求解出由 h+2 个 2 元状态序列组成的最优状态序列

$$B\hat{q}_1,\hat{q}_1\hat{q}_2,\hat{q}_2\hat{q}_3,\ldots,\hat{q}_{h-1}\hat{q}_h,\hat{q}_hE\ （\hat{q}_i\in Q）$$

显然，它们唯一地确定了每个观察所对应的状态。

SSD 模型与 HMM 模型主要有 3 点不同：

首先，在n阶HMM中，与t时刻的可能状态 q_t 相关联的观察，只考虑了 o_t；但在n元SSD模型中，则要考虑包含O_t的n个基元（词性标注中为词）所构成的序列。每一个可能状态序列的集合由于受到n个观察值共现的约束，其规模会大大减小，从而模型的搜索范围大大压缩。

第二，n 阶 HMM 中，涉及 t 时刻的状态 q_t 和观察值 o_t 的概率只有 $P(o_t\mid q_t)$；而在 n 元 SSD 模型中，则有 n 个发射概率：$P(o_{t-n+1}\ldots o_t\mid q_{t-n+1}\ldots q_t),\cdots,P(o_t\ldots o_{t+n+1}\mid q_t\ldots q_{t+n+1})$。如此，观察值的前后联系将对状态的判断形成约束。

第三，n 阶 HMM 中计算 n 个状态的序列到下一个状态的转移概率 $P(q_i\mid q_{i-n},\ldots q_{i-1})$；n 元 SSD 模型则计算的是相邻且搭接的两个 n 元状态序列之间的转移概率。当搭接部分相同时，即满足搭接约束条件时，这个概率同 n 阶 HMM 中的概率是相同的；当不满足约束条件时，转移概率无定义。这一约束条件剪裁掉了大量的搜索路径，进一步提高了解码的速度。

下面通过实例说明 2 元 SSD 模型求解最优状态序列的过程，从中可以发现，该句通过符号解码，不必进行概率计算就可以得到最终的最优词性序列，如下表所示：

表 1 SSD 模型解码结果

观察值	*开始*-*开始*	*开始*-领导	领导-强调	强调-深入	深入-细致	细致-的	的-工作	工作-作风	作风-*结尾*	*结尾*-*结尾*
S1	**B-B**	**B-n**	**n-v**	**v-a**	**a-a**	**a-u**	**u-v**	**vn-n**	**n-E**	**E-E**
S2		**B-vn**		**v-ad**	**ad-ad**		**u-n**			
S3		**B-v**					**u-vn**			

图中阴影部分的节点是由于不满足前后搭接约束条件而被剪裁掉的节点，当这些节点剪裁掉后，剩下的只有唯一一条可能路径，这也是最终所要求解的最优路径。

在实际标注过程中，并不一定每次都能够通过符号解码获得唯一可能路径。当符号解码后

的可能路径不唯一时就需要进行数值计算，利用 Viterbi 算法进行数值解码，然后得到最优状态路径。

3. 参数估计及稀疏数据处理策略

SSD 模型需要估计的参数有两个：（1）状态转移参数 Pt；（2）状态发射参数 Po 。我们采用最大似然法估计相关参数，篇幅所限不给出具体过程。

SSD 模型采用回退策略解决数据稀疏问题，设某个 n 元词序列 $w_{j-n+1}...w_j$ 未在词表中出现，则根据回退策略取 $w_{j-n+1}...w_j$ 的后 n-1 个词组成 n-1 元词序列 $w_{j-n+2}...w_j$ 作为替代序列，如果该序列仍然未在词表中出现，则继续回退，直至成为 2 元词序列。回退到 s 元词序列时，就使用 s 元词表中给出的词性序列。但如果 $w_{j-1}w_j$ 仍未在 2 元词表中出现，则不再回退到单个词，而将词 w_{j-1} 与词 w_j 的所有可能词性组合作为 $w_{j-1}w_j$ 的词性序列。

4. 评价方法

(1) 总体标注正确率

$$\text{总体标注正确率}=\frac{\text{正确标注的词数}}{\text{总词数}}*100\ \%$$

(2) 兼类词标注正确率

$$\text{兼类词标注正确率}=\frac{\text{正确标注的兼类词数}}{\text{兼类词总数}}*100\ \%$$

(3) 优化幅度

$$\text{优化幅度}=\frac{SSD\text{模型标注正确率}-\text{HMM模型标注正确率}}{1-\text{HMM模型标注正确率}}*100\%$$

5. 实验设计及结果

5.1 语料与预处理

训练语料与测试语料均来自北京大学标注的 1998 年上半年人民日报，具体划分为如下：

表 2 语料划分

组别	语料类别	语料内容	语料规模（词）
1	训练语料	1998 年 2 月人民日报	1082238
2		1998 年 2-3 月人民日报	2238746
3		1998 年 2-4 月人民日报	3679508
4		1998 年 2-5 月人民日报	4921648
5		1998 年 2-6 月人民日报	6166046
6	开放测试语料	1998 年 1 月人民日报	1049414
7	封闭测试语料	1998 年 2 月人民日报	1082238

实验采用两种方法，一种方法是利用 2 阶 HMM 进行标注，另一种方法是利用 2 元 SSD 模型进行标注，然后对结果进行对比分析。

在标注之前首先根据标注语料的标注结果对训练语料与测试语料进行了预处理，将姓名、地名、机构名、数字、时间等进行了归并，所有姓名（不区分姓与名）以“*姓名*”表示，地名以

“*地名*”表示，机构名以“*机构名*”表示，数字以“*数字*”表示，时间以“*时间*”表示，这样处理后可以排除专名识别对于比较不同模型标注性能的影响。

5.2 实验结果

表3 封闭测试结果

	总体正确率	兼类词正确率
2阶HMM模型	95.56%	94.34%
2元SSD模型	97.08%	95.06%
优化幅度	34.23%	12.72%

（注：训练语料为1998年2-6月人民日报，测试语料为1998年2月人民日报）

表4 不同规模训练语料的开放测试总体正确率结果

训练语料规模	1个月	2个月	3个月	4个月	5个月
2阶HMM模型总体正确率	93.73%	94.09%	94.43%	94.59%	94.70%
2元SSD模型总体正确率	94.34%	94.96%	95.33%	95.54%	95.67%
优化幅度	9.73%	14.72%	16.16%	17.56%	18.30%

（注：测试语料为1998年1月人民日报）

表5 不同规模训练语料的兼类词标注正确率结果

训练语料规模	1个月	2个月	3个月	4个月	5个月
2阶HMM模型兼类词标注正确率	89.87%	90.05%	90.56%	91.01%	91.50%
2元SSD模型兼类词对标注正确率	91.36%	91.99%	92.45%	92.89%	93.31%
优化幅度	14.71%	19.50%	20.02%	20.91%	21.29%

（注：测试语料为1998年1月人民日报）

表6 小规模训练大规模测试的结果

训练语料	测试语料	测试语料规模（词）	2阶HMM	2元SSD	3元SSD
1998年1月人民日报	1998年2月人民日报	1082238	93.93%	95.12%	95.14%
	1998年2-3月人民日报	2238746	93.76%	94.94%	94.95%
	1998年2-4月人民日报	3679508	93.61%	94.79%	94.80%
	1998年2-5月人民日报	4921648	93.40%	94.61%	94.62%
	1998年2-6月人民日报	6166046	93.34%	94.53%	94.54%

我们还利用1998年2-3月人民日报语料作为训练语料，以1998年1月份人民日报语料作为测试语料，检验SSD模型在完全稀疏条件下的标注性能。所谓完全稀疏，是指在利用n元SSD模型标注时，不使用n元词表，而只使用1元至n-1元词表，这使得测试语料中出现的所有n元词序列都成为稀疏词序列，这是n元SSD模型可能遇到的最稀疏情况，这时的标注性能可以认为是n元SSD模型的性能底线，测试结果如下：

表7 完全稀疏条件下的SSD模型标注结果对比

	2阶HMM	完全稀疏2元SSD模型	2元SSD模型	完全稀疏3元SSD模型	3元SSD模型
正确率	94.09%	94.09%	94.96%	94.96%	94.97%

从上表结果及错误分析发现，完全稀疏的2元SSD模型标注正确率与2阶HMM的标注正确率等同且错误完全一样；完全稀疏的3元SSD模型的标注正确率与2元SSD模型的标注正确率等同且错误完全一样。这证明了，n元SSD模型对于稀疏数据的处理策略保证了当n增长的情况下，模型不会因为数据稀疏问题而造成性能的降低，反而会随着n的增长，模型的语境观察

范围得到，其性能会得到不同程度的提高。

为了与判别模型在词性标注上的性能进行对比分析，我们选用最大熵模型进行实验，结果如下表。

表 8 最大熵模型标注结果比较

	2 阶 HMM 模型	2 元 SSD 模型	最大熵模型
正确率	94.23%	94.98%	95. 69%
训练时间（秒）	52	73	16560
标注时间（秒）	207	478	659

上述结果说明，在当前的训练规模条件下，最大熵模型的标注正确率要高于 SSD 模型和 HMM 模型，显示出判别模型在利用语境信息方面的优势，但是最大熵模型的训练时间远高于其他两种模型，而 SSD 模型的训练时间虽多于 HMM 模型，但是二者相差不过 20 秒左右，基本在同一个数量级上，且 SSD 模型的标注正确率高于 HMM，尽管低于最大熵模型，但其保持了 HMM 简单快捷的优势，又较 HMM 的标注正确率有较大幅度提高，具有一定的实用价值。

6. 讨论与展望

SSD 模型训练复杂度较判别模型要低，解码速度较快，因此能够更灵活方便地根据实际需求，迅速训练并提供所需语言模型，同时，SSD 模型还克服了 HMM 模型的强独立性假设的不足，能够利用更多的观察信息，保证较高的标注正确率。同时，SSD 模型也并非绝对不能够利用语境中的其他特征信息进行状态判断，而是有可能将其他有用信息也集成到模型之中，因此，我们下一步的工作将重点将研究如何将丰富的语境信息合理地集成到 SSD 模型之中，使其性能得到进一步提高。

参 考 文 献

[1] Daniel Jurafsky, James H. Martin. Speech and Languge Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. USA:Prentice Hall, 2000

[2] Doug Cutting, Julian Kupiec, Jan Pedersen, Penelope Sibun. A Practical Part-of-Speech Tagger. In Proceedings of the Third Conference on Applied Natural Language Processing, 1992:133-140.

[3] Adwait Ratnaparkhi. A maximum entropy model for Part-of-speech Tagging[C]. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 1996:133-141.

[4] 梁以敏，黄德根 基于完全二阶隐马尔可夫模型的汉语词性标注[J]. 计算机工程, 2005, 05

[5] 屈刚，陆汝占 一个改进的汉语词性标注系统[J]. 上海交通大学学报，2003，06

[6] 洪铭材，等 基于条件随机场(CRFs)的中文词性标注方法[J]. 计算机科学, 2006,33（10）

[7] 姜维，等 基于条件随机域的词性标注模型[J]. 计算机工程与应用，2006, 21