汉语词性信息对句法分析的贡献

——基于信息论的统计研究

提　要

本文中，我们利用了宾州树库（Penn Treebank）中文部分的内容，将其进行多种词性标准下的词性标注,并衡量词性信息对句法成分决策的贡献。以上考察从统计学的角度上验证了汉语词法信息和句法信息的差异,表明了词性信息对句法分析的贡献：有贡献但是较为有限，而且主要集中在最内层句法结构中,摈弃在特定若干词性上几乎没有作用。

关键词：词性 句法成分 信息熵 信息贡献 信息论

汉语词性标注是中文信息处理的基石性工作。而句法分析则为诸多上层应用和研究提供支撑。词性和句法成分两者可以视作语料库中的两种重要信号。前者携带信息量的大小和对后者决策的信息贡献将为句法分析效果的提升提供指导，这也是本文考察的重点。

1. 信息熵与“词性-句法结构对”系统

熵（Entropy）是热力学中衡量一个封闭系统内部复杂程度的物理标量。在信息论中信息熵（Information Entropy, IE）被用来描述传播中信息的不确定性。当一个信号具有较大量的信息，则其熵值较低,低信息量的信号则反之。信息熵的度量单位是比特（bit），其计算公式如下[1]：

$$IE = -\sum p(x_i)\log_2(p(x_i))$$

其中 $p(x_i)$为信号的一个取值出现的概率。在计算语言学常见的应用中，如果以语料库为一个封闭系统，该语料库的字集或词集可以视为一种信号，则出现的每个字或词为就为该信号的一个取值。

在本文的工作中我们希望通过信息熵来考察词性对句法分析的信息贡献,则需要考虑到词性和句法结构两个因素。因而我们设计了“词性-句法结构对”这个系统。可以形式化的表示为<POS, CHUNK>这样的二元组。其中 POS 为词性（Part Of Speech)，如形容词、基本名词等。CHUNK 代表一个词所属于的句法结构（syntax chunk），如名词性短语或者动词性短语等。这里我们将句法结构视作信号 CHUNK。显然，对于语料库中一个特定词性的词，它所隶属的句法结构就是“CHUNK”信号的一种取值。

对于一种词性，标注为该词性的所有词，在一个确定的语料库中，共计多少次隶属于多少种句法结构就可以帮助计算其信息熵。本文的工作使用最大似然估计来获得一个 CHUNK 信号取某一值的概率。

$$p(x_i) = \frac{f(x_i)}{\sum f(x_i)}$$

其中 x_i 为 CHUNK 信号的一种取值，f(x)是取值为 x 的词在语料库中出现的个数。

对于一个特定语料，在确定的词性标注标准下的每个词性都有一个“词性-句法结构对”系统。每个系统也都有自己的熵值。按照信息熵的数学表达，当熵值较小时，表明该词性在句法结构中的分布不平均。其倾向于集中出现在某种或某几种句法结构中。而当熵值较大的时候，表明该词性在不同的句法结构中分布很平均（或者说很杂乱），即通过该词性很难判断被标注为该词性的词（或符号）可能属于什么句法结构。以标点符号为例。在许多词性标注标准中，标点符号是一种词性，通常被标注为符号“w”。如果标点符号平均出现在各个句法结构中，则标点符号的存在本身就不能成为判断一个包含标点符号的字符串属于什么样句法结构的依据，即它为句法分析所提供的信息量少（这是根据信息熵的物理含义说的）。也就可以进一步认为一个符号是否被正确的标注为标点符号，对句法分析的结果而言并不重要。

2. 内层信息熵分析实验

本文实验中所选择的语料是宾夕法尼亚大学标注的句法树库（Penn Tree Bank, 简称宾州树库）的中文部分 3223 个句子[2]，其全部来自于新华社新闻。在实验中我们共选取了四种词性标注标准：北京大学计算语言学研究所词性标注标准（后文简称北大标准）[3]、中国科学院计算技术研究所词性标注标准（后文简称计算所标准）[4]、宾州树库标准和宾州树库英文标准。其中本文工作中采用的英文树库为中文树库的句对齐翻译版，并且按照英文词性标准和句法标准进行了标注。英文语料将成为实验中的外语参照组。

由于信息熵本身的绝对值大小并无参照意义，实验中除去外语对照组，我们还设定了随机对照组，模拟“随机极端情况”，以作为信息量大小的参照。随机对照组通过对语料库中的词进行某种标准下的随机标注来实现（句法树并未改变，只有词性是随机标注）。其信息熵为“随机词性－句法对”系统的熵值。这是一种极端情况：显然随机标注词性和不标注词性是一样的。表 1 为各词性标准下参照组的平均熵值。

还是以标点符号（北大标准中的标注符号为“w”）为例。我们普遍认为标点符号对于句法分析没有特别的作用，而在统计中其信息熵为 2.576bit，而对同一份语料中的标点符号进行随机词性标注后的熵值为 2.872bit。两者十分接近，也就是说标点符号对句法分析的信息贡献很低。

词性标准	随机对照组的平均熵值（bit）
北大标准	2.849

计算所标准	2.831
宾州树库标准	2.851
宾州树库英文标准	2.849

表 1. 随机对照组的平均熵值

由于句法分析的结果是句法树。本部分实验里“词性-句法结构对”中的句法结构为被标注词所隶属的最内层句法结构。如下例[1]中“遍布”一词，其词性为其他动词（宾州树库标准，VV），其所隶属的最内层句法结构为动词短语（VP）。在这里我们使用的是宾主树库的句法功能标记作为 CHUNK 信号的值。如 NP-SBJ 和 NP 在实验中被视为不同的结构。

例 1.

```
( (IP (IP (NP-SBJ (NN 建筑))
          (VP (VC 是)
              (NP-PRD (CP-APP (IP
(NP-SBJ (-NONE- *pro*))
                  (VP (VV 开发)
                      (NP-PN-OBJ (NR
浦东))))
                  (DEC 的))
                  (QP (CD 一)
                  (CLP (M 项)))
                  (ADJP (JJ 主要))
                  (NP (NN 经济)
                  (NN 活动)))))
          (PU ，)
          (IP (NP-SBJ (-NONE- *pro*))
              (VP (DP-TMP (DT 这些)
                  (CLP (M 年)))
                  (VP (VE 有)
(IP-OBJ (NP-SBJ (NP (QP (CD 数百)
                  (CLP (M 家)))
                  (NP (NN 建筑)
                  (NN 公司)))
                  (PU 、)
                  (NP (QP (CD 四千余)
                  (CLP (M 个)))
                  (NP (NN 建筑)
                  (NN 工地))))
              (VP (VV 遍布)
                  (PP-LOC (P 在)
                  (LCP (NP (DP (DT 这)
                  (CLP (M 片)))
                  (NP (NN 热土)))
                  (LC 上))))))))
          (PU 。)) )
```

[1] 该例内容为“建筑是开发浦东的一项主要经济活动，这些年有数百家建筑公司、四千余个建筑工地遍布在这片热土上。”

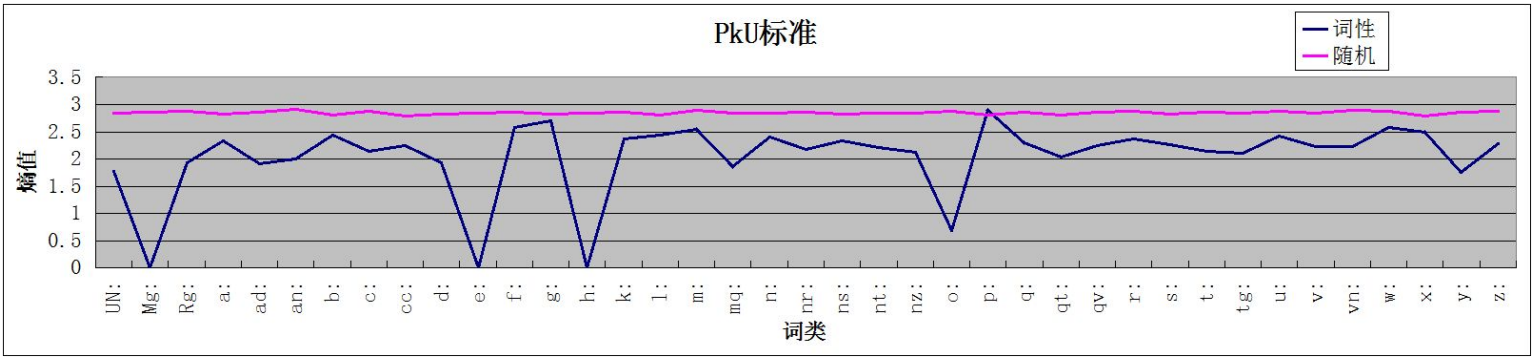

图 1. 北大标准下内层实验信息熵

表 2.为北大标准（PkU 标准）中介词（符号为“p”）在各种句法结构中的出现情况。我们可以观察到其在众多句法结构中都有出现，而且高频的若干种结构中出现频次相差不大。因而借由其词性本身预测其句法隶属是十分困难的，即对句法分析的信息贡献不大。

如果考察所有词性，则图 1 为使用北大标准（PkU 标准）对文本进行词性标注后度量“词性-句法成分对”信息熵的结果。对照组（粉色曲线）为使用北大标准对文本进行随机标注后“随机词性-句法成分对”的信息熵。通过上图我们可以发现许多词性在句法成分中出现的情况接近甚至达到了这种“随机极端情况”。如语素（符号为“g”）、介词（符号为“p”）、数词（符号为“m”）等。这表明这些词性本身对预测或判别其所在符号串的句法结构是帮助很小，甚至是没有帮助的。

句法结构	频次	句法结构	频次	句法结构	频次
PP	559	CP	17	IP-HLN	1
VP	538	FRAG	16	NP-ADV	1
NP	508	NP-PN-OBJ	15	NP-TTL-SBJ	1
PP-LOC	506	PP-PRD	10	UCP-APP	1
PP-DIR	272	QP-PRD	10	NP-TTL-OBJ	1
PP-TMP	242	QP-OBJ	10	LCP-OBJ	1
PP-MNR	220	IP-OBJ	9	PP-TTL	1
NP-PN	216	NP-APP	7	IP-Q	1
ADVP	133	QP-EXT	7	NP-PN-IJ	1
NP-SBJ	84	VCD	6	NP-OBJ-TTL	1
IP	83	NP-PN-APP	6	QP-ADV	1
PP-PRP	63	VCP	4	NP-TTL-PN	1
CLP	60	NP-PN-TPC	4	NP-TPC	1
QP	57	VP-PRP	4	DP-TMP	1
PP-BNF	53	CP-APP	4	NP-IO	1

PP-ADV	50	LCP-TMP	4		
NP-PN-SBJ	49	DVP	3		
NP-OBJ	44	NP-PN-LOC	3		
NP-TMP	40	NP-TTL	3		
ADJP	36	IP-PRD	3		
PP-LGS	36	NP-PN-IO	2		
DP	31	NP-PRD	2		
DNP	27	PRN	2		
LCP	23	PP-EXT	2		
VRD	19	NP-LOC	2		

表 2. 介词在各种句法结构中的分布情况

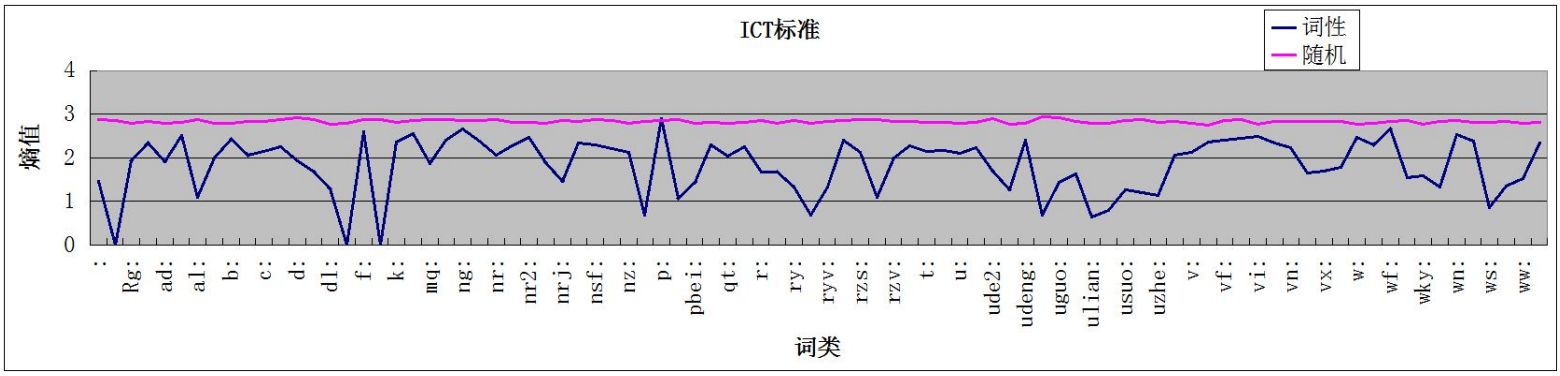

图 2. 计算所标准下内层实验信息熵

类似的使用计算所词性标准（ICT 标准）对文本进行词性标注之后，我们获得的“词性-句法成分对”的信息熵如上图所示。计算所标准的词性类别较多较细（共 99 个符号），上图仅部分显示。和北大标准下的工作相同，粉色曲线为随机标注的对照组。在该实验中，介词（符号为“p”）的信息熵值为 2.903bit。这和随机对照组的熵值几乎一致。而方位词（符号为“f”）、处所指示词（符号为“rzs”）、省略词（符号为“ws”）和不及物动词（符号为“vi”）的熵值也都接近了“极端随机情况”。

可见在句法成分的预测和判别中，计算所标准和北大标准的表现基本一致，并无太大差别。另外我们还考察了宾州树库自己的词性标注情况，其熵值如下图所示。

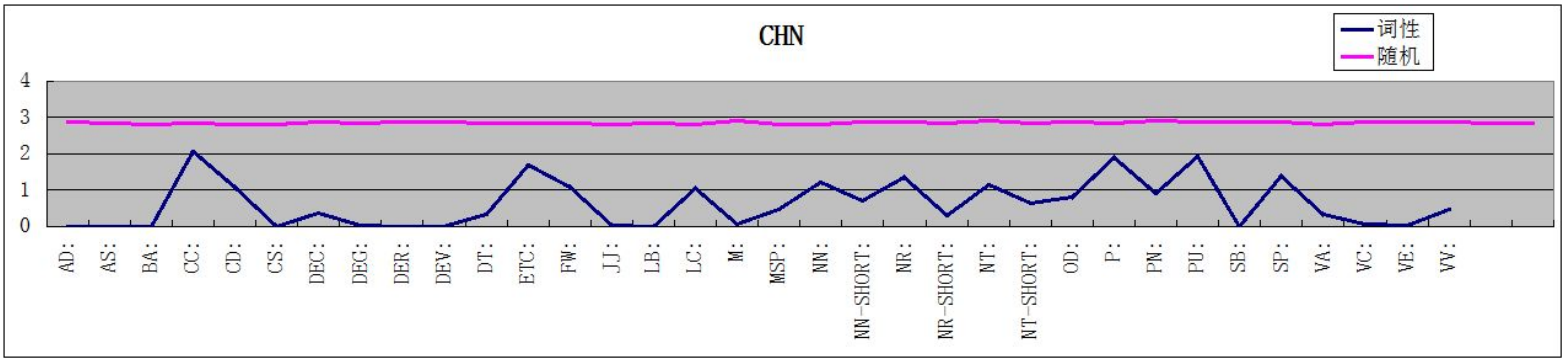

图 3. 宾州树库中文词性标注下内层实验信息熵

可以看到，这种情况和前面的情况有非常大的不同。其最高熵值出现在并列连词（符号为“CC”），仅为 2.069bit。我们可以说宾州中文树库自己的词性标注体系更加适合于自己的句法分析结果，即对句法成分，其词性标注结果有着更好的预测和判别能力。

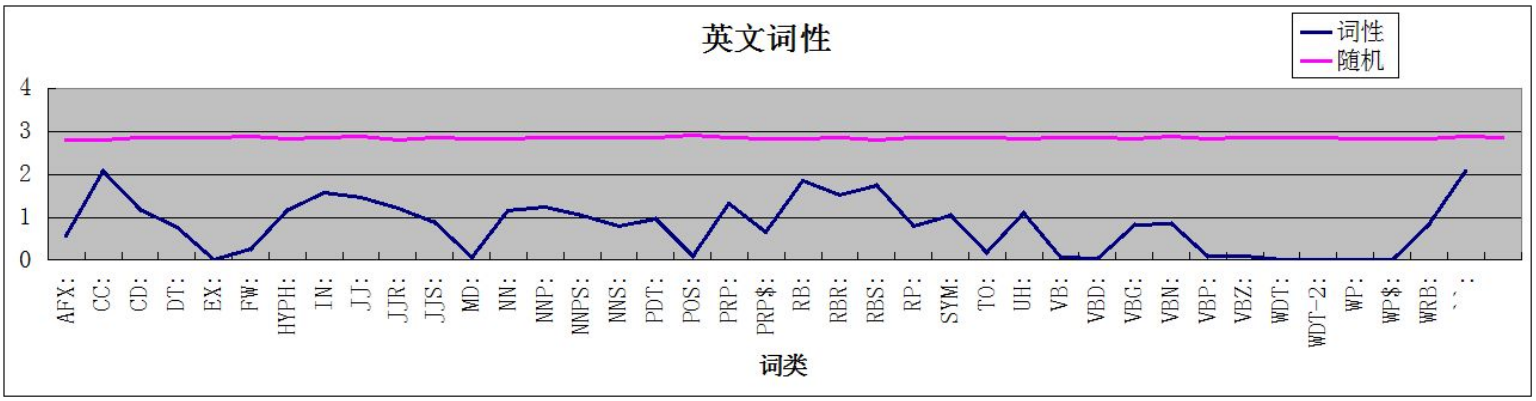

图 4 . 宾州树库英文词性标注下内层实验信息熵

类似的，我们也考察了同为宾州树库序列的宾州树库英文版。其“词性-句法成分对”的结果与图 3 相仿。最高的熵值出现在并列连词（符号为“CC”）处，2.076bit，和标点符号中的双引号（符号为“ ”），2.066bit。其余词性的熵值均低于 2bit。

我们就四种词性标注标准里超过随机对照组熵值 80%的词性种数进行了统计，如表 3 所示（每个词性系统超过对照组在该词性上随机标注熵值的 80%而非超过对照组平均值的 80%）。可以看出北大标准、计算所标准与宾州树库中英文标准之间在预测句法结构方面存在着较大的差异。并且其预测力相对较差。

词性标准	超阈值词性种数	占词性标准
北大标准	17	43.6%

计算所标准	27	27.3%
宾州树库标准	0	0
宾州树库英文标准	0	0

表 3. 四种词性标注中高熵值词性的种数（内层）

综合以上的数据，我们倾向于认为并非词性标注本身不适应于句法分析，或者更激进的说词性划分不适用于汉语，而是不同的词性系统可能对于不同的句法分析系统有其自身的特异性。从内层实验中我们还可以看出词性标准对句法分析所能提供的信息是有限的。在有的标准中特定若干词性对句法结构的预测几乎没有作用。

而在北大标准和计算所标准中共同的高熵值词性有 13 种，分别是形容词、区别词、方位词、后缀、数量词、普通名词、地名、介词、非词字符串（如网址）和标点符号。因为计算所标准和北大标准的渊源关系，它们共享了大部分的大词性。而北大标准没有子词性。计算所标准中有 27 种高熵值词性，和北大标准的高熵值词性重合的有 13 类，不足一半。则我们可以认为高熵值系统多出现在某一大类下的子类中。因而我们倾向于较细的词性划分标准有助于降低高熵值词性的比例，提高对句法分析工作的信息贡献。

3. 二层信息熵分析实验

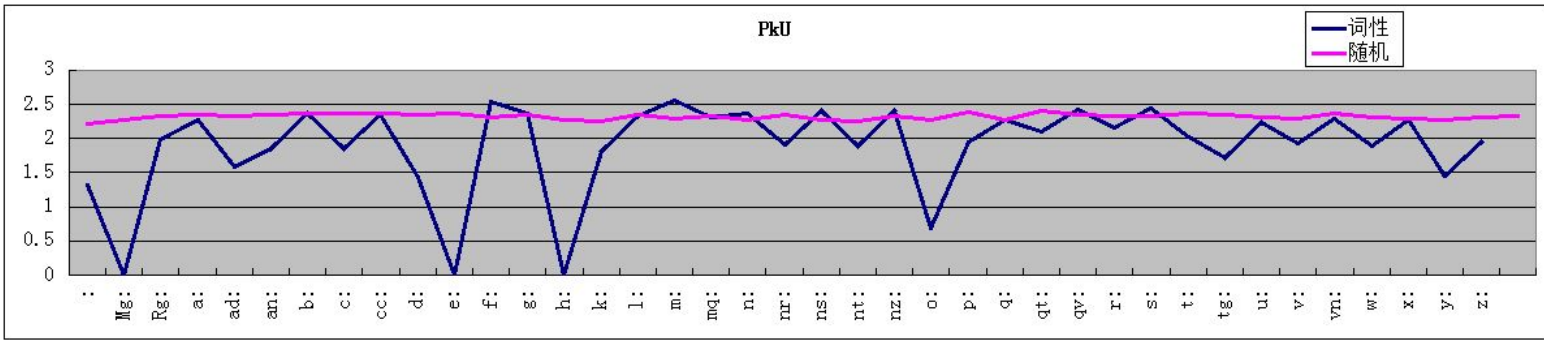

图 5 . 北大标准下二层实验信息熵

前文实验为“词性-句法成分对”的最内层分析，即我们只分析了该词所属的最内层句法结构。本部分中我们考察了被考察词所属的第二层句法结构（如果存在的话）。如例 1 中“数百”一词的二层句法结构为名词性短语（NP）。“词性-句法结构对”系统的信息熵计算和前文所述一致。这一部分也以同样的方法设置了“随机极端情况”的对照组。其平均熵值如表 4 所示。

词性标准	随机对照组的平均熵值（bit）
北大标准	2.323
计算所标准	2.305
宾州树库标准	2.324
宾州树库英文标准	2.323

表 4. 二层句法结构随机对照组的平均熵值

图 5 是北大标准下的二层句法结构的熵值情况。可以看出有更多词性的熵值达到或接近了随机参照组的水平：如普通名词（符号为“n”）、地名（符号为“ns”）和形容词（符号为“a”）等。类似的现象也出现在计算所标准下，如图 6 所示。

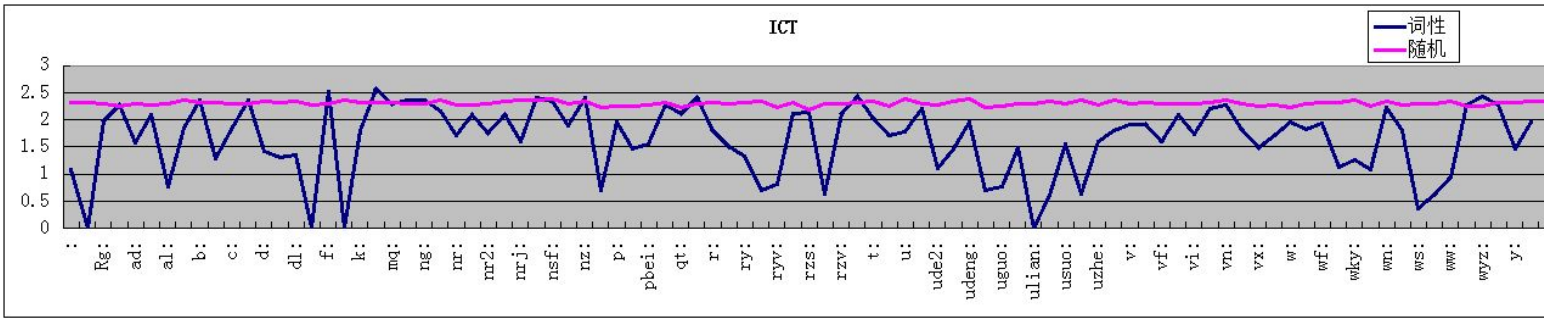

图 6. 计算所标准下二层实验信息熵

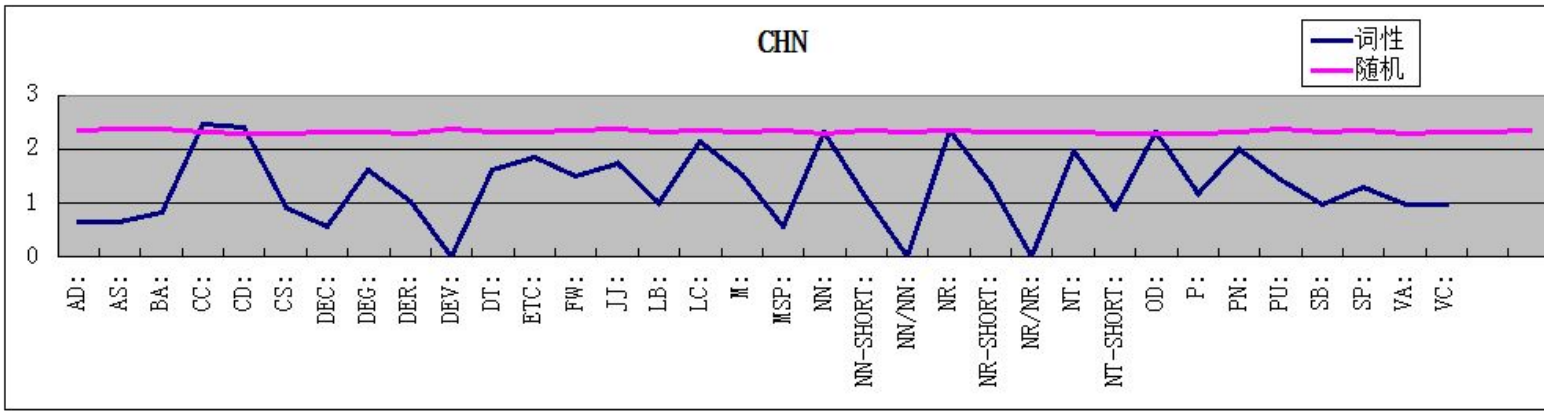

图 7 . 宾州树库中文标准下二层实验信息熵

再次观察宾州树库自身词性系统在二层实验中的表现。其状况和北大标准下的情况已经十分接近。很多词性的信息熵也已经接近并达到了随机对照组的数

值，如定位词（符号为“LC”）、普通名词（符号为“NN”）和序数词（符号为“OD”）等。

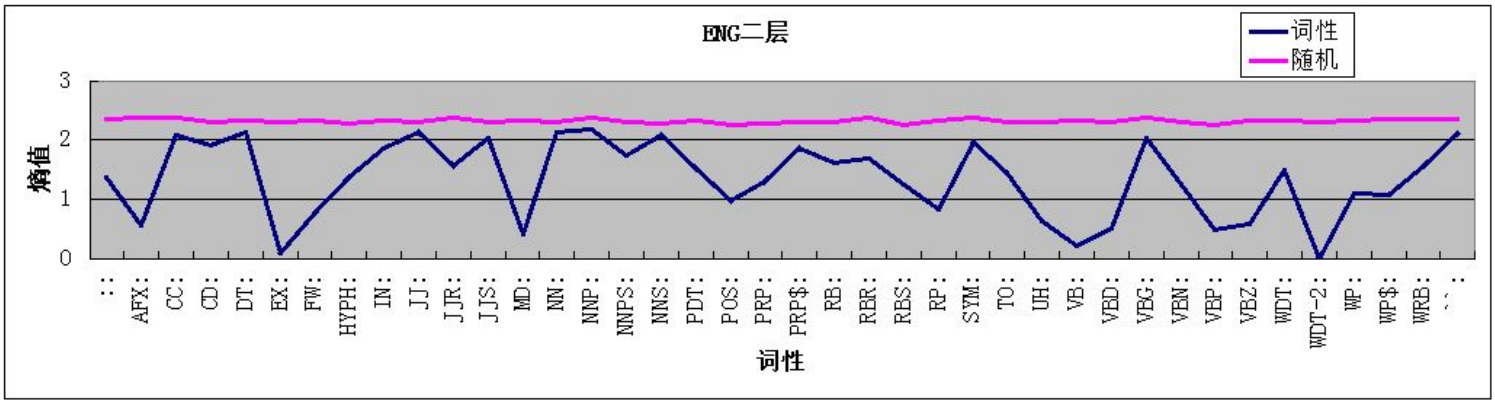

图 8．宾州树库英文标准下二层实验信息熵

英语组的情况也十分相似，包括形容词（符号为“JJ”）和普通名词（符号为“NN”）在内的许多词性都接近了对照组的完全随机状态。综合四种词性标准在二层实验中的数据，同样可以考察超过各自随机对照组熵值 80%的词性种类（如表 5 所示）。

词性标准	超阈值词性种数	占词性标准
北大标准	27	69.2%
计算所标准	42	42.4%
宾州树库标准	8	22.2%
宾州树库英文标准	13	32.5%

表 5. 四种词性标注中高熵值词性的种数（二层）

可以看出宾州树库自身词性系统和北大/计算所系统在“词性-句法成分对”上的信息熵的差异已缩小了不少。词性标注对句法结构的信息贡献进一步减小，因此我们倾向于认为，词性对句法结构预测和判断的信息贡献不仅有限而且局限于最内层中。

4. 结论与展望

本文通过构造“词性-句法结构对”系统实现了词性对句法分析基于信息论的信息贡献度量。通过对最内层和二层句法结构的考察，我们认为词性信息对句

法分析有信息贡献，但较有限。一些词类，尤其是重要词类如普通名词和形容词、地名等对句法分析几乎没有信息贡献。而它们在中文词性标注中是较难的课题。因而建议进一步提高中文句法分析的准确性不必要在提高词性标注方面做过多的努力。

表 2 和表 4，以及图 7、8，图 3、4 的对比显示词性标准在中英文语料上的信息贡献差别不大，因而不能激进的认为词性划分不适合对汉语的分析。

通过北大标准和计算所标准的比较，可以看出划分较细的标注标准可以为上层应用提供更好的信息贡献。而这也是符合我们的一般直觉的。内层和外层句法结构实验的比较使我们有理由猜测层次越深，词性信息的作用可能越少。这一点还需要更多层次的句法分析来进行验证。而为了克服数据稀疏，这样的实验无疑需要更大规模的树库资源支持。

注意到本文的实验中使用的是宾州树库的“句法-功能”混合标记而非单纯句法标记。相信在后者标记下也可以得到类似结论。更细致的具体实验和分析可以留待未来工作中做进一步的探索。

参考文献

[1]Shannon, C.E. (1948), "A Mathematical Theory of Communication", Bell System Technical Journal, 27, pp. 379–423 & 623–656, July & October, 1948.

[2]Marcus M P, Marcinkiewicz M A, Santorini B. Building a large annotated corpus of English: The Penn Treebank[J]. Computational linguistics, 1993, 19(2): 313-330.

[3]俞士汶, 段慧明, 朱学锋, 等. 北大语料库加工规范: 切分• 词性标注• 注音[J]. 汉语语言与计算学报, 2003, 13(2): 121-158.

[4]Zhang H P, Yu H K, Xiong D Y, et al. HHMM-based Chinese lexical analyzer ICTCLAS[C]//Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17. Association for Computational Linguistics, 2003: 184-187.