# 汉语词性标注任务

## 任务目标

实现一个汉语的词类自动标注系统。

## 国内外现状

中文词性标注是自然语言处理领域中诸多应用的基础，比如信息检索、机器翻译等后续任务，都可以从一个好的中文分词词性标注模型中获益。国内外研究人员都比较重视词性标注的意义，成功提出或者设计出了许多词性标注模型。总的来说，词性标注算法有两大类，一类是基于规则的方法，一类是基于统计学的方法。但就目前的研究形势来看，基于统计学的方法比较热门，而我们自己的词性标注实现也是基于统计学的方法，接下来就简要介绍下基于统计学方法的词性标注目前国内外的研究现状。

2002年左右，薛念文等人[1]提出了基于字标注的中文分词，将中文分词建模为一种字序列标注的问题。序列标注问题已经在英语等其它语言的词性标注问题中取得了成功。这一基于字的模型抓住了汉语主要以字为语素构词的特点，在汉语语言学研究中也有字本位的研究。自此，引入机器学习方法的字标注模型不断成熟，被应用在分词和词性标注任务中。

黄伟道等人[2]使用最大熵的机器学习方法进行基于字标注的中文分词。同 时，他们扩展了字标注的标注集，使得进行一次字标注就可以同时得到词性标注 结果。他们的实验表明，同时考虑分词词性标注的字标注模型的分词结果，要好 于只使用分词字标注的模型得到的分词结果，但其速度远远慢于只使用分词字标 注的模型。Peng等人[3]使用条件随机场（CRF，Conditional Random Field）进行基于字标注的中文分词，将标注间的上下文约束很好的融入了模型，进一步高了效果，这种方法在当时的评测中也取得了很好的效果。Li 等人[4]使用感知器（perceptron）进行基于字标注的中文分词，较之条件随机场模型，感知器模型中不使用概率概念，训练较为简单、快速。类似的，姜文斌等人[5] 沿用黄伟道等人的标注方法，使用感知器进行基于字序列的词性标注。Qian等人[6]使用包含虚拟节点的改进条件随机场模型，试图降低联合模型的时间复杂度，并能更灵活的使用特征。

字标注模型的特征基于的是字符和字符的二元组等，它的一个不足是不能直接利用更复杂的词一级的特征。因此有别于字标注方法的另一种方法就是基于词

的方法（ word-based），这一类模型能够同时利用字和词的特征，通常模型更为复杂，速度更慢。微软的 Gao 等人[7] 使用一种基于词的信源信道模型进行中文分词和词性标注。 Zhang 等人[8] 提出了基于词的中文分词模型，训练算法使用的是感知器。因为使用了词一级的特征，这些模型的相对于对应的字标注方法，性能有一定的提高。然而同样也是因为使用了词一级的特征，导致模型更为复杂，解码速度更慢，即使进行了优化之后，解码的速度也不理想。

以上谈到的都是使用一个单一的模型进行分词或者词性标注，近年来，多模型组合的方式也被广泛应用。最基本的方式就是先进行分词、再在分词结果的基础上进行词性标注。张华平等人[9]提出了一个基于层叠马尔可夫模型的中文分词词性标注系统，并衍生出了一个实用的中文词法分析系统 ICTCLAS。在这个模型中分若干层次，每个层次都是一个马尔可夫模型，负责进行某一特定的标注、识别任务。相邻两个模型之间数据的压缩表示使用的是一种广义的词图（word lattice）。在这样的词图中，某些命名体被拆分，其中的每个字构成词图中的一条边。

Sun[10]使用 stacked learning 的机器学习框架进行分词词性标注。 Stacked learning 是以若干弱分类器的输出为特征，训练一个更强的分类器的框架。具体的，她使用了三个简单的分词或词性标注模型，分别是基于词的分词模型、基于字的分词模型和局部的基于字的词性标注模型，三个模型并行连接分别处理原始句子。然后，将他们的输出整合成带词性标注的亚词的序列，再送入一个相对复杂的模型中得到最后的结果。

## 本小组的相关工作

### 语料库：

采用北大新闻分词词性标注语料（人民日报1998年1月）。

### 系统流程：

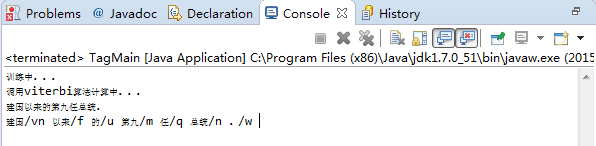
采用一阶马尔可夫模型计算参数，采用隐马尔科夫模型及维特比算法解码。

#### 训练模型参数

1. 数据预处理：
   1. 语料中，每一行前面有诸如“19980101-01-001-016/m”的日期及其词性。但在训练中这个信息不需要，因此将其去除；
   2. 语料中存在诸如“[人民/n 大会堂/n]ns”的词，需要将“[”和“]ns”去除。否则会产生“[人民”这样的词，和“n]ns”这样的词性。
   3. 语料中，大部分情况下词-词性对之间是用单个空格分隔的，但是也存在用多个空格分隔的情况。需要统一分隔符。
   4. 语料中存在空行，需要去除。
2. 计算状态转移矩阵的值：
   1. 采用一阶马尔科夫模型，即假设当前词的词性只与前一个词的词性有关，而与其它词的词性无关。
   2. 统计出每种词性总的出现次数。比如v出现的总次数为c(v)
   3. 统计出任意两个词性前后依次出现的次数。比如v、n依次出现的总次数是c(n|v)
   4. 用极大似然法计算状态转移矩阵中的各项的值。比如q(n|v) = c(n|v)/c(v)。
3. 计算混淆矩阵的值：
   1. 对每个词，分别统计它标注为各种词性的次数。比如c(人/n)=100，c(人/m)=1等等。
   2. 统计出每种词性总的出现次数。比如n出现的总次数为c(n)。
   3. 对所有词性，用极大似然估计法计算该词性出现时，出现某个词的概率。比如q(人|n)= c(人/n)/ c(n)。得到混淆矩阵。

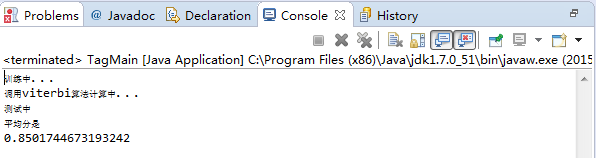
#### 预测词序列中词的词性

1. 输入测试词序列：
   1. 将句子人工分隔为词序列，并保持原有的顺序。
   2. 要求词序列中的词必须在训练语料中出现过。
2. 维特比算法解码并回溯找到最优解
   1. 利用混淆矩阵和状态转移矩阵，用维特比算法解码。维特比算法见附录。
   2. 回溯找到最优解
3. 返回得到的结果：结果举例



### 评估：

由于语料缺乏，并且为了避免测试时出现未登录词，因此我们使用的是训练语料作为测试语料。

1. 评估方法：先计算每一句话的正确率（标注正确的词数除以总词数），再对所有句子的正确率求平均，得到平均正确率。
2. 平均正确率是85.02%  
   

### 缺陷：

1. 没有处理未登录词。如果输入的词在训练语料中没有出现过，就无法给出整个句子的词性
2. 断句不合理。程序中只是按照语料中的一行作为一个句子。实际上一行应该是一篇文档，包含多个句子。
3. 未调用分词。程序所需的输入是字符串数组，该数组按顺序包含了一句话中的每个词。

## 人员分工

## 参考文献

[1] Xue N, Converse S. Combining classifiers for Chinese word segmentation. Proceedings of the 1st SIGHAN Workshop on Chinese Language Processing, 2002. 63 – 70

[2] Ng H T, Low J K. Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? WordBased or Character-Based? In: Lin D, Wu D, (eds.). Proceedings of EMNLP  
2004, Barcelona, Spain: Association for Computational Linguistics, 2004. 277 – 284.

[3] Peng F, Feng F, McCallum A. Chinese Segmentation and New Word Detection using Conditional Random Fields. Proceedings of Proceedings of Coling 2004, Geneva, Switzerland:COLING, 2004. 562 – 568.

[4] Li Y, Miao C, Bontcheva K, et al. Perceptron Learning for Chinese Word Segmentation.Proceedings of Proceedings of Fourth SIGHAN Workshop on Chinese Language processing (Sighan-05), 2005. 154 – 157.

[5] Jiang W, Mi H, Liu Q. Word Lattice Reranking for Chinese Word Segmentation and Part-ofSpeech Tagging. Proceedings of Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, UK: Coling 2008 Organizing Committee,2008. 385 – 392.

[6] Qian X, Zhang Q, Zhou Y, et al. Joint training and decoding using virtual nodes for cascaded segmentation and tagging tasks. Proceedings of Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010. 187 – 195.

[7] Gao J, Li M, Huang C N. Improved source-channel models for Chinese word segmentation. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, 2003. 272–279.

[8] Zhang Y, Clark S. Chinese Segmentation with a Word-Based Perceptron Algorithm. Proceedings of ACL-07:, Prague, Czech Republic: Association for Computational Linguistics, 2007. 840–847.

[9] Zhang H P, Liu Q, Cheng X Q, et al. Chinese lexical analysis using hierarchical hidden markov model. Proceedings of Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17. Association for Computational Linguistics, 2003. 70.

[10] Sun W. A Stacked Sub-Word Model for Joint Chinese Word Segmentation and Part-of-Speech  
Tagging. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA: Association for Computational Linguistics, 2011. 1385 – 1394.

[11] <http://www.52nlp.cn/hmm-learn-best-practices-one-introduction>

[12] <http://www.cs.columbia.edu/~mcollins/hmms-spring2013.pdf>

## 附录：维特比算法

Viterbi算法是一种动态规划算法，寻找最可能的隐含状态序列。

Viterbi算法是在给定模型和观察序列的条件下求最大的状态序列：



定义Viterbi变量是在时间t时，模型沿着某一条路径到达，并输出观察序列的最大概率：



Viterbi算法如下所示：

1. 初始化：

最大概率路径变量

1. 递推计算：



1. 结束：



1. 通过回溯得到状态序列

