基于三元统计模型的汉语分词及标注一体化研究

高山　张艳　徐波　宗成庆　韩兆兵
中国科学院自动化研究所模式识别国家重点实验室 北京 100080
E-Mail: {sgao, yzhang, xubo, cqzong, zbhan}@nlpr.ia.ac.cn

摘要：汉语的分词及词性标注是汉语语言处理的基础。虽然，该领域在过去十年已经有了很大进展，但高精度的面向大规模真实文本的分词及标注仍然存在许多困难。本文提出一种基于三元统计模型的汉语分词标注的方法，旨在并行考虑词性及词汇的三元概率模型，兼顾词及词性之间的搭配，实现分词和 78 类二级词性标注的整体最优，实验结果显示该方法获得很高的正确率。

关键词：分词，词性标注，三元统计模型

The Research on Integrated Chinese Words Segmentation and Labeling based on Trigram Statistical Model

Shan GAO, Yan ZHANG, Bo XU, ChengQing ZONG, ZhaoBing HAN,
National Lab of Pattern Recognition
Institute of Automation,Chinese Academy of Sciences,Beijing 100080
E-Mail: {sgao, yzhang, xubo, cqzong, zbhan}@nlpr.ia.ac.cn

Abstract: The Chinese words segmentation and labeling are basis of the Chinese language processing. Although many progress in this field have been made, there are still many difficulties in dealing precisely with the large amount real text. In this paper, the method based on trigram statistical model is prescribed. It aims to calculate simultaneously the trigram statistical models of both language and part-of-speech, taking their relations into account, to simultaneously segment and label with 78 second types. The outcome with high accuracy is gained from experiments.

Key words: Segmentation, Part-of-Speech tagging, Trigram statistical model

1. 引言

分词和词性标注是语料库加工的重要组成部分，随着计算机对大量文本处理的需要，对分词及词性标注的正确率的提高也日益迫切，并且这方面的研究也正在广泛地进行

[2,3,4,5,16]。分词主要问题是在歧义点的处理[1]。如“他||将||来||北京”错误地分成“他||将来||北京”，等等。此外，文本中人名，地名等专有名词应作为一个词，但实际中常常会被错误地划分开来。

词性标注的主要问题是词性兼类[2]。词性标注的难点就是兼类词的消歧。汉语缺乏形态变化，词的应用非常灵活，可以充当不同的句子成分，因此汉语的词类兼类特别多，特别复杂。自然语言中的词性兼类是普遍存在的现象。我们对汉语词类兼类现象进行了统计，兼类词在词典中占总词数的比例为 10.88%，而兼类词在语料库文本中所占总词次的比例为 25.76%。结果表明尽管词典中兼类词数量所占比例不算太高，但是在语料文本中出现次数却不低。这说明许多常用词是兼类词。

关于汉语的词性标注的研究，主要的方法有：（1）基于规则的方法。Greence 和 Rubin 开发的 TAGGIT 标注系统，利用 3300 多条上下文规则，对百万次的 BROWN 语料进行标注。90 年代以来，另一种新的基于规则的词性标注系统采用了 Brill 方法。这种方法使用基于转换的错误驱动方法进行词性标注处理。（2）基于统计的方法。80 年代初，Mashall 提出了 LOB 语料库标注算法 CLAWS，其算法的时间复杂度是指数级的。DeRose 等人对统计方法进行了改进，设计了 VOLSUNGA 标注系统。（3）混合方法，北京大学提出了一种先规则后统计的规则和统计相结合的标注算法[15,17]。哈尔滨工业大学把置信区间的方法引入到词性标注中[13,16]。

本文系统地描述了中科院自动化所在参加“973”项目评测时研制的分词和标注系统，该系统在用大量真实文本测试时得到了满意的结果。本系统预先从训练集中获取语法规则和单词的频度信息，及三元语言模型，实现高精度的分词和标注。本文第二节描述了三元语法及语言统计模型；第三节描述了分词和标注的整体最优方法[18][19]；然后给出实验结果及对结果的分析。

2. 分词和词性标注统计模型

由于汉语中字和字之间没有分隔，汉语分词就是每个相连的孤立字之间的前后划分。而文本中的标点符号可作为此处的天然分割。利用这些标点符号，整个文本被标点分割为若干条短句。设短句 S 是由单词串 $W=w_1,w_2,\ldots\ldots,w_n$ 组成，词 w_i 的词性为 t_i，则该短句相应的词性标注可表达为 $T=t_1,t_2,\ldots\ldots,t_n$。在 S 所对应的各种分词和标注形式，寻找 T 和 W 的联合概率 $P(W,T)$为最好的词切分和标注的组合。

2.1 基于三元词性统计模型分词标注

如应用隐马尔可夫链[6]来描述一个完整句子中的词性的变化，每种词性对应一种状态，状态的转移概率代表词性之间的搭配关系。这样，在生成一个句子时，系统不断地由

一个状态转移到另外一个状态，每一状态都产生一个输出，由此得到句子中相应的每个词，直至整个句子输出完毕。

由隐马尔可夫模型可近似将 $P(W,T)$简化表示为：[7]

$$P(W,T)= P(W/T)\ P(T) \approx \prod_{i=1}^{n} p(w_i \mid t_i)\, p(t_i \mid t_{i-1},t_{i-2}) \qquad ①$$

$P(w_i \mid t_i)$的意思是，在整个语料中，在词性 t_i 的条件下，单词 w_i 出现的概率；$p(t_i \mid t_{i-1},t_{i-2})$的意思是前两个单词的词性是 t_{i-1}，t_{i-2}，当前词的词性是 t_i 的概率。

在切分句子时，列出所有可能的分法，用词典中单词出现的概率和语法规则中词性和词性的连接概率，计算每个切分方法的概率总和，概率值最大的一个即为输出结果。

为了获取 $p(w_i \mid t_i)$及 $p(t_i \mid t_{i-1},t_{i-2})$，事先以大规模的汉语标注语料库做训练，抽取词性的语法规则和单词的频度信息。对已经分词和标注过的语料统计每个词的不同词性出现的次数 $C(w_i,\ t_i)$ 及每种词性 t_i 在文本中出现过的总次数 $C(t_i)$，根据最大似然估计算法计算得：

$$p(w_i \mid t_i) = C(w_i,\ t_i) / C(t_i)$$

类似地，以三元为单位，统计文本中的前后词性组合的频度 $C(t_{i-2},\ t_{i-1})$及 $C(t_{i-2},\ t_{i-1},\ t_i)$，计算 $p(t_i \mid t_{i-1},t_{i-2})$：

$$p(t_i \mid t_{i-1},t_{i-2}) = C(t_{i-2},\ t_{i-1},\ t_i) / C(t_{i-2},\ t_{i-1})$$

分别将 $p(w_i \mid t_i)$及 $p(t_i \mid t_{i-1},t_{i-2})$存放在一专门文件中。考虑到数据的稀疏问题，需要规模较大的语料来做训练，我们进行了平滑处理，即对于训练中不存在的三元词性组合，也就是当 $C(t_{i-2},\ t_{i-1},\ t_i)$为零时，给 $p(t_i \mid t_{i-1},t_{i-2})$赋一个很小的值。

2.2 基于三元语言统计模型分词及标注

采用三元的语言模型就是采用统计的方法来揭示语言的内在规律。假设第 n 个词出现与前面的 $n-1$ 个词相关，则整句的概率就是各个词出现概率的乘积。采用三元语言模型[6][8]，公式 $P(W, T)$的另一个简化变形为：

$$P(W,T)= P(T \mid W)\ P(W) \approx \prod_{i=1}^{n} p(t_i \mid w_i)\, p(w_i \mid w_{i-1},w_{i-2}) \qquad ②$$

$p(w_i \mid w_{i-1},w_{i-2})$是采用三元文法的语言模型。因为$(w_i ,w_{i-1},w_{i-2})$的组合种类很多，所以要求用大量的语料训练 $p(w_i \mid w_{i-1},w_{i-2})$。由于数据稀疏的原因，在 $p(w_i \mid w_{i-1},w_{i-2})$ 无法直接求得时，则回退求 $p(w_i \mid w_{i-1},)$ 直至求 $p(w_i)$ 来计算代替三元概率[8]。$p(t_i \mid w_i)$反映的是每个词最大可能出现的词性。$p(w_i \mid w_{i-1},w_{i-2})$对分词有很大的贡献，但当分词确定后，它对词性标注无任何影响。而 $p(t_i \mid w_i)$并没有体现前后词性的搭配关系，只是说明该词最大可能的词性。实验结果也表明公式②比公式①有更强的约束能力，但公式②对分词却有很大的贡献。

3. 分词及词性标注的整体最优

应用三元语言统计模型和三元词性统计模型，即第二节中的公式①和②，综合语言及词性统计特性，一种实现分词和标注整体最优的公式为：

$$P^*(W,T)=\alpha \prod_{i=1}^{n} p(w_i \mid t_i)\, p(t_i \mid t_{i-1},t_{i-2}) + \beta \prod_{i=1}^{n} p(t_i \mid w_i)\, p(w_i \mid w_{i-1},w_{i-2})$$

从语言模型得到的结果分析可知，$p(t_i|w_i)$对分词无帮助，且在分词确定后对词性标注又会增添偏差。故取②中的分词部分，舍弃标注部分，可有

$$P^{**}(W,T)=\prod_{i=1}^{n} p(w_i \mid t_i)\, p(t_i \mid t_{i-1},t_{i-2}) + \lambda \prod_{i=1}^{n} p(w_i \mid w_{i-1},w_{i-2}) \qquad ③$$

其中 λ 为加权系数。公式③与公式①相比，在标注方面无任何改变，但是 $\lambda \prod_{i=1}^{n} p(w_i \mid w_{i-1},w_{i-2})$ 进一步提高了分词的正确性。由于在实际计算中公式③中前一项 $\prod_{i=1}^{n} p(w_i \mid t_i)\, p(t_i \mid t_{i-1},t_{i-2})$ 与后一项 $\lambda \prod_{i=1}^{n} p(w_i \mid w_{i-1},w_{i-2})$ 的数值大小相差很大，用 λ 平衡二者关系。可以根据词性 t 的种类个数及词典中词 w 的个数，调节 λ。一般 λ 取词典中词 w 的个数 / 词性 t 的种类个数。

这样，实现分词及词性标注整体最优就是同时找到了词序列 $w_1w_2\ldots\ldots w_n$ 及词性序列 $t_1t_2\ldots\ldots t_n$ 使 $\prod_{i=1}^{n} p(w_i \mid t_i)\, p(t_i \mid t_{i-1},t_{i-2}) + \lambda \prod_{i=1}^{n} p(w_i \mid w_{i-1},w_{i-2})$ 最大，即 $max(P^{**}(W,T))$ 的寻找过程。

例如，对句子“她出生在辽宁”做切分，从词典中词的词性属性看，可以有多种切分结果。我们根据 $p(w_i \mid t_i)$，$p(w_i \mid w_{i-1},w_{i-2})$ 及 $p(t_i \mid t_{i-1},t_{i-2})$ 计算每条路径的概率，可得出概率最大的一条路径为：“她[代]出生[动]在[介]辽宁[地名]。”其切分过程如图 1 所示：

为了提高精度和效率，首先对真实文本进行一次预处理，对文本中的姓名，地名，及数字串分别识别出来。应用统计方法[9][10]，对汉语中常见的姓氏、人名进行统计，并考虑人名的前后词语的常见的搭配（比如，人名后经常接称呼，身份，职业等等）。同样，再进一步把文本中的地名，及数字串分别识别出来，并用特殊符号分别给与标注，在对文本做了初级切分后，再利用公式③，结果会有显著的提高。

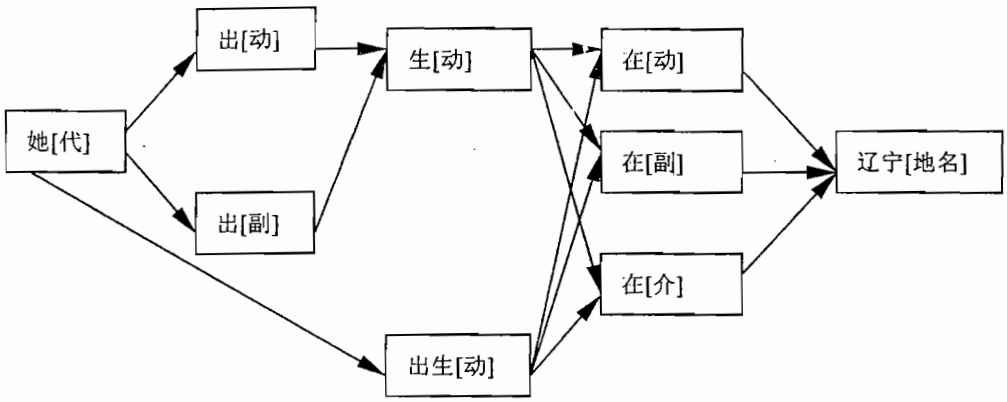

图 1 例句切分过程及结果

4. 实验结果及分析

采用近 50000 个词的词典做实验，以人民日报为语料进行切分和标注。我们将汉语词性分成 19 大类（一级标注），78 个小类（二级标注）[11]，采用 13M 已经标注好的语料训练统计 $P(w_i|t_i)$及 $p(t_i|t_{i-1},t_{i-2})$。经过对待切分的文本进行预处理后，分别采用公式①及公式③针对不同长度的文本进行来集内和集外两组不同的实验：实验中包含文本 1－7，分别对应集内的 1284，4265，9681 个词的文本和集外的 719，4644，5627，13166 个词的文本。实验结果的对照表如下所示：

实验 1：采用公式①，未加入语言模型：

训练集集内分词平均正确率为 97.78%。在分词正确的基础上，二级标注平均正确率为 93.24%，一级标注平均正确率为 96.33%。

待测文本	文本长度（词的个数）	分词正确率（%）	二级标注正确率（%）	一级标注正确率（%）
文本 1	1284	97.8	93.2	96.5
文本 2	4265	98.2	93.8	96.8
文本 3	9681	97.6	93.0	96.1

表 1 未加入语言模型的集内结果

训练集集外分词平均正确率为 96.79%。二级标注平均正确率为 93.10%，一级标注平均正确率为 96.32%。

待测文本	文本长度	分词正确率	二级标注正确率	一级标注正确率

	（词的个数）	（%）	（%）	（%）
文本 4	719	97.3	93.1	96.7
文本 5	4644	97.1	92.9	96.1
文本 6	5627	96.9	93.3	96.5
文本 7	13166	96.6	93.1	96.3

表 2　未加入语言模型的集外结果

实验 2：添加语言模型应用公式③

这里采用 110M 语料训练 $p(w_i | w_{i-1},w_{i-2})$。与实验 1 采用相同的语料做测试。

训练集集内分词平均正确率为 99.48%。二级标注平均正确率约为 93.21%，一级标注平均正确率约为 96.28%。

待测文本	文本长度（词的个数）	分词正确率（%）	二级标注正确率（%）	一级标注正确率（%）
文本 1	1284	99.5	93.2	96.6
文本 2	4265	99.2	93.7	96.6
文本 3	9681	99.6	93.0	96.1

表 3　加入语言模型的集内结果

训练集集内分词平均正确率为 98.06%。二级标注平均正确率约为 93.07%，一级标注平均正确率约为 96.32%。

待测文本	文本长度（词的个数）	分词正确率（%）	二级标注正确率（%）	一级标注正确率（%）
文本 4	719	99.1	93.2	96.8
文本 5	4644	98.3	92.7	96.1
文本 6	5627	97.9	93.3	96.5
文本 7	13166	98.0	93.1	96.3

表 4　加入语言模型的集外结果

实验发现分词出错主要集中在前后词歧义点上。如：“南京市长江大桥”中会出现“南京||市长”的错误。比较两组实验可以看出，加入语言模型后分词正确率有了明显的提高，尤其在训练集内效果更加明显，集内从 97.78%提高到 99.48%，集外从 96.79%提高到 98.06%。

5. 结束语

词性标注是语言信息处理的基础，为汉语句法分析提供了可靠依据，在机器翻译、语

音合成、语音识别等方法具有广阔的应用前景。本文提出一种基于三元统计模型的汉语分词标注一体化的方法，既考虑了词性及词汇的三元概率模型，同时又兼顾了词及词性之间的搭配关系。此方法有效地提高了分词和标注地正确率。另外词典的建立也影响分词的结果。对于词典中未登录的词，出现时将无法正确划分，所以，词典应尽量完善一些。但如果词典规模增大将要求更大的训练语料，还会增加歧义点及降低运行效率。标注的错误主要出现在插入语和句式尤为复杂的长句子中，这需要加入规则来改进标注算法。下一步需要研究未登录词的识别[12]及及基于规则的标注[13]研究。

参考文献

[1]刘开瑛，现代汉语自动分词系统中几个问题的讨论，计算机开发与应用，1998

[2]. 王素格，张永奎，刘开瑛，《汉语词性标注中兼类词排歧算法探讨》，计算语言学联合学术会议（JSCL-99），1999，北京

[3].冯志伟，中文信息处理与汉语研究，商务印书馆，1992

[4]刘源等，信息处理用现代汉语分词规范即自动分词方法，清华大学出版社，广西科学技术出版社，1994

[5]刘开瑛，郑家恒，赵军，语料库词类自动标注算法研究，机器翻译研究进展，电子工业出版社，1992

[6] Leonard E. Baum et al, A Maximization Technique Occurring in Statistical Analysis of Probabilistic Functions of Markvov Chain The Annals Mathematical Statistics, 1970, Vol.41. No.1, pp.164~171

[7] 北研二等合著，《音声言语处理》，森北出版株式会社

[8] F.Jelinek, "Self-organized language modeling for speech recognition," in Readings in Speech Recognization, A Waibel and K.F. Lee, eds., Morgan-Kaufmann, San Mateo, CA, 1990, pp450-506

[9] 孙茂松等《中文姓名的自动辨识》，中文信息学报， Vol. 9, No. 2, 1995

[10] 孙茂松等《中文地名的自动辨识》，计算语言学进展与应用，清华大学出版社，1995

[11]《信息处理用现代汉语词类标记集规范》，国家语言文字应用研究所计算语言学研究室

[12] Jian-Yun Nie 等，Unknowing Word Detection and Segmentation of Chinese Using Statistical and Heuristic Knowledge, Communications of CLSIPS, VOL.5, NO.1&2, DEC.1995, page 47-57

[13]. 张民，李生，赵铁军，张艳风，统计与规则并举的汉语词性自动标注算法，软件学报，1998，9(2):134-138

[14]. Brill E, Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging, *Computational Linguistics*, 1995, 21(4): 543-565

[15]. 周强，规则与统计相结合的汉语词类标注方法，中文信息学报，1995，9(2): 1-10

[16]. 张民，李生，赵铁军，基于评价的汉语词性纯概率标注算法，计算机研究与发展，1998，35(4)：349-352

[17]. 周强，俞士汶，一种切分与词性标注相融合的汉语语料库多级处理方法，计算语言学研究与应用，北京语言学院出版社，1993，126-131

[18] 白拴虎硕士论文，《自然语言处理研究》黄昌宁

[19] 孙茂松等 1997 Proc. of ANLP'97