对中文序列标注问题的调研

赖泽强，刘文卓，钱泽，谭超

北京理工大学 计算机学院

摘　　要:针对中文分词，我们实现了基于词典的方法（最大匹配法），和基于统计的方法（Uni-Gram模型以及隐式马尔可夫模型）。对于词性标注，我们同样采用隐式马尔可夫模型进行词性序列的标注。与此同时，我们测试神经网络算法在分词，词性标注，命名实体识别问题的效果。

关　键　词：中文分词，词性标注，命名实体识别，隐式马尔可夫模型，神经网络

1　引言

作为中文自然语言处理中的基础工作，中文分词，词性标注以及命名实体识别虽然较其他任务，如句法分析，观点抽取，更为简单，但其重要性仍然不可忽视。

本篇文章将就中文分词，词性标注及命名实体识别的基础算法进行探讨，并通过我们的实验验证不同算法的效果差异。

针对中文分词，我们实现了基于词典的方法（最大匹配法），和基于统计的方法（Uni-Gram 模型以及隐式马尔可夫模型）。对于词性标注，我们同样采用隐式马尔可夫模型进行词性序列的标注。与此同时，我们测试神经网络算法在分词，词性标注，命名实体识别问题的效果。

本文后续部分组织如下。第二节给出了中文分词的问题描述，对三种算法的简单重述以及具体的实验结果；第三节则给出词性标注的问题描述，其对应算法的描述以及实验结果；第四节给出神经网络序列标注的算法与结果；第五节就本文内容进行了总结。

2　中文分词

2.1 问题重述

中文分词指的是将一个汉字序列切分成一个个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。[1]

2.2 基于词典：最大匹配法

正向最大匹配：正向将待分词文本中的几个连续字符与词典匹配，如果匹配成功，则切分出一个词。

逆向最大匹配：逆向将待分词文本中的几个连续字符与词典匹配，如果匹配成功，则切分出一个词。

双向最大匹配：根据正向和逆向的切分结果以及中文分词一般原则，选出切分效果更好的结果。

中文分词的一般原则：切分出词的数量越少越好，切分单字的数量越少越好。

2.3 基于统计：Uni-Gram 模型

N-gram模型的介绍:N-gram是基于一个假设：第n个词出现与前n-1个词相关，而与其他任何词不相关。整个句子出现的概率就等于各个词出现概率的乘积。各个词的概率通过语料中的统计计算得到。假设句子T是有词序列w1,w2,w3...wn组成，用公式表示N-Gram语言模型如下：

$$P(T)=P(W1W2…Wn)=P(W1)P(W2|W1)P(W3|W1W2)…P(Wn|W1W2…Wn-1)$$

当n = 1时，即为Uni-Gram模型，公式如下：

$$P(T) =P(W1)P(W2)P(W3)…P(Wn)$$

该模型只考虑了每个单独的标注出最可能出现的情况，并未考虑一个标注的上下文对其的影响。

2.4 基于统计：隐式马尔可夫模型

马尔可夫模型中有两个序列，一条是显式的字串序列 W，也就是我们输入的句子，另一个是隐式的标注序列 T。这两个序列的长度是相等的，并且一一对应。在中文分词的任务中，一个字对应一个标注。

[1] 摘自百度百科[中文分词]

中文分词的隐式马尔科夫模型是基于字标注的模型。一共有四种标注：B（词的开头），M（词的中部），E（词的结尾），S（单字）。

马尔可夫模型中，一个标注序列在给定字串序列情况下出现的概率P(T|W)公式如下：

$$P(T|W) = P(t1|t0)\times P(W1|t1) \times P(t2|t1) \times P(W2|t2) \times \ldots \times P(tn|tn-1)\times P(Wn|tn)$$

式中：P(ti|ti-1)表示第i-1个标注状态转移到第i个标注状态的概率， P(Wi|ti)表示第i个状态标注下，对应字串的字出现的概率。

对于给定的字串序列，我们需要求出概率最大，也就是最可能出现的标注序列。

维特比算法的实质是动态规划的求解方法。从字串的开头进行计算，对于每一个字的每一个可能的标注状态，记录下从前一个字的任一标注达到该标注状态的最大值。这样一直记录到最后一个字,所得到的即为整个句子的全局最大概率。

2.5 实验

数据集：以下实验使用的数据集为Bakeoff 2005的PKU数据集。

结果：最大匹配法，Unigram以及隐式马尔可夫模型如下列各表所示：

表 1 最大匹配法结果展示

模型	Precision	Recall	F1	速度(字/s)
前向	0.843	0.907	0.874	32000+
后向	0.845	0.909	0.876	10000+
双向	0.845	0.909	0.876	7600+

表 2 n-gram模型结果展示

模型	Precisio	Recall	F1	速度(字/s)
Unigram	0.844	0.922	0.881	8400+
HMM	0.777	0.792	0.785	20000+

2.6 分析

1）特殊字符处理:

分析五个模型的分词结果，我们发现不管是基于词典的最大匹配法，基于统计的隐式马尔可夫模型，还是二者混合的 Unigram 模型，均无法有效的将时间，数字，人名和地名准确的切分处理。

其原因在于最大匹配是基于字典的切分方式，当遇到字典中未出现的词语时，最大匹配法无法正确的切分。

因此我们增加了基于规则的数字，日期匹配算法，通过实验，我们取得了 3%的 F1 值提升。具体结果如下表所示。

2）隐式马尔可夫模型

从结果来看，隐式马尔可夫模型的表现较差。

表 3 加入日期数字匹配结果展示

模型	Precisio	Recall	F1	速度(字/s)
前向	0.892	0.932	0.912	10000+
后向	0.894	0.934	0.914	9000+
双向	0.895	0.935	0.914	5000+
Unigram	0.893	0.946	0.919	3600+

不管是精确率，召回率还是 F1 值都显著低于最大匹配法和 Unigram 模型。

经过分析，我们认为可能有以下几个原因：

1. 隐式马尔可夫模型是基于字标注的模型，它没能很好的利用词典的信息进行标注。
2. 隐式马尔可夫模型对于其发射矩阵，转移矩阵，初始矩阵的空值较为敏感，简单的平滑处理不能很好的体现真实情况。
3. 训练语料与测试中的句子长度也可能是其表现较差的原因。

因此我们认为：尝试整合词典，采用更为复杂的平滑方法，长句化为几个短句进行处理，均可能提升隐式马尔可夫模型的分词效果。不过由于时间关系，我们还未进行实际的实验。

3 词性标注

3.1 问题重述

词性标注，又称为词类标注或者简称标注，是指为分词结果中的每个单词标注一个正确的词性的程序，也即确定每个词是名词、动词、形容词或者其他词性的过程。

3.2 隐式马尔可夫模型

模型与中文分词的隐式马尔可夫模型基本一致。输入序列有所不同，词性标注的输入序列是切分好的词串序列，对应的隐式序列为词性标注的序列。

3.3 实验

数据集：我们使用人名日报标注语料库 1998 年进行实验。前 80%作为训练语料，后 20%作为测

试语料。

结果：隐式马尔可夫模型的词性标注结果如下表所示

表 4 HMM词性标注结果展示

模型	Precision	Recall	F1	速度(字/s)
HMM	0.84	0.84	0.84	8400+

3.4 分析

1）稀疏矩阵的平滑：

构建初始、转移、发射矩阵时，语料库中未出现的情况会导致矩阵的许多元素为零，进一步会导致用维特比算法求解时某些路径的概率直接为 0，这显然不是我们想要的结果，因为有些情况语料库中未出现，并不代表这种情况不存在。因此，我们需要进行稀疏矩阵的平滑，对零元素重新估计赋值。

我们采用了两种平滑的方法：简单的加一平滑和 good-turing 平滑。

Good-turing 平滑的基本思想是用出现次数为 r+1 的元素去重新估计出现次数为 r 的元素，重新估值的公式为：

Rnew = (r+1)Nr+1/Nr

在具体的实现过程中，我们只重新估计了零元素，并设定阈值，使得零元素的重新估值不会超过 1。

测试时，采用 good-turing 平滑相较于简单的加一平滑,准确率提升了 1%至 2%。

2）分词误差的传播

由于我们要求输入为分好词的句子，因此分词的准确率将会影响到词性标注的准确率。这个误差的传播可以采用联合模型，即同时进行分词与词性标注来解决。

4 基于神经网络的序列标注

事实上，不管是中文分词，词性标注还是命名实体识别，它们本质上均可以转换为序列标注的问题。对于中文分词,标注为 BMES,词的开头，中间，结尾以及单字成词；对于词性标注，则是某个标注的开头，中间与结尾；而对于命名实体识别，则是某种命名实体的开头，中间与结尾。

4.1 BiLSTM-CRF

众所周知,双向 LSTM 是一个针对包含序列的问题的常用神经网络。针对一个输入序列，双向 LSTM 会输出一个等长的输出序列。

得益于深度神经网络强大的表示能力，双向 LSTM 在序列标注上也能取得较好的结果。但是由于双向 LSTM 只考虑了输入序列的前后约束关系，而没有考虑输出序列的约束关系。其序列标注效果仍然收到制约。如，在分词上，输出序列 BBE 是不应该出现的，因为 B 后面不可能紧跟着一个 B，但是 LSTM 无法很好的学习到这种特征。

CRF 的加入，弥补这方面的不足。通过将 CRF 接在双向 LSTM 的输出之上,模型则能够学会将上面不合法的情形排除掉。

4.2 实验

1）数据集：

- 中文分词：Bakeoff 2005 PKU
- 词性标注：人名日报标注语料 1998
- 命名实体识别：人名日报标注语料 1998，共三种实体类型，地名，人名，机构名。

2）结果

表 5 BiLSTM-CRF分词效果展示

模型	Precision	Recall	F1	速度(字/s)
BiLSTM-CRF	0.885	0.870	0.877	-

表 5 BiLSTM-CRF命名实体识别效果展示

模型	Precision
BiLSTM-CRF	0.93

命名实体识别展示的准确率为标记的准确率，即 B，I，O 的准确率。

2）分析

- 实际上，从标记的准确率来看，BiLSTM-CRF 可以达到 95%的准确率，但是从分词的准确率来看，其效果并不比最大匹配法要好很多。为了让神经网络发挥其威力，一个庞大的数据集和精细的调教仍需十分重要。

- 命名实体识别也存在同样的问题，虽然从标记的角度看准确率很高，但从命名实体整体识别率来看，效果较差。

5 结 论

1）在分词上，简单的最大匹配法无法较好的处理未登录的词的问题。针对特殊字符，制定相

应的匹配规则能够解决部分未登录词的问题。而整合基于统计的 n-gram 模型，则能够解决部分歧义的问题。

2）单纯的隐式马尔可夫模型无法在分词上取得非常好的效果，其训练需要较为细致的编码调节以及大量的数据。

3）神经网络的方法属于端到端的模型，拥有较强的标注能力，但对计算能力有较高的要求。

参考文献（References）

[1] 基于keras的BiLstm与CRF实现命名实体标注 https://www.cnblogs.com/vipyoumay/p/ner-chinese-keras.html

[2] 基于统计学的的分词 https://applenob.github.io/statistics_seg.html

[3] python中文分词 http://www.isnowfy.com/python-chinese-segmentation/

[4] 用双向最大匹配法进行中文分词 https://blog.csdn.net/PKU_ZZY/article/details/54730972

[5] 用HMM模型进行中文分词 https://blog.csdn.net/PKU_ZZY/article/details/56479627

[6] 用HMM模型进行中文分词 https://blog.csdn.net/PKU_ZZY/article/details/56479627

[7] 用HMM模型进行中文分词 https://blog.csdn.net/PKU_ZZY/article/details/56479627

[8] 词性标注调研 http://heshenghuan.github.io/2016/03/23/词性标注调研

[9] 基于一阶HMM的中文词性标注（Java实现） https://blog.csdn.net/rm_wang/article/details/50838243