NLP实验二 汉语命名实体自动识别系统

SA18225293 彭辰铭

一、实验目的

1.熟悉国内外汉语命名实体自动识别技术的进展

2.独立完成汉语命名实体自动识别系统

二、命名实体识别方法综述

命名实体是命名实体识别的研究主体，一般包括三大类（实体类、时间类和数字类）和七小类（人名、地名、机构名、时间、日期、货币和百分比）命名实体。

评判一个命名实体是否被正确识别包括两个方面：实体的边界是否正确和实体的类型是否标注正确。 命名实体识别的主要技术方法分为：

基于规则和词典的方法、基于统计的方法、二者混合的方法等。

1. 基于规则和词典的方法

基于规则的方法多采用语言学专家手工构造规则模板,选用特征包括统计信息、标点符号、关键字、指示词和方向词、位置词(如尾字)、中心词等方法，以模式和字符串相匹配为主要手段，这类系统大多依赖于知识库和词典的建立。

基于规则和词典的方法是命名实体识别中最早使用的方法，一般而言，当提取的规则能比较精确地反映语言现象时，基于规则的方法性能要优于基于统计的方法。

但是这些规则往往依赖于具体语言、领域和文本风格，编制过程耗时且难以涵盖所有的语言现象，特别容易产生错误，系统可移植性不好，对于不同的系统需要语言学专家重新书写规则。基于规则的方法的另外一个缺点是代价太大，存在系统建设周期长、移植性差而且需要建立不同领域知识库作为辅助以提高系统识别能力等问题。

2. 基于统计的方法

基于统计机器学习的方法主要包括：隐马尔可夫模型(HiddenMarkovMode,HMM)、最大熵(MaxmiumEntropy,ME)、支持向量机(SupportVectorMachine,SVM)、条件随机场(ConditionalRandom Fields,CRF)等。

最大熵模型结构紧凑，具有较好的通用性，主要缺点是训练时间复杂性非常高，有时甚至导致训练代价难以承受，另外由于需要明确的归一化计算，导致开销比较大。

条件随机场为命名实体识别提供了一个特征灵活、全局最优的标注框架，但同时存在收敛速度慢、训练时间长的问题。

一般说来，最大熵和支持向量机在正确率上要比**隐马尔可夫模型**高一些，但是隐马尔可夫模型在训练和识别时的速度要快一些，主要是由于在利用Viterbi算法求解命名实体类别序列的效率较高。隐马尔可夫模型更适用于一些对实时性有要求以及像信息检索这样需要处理大量文本的应用,如短文本命名实体识别。

基于统计的方法对特征选取的要求较高，需要从文本中选择对该项任务有影响的各种特征，并将这些特征加入到特征向量中。依据特定命名实体识别所面临的主要困难和所表现出的特性，考虑选择能有效反映该类实体特性的特征集合。主要做法是通过对训练语料所包含的语言信息进行统计和分析，从训练语料中挖掘出特征。有关特征可以分为具体的单词特征、上下文特征、词典及词性特征、停用词特征、核心词特征以及语义特征等。

基于统计的方法对语料库的依赖也比较大，大数据时代以前可以用来建设和评估命名实体识别系统的大规模通用语料库比较少。

3. 混合方法

(1)统计学习方法之间或内部层叠融合。

(2)规则、词典和机器学习方法之间的融合，其核心是融合方法技术。在基于统计的学习方法中引入部分规则，将机器学习和人工知识结合起来。

(3)将各类模型、算法结合起来，将前一级模型的结果作为下一级的训练数据，并用这些训练数据对模型进行训练，得到下一级模型。

三、实验原理

词性标注：

给一个句子中的每一个单词注明磁性。

特征函数：

那么当我们标注完一个句子后，需要对其进行打分。比如 因为 动词后是动词 相对出现概率率较小，我们就给它减分。这里的我们的 **动词后是动词** 就是一个**特征函数。**

我们可以定义一个特征函数集合，用这个特征函数集合来为一个标注序列打分，并据此选出最靠谱的标注序列。

也就是说，每一个特征函数都可以用来为一个标注序列评分，把集合中所有特征函数对同一个标注序列的评分综合起来，就是这个标注序列最终的评分值。

定义crf中的特征函数

现在，我们正式地定义一下什么是CRF中的特征函数，所谓特征函数，就是这样的函数，它接受四个参数：

- 句子s（就是我们要标注词性的句子） - i，用来表示句子s中第i个单词 - l_i，表示要评分的标注序列给第i个单词标注的词性 - l_i-1，表示要评分的标注序列给第i-1个单词标注的词性

它的输出值是0或者1,0表示要评分的标注序列不符合这个特征，1表示要评分的标注序列符合这个特征。

（这里特征函数仅仅依靠**当前单词的标签**和**它前面的单词的标签**对标注序列进行评判，这样建立的CRF也叫作**线性链CRF**，这是CRF中的一种简单情况）

定义好一组特征函数后，我们要给每个特征函数f_j赋予一个权重λ_j。现在，只要有一个句子s，有一个标注序列l，我们就可以利用前面定义的特征函数集来对l评分。

$$score(l|s) = \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i, l_{i-1})$$

对这个分数进行**指数化和标准化**，我们就可以得到标注序列l的概率值**p(l|s)**，如下所示：

$$p(l|s) = \frac{exp[score(l|s)]}{\sum_{l'} exp[score(l'|s)]} = \frac{exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i,l_{i-1})]}{\sum_{l'} exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l'_i,l'_{i-1})]}$$

$$p(l|s) = \frac{exp[score(l|s)]}{\sum_{l'} exp[score(l'|s)]} = \frac{exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i,l_{i-1})]}{\sum_{l'} exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l'_i,l'_{i-1})]}$$

几个特征函数的例子

前面我们已经举过特征函数的例子，下面我们再看几个具体的例子，帮助增强大家的感性认识。

$$f_1(s,i,l_i,l_{i-1}) = 1 \ f_1(s,i,l_i,l_{i-1}) = 1$$

当l_i是“副词”并且第i个单词以“ly”结尾时，我们就让f1 = 1，其他情况f1为0。不难想到，f1特征函数的权重λ1应当是正的。而且λ1越大，表示我们越倾向于采用那些把以“ly”结尾的单词标注为“副词”的标注序列

$$f_2(s,i,l_i,l_{i-1}) = 1 \, f_2(s,i,l_i,l_{i-1}) = 1$$

如果i=1，l_i=动词，并且句子s是以“？”结尾时，f2=1，其他情况f2=0。同样，λ2应当是正的，并且λ2越大，表示我们越倾向于采用那些把问句的第一个单词标注为“动词”的标注序列。

$$f_3(s,i,l_i,l_{i-1}) = 1 \, f_3(s,i,l_i,l_{i-1}) = 1$$

当l_i-1是介词，l_i是名词时，f3 = 1，其他情况f3=0。λ3也应当是正的，并且λ3越大，说明我们越认为介词后面应当跟一个名词。

$$f_4(s,i,l_i,l_{i-1}) = 1 \, f_4(s,i,l_i,l_{i-1}) = 1$$

如果l_i和l_i-1都是介词，那么f4等于1，其他情况f4=0。这里，我们应当可以想到λ4是负的，并且λ4的绝对值越大，表示我们越不认可介词后面还是介词的标注序列。

总结

为了建一个条件随机场，我们首先要定义一个特征函数集，每个特征函数都以整个句子s，当前位置i，位置i和i-1的标签为输入。然后为每一个特征函数赋予一个权重，然后针对每一个标注序列l，对所有的特征函数加权求和，必要的话，可以把求和的值转化为一个概率值。

这里的特征函数也与我们的特征模板相对应起来。

四、实验结果

```
输入训练数据的大小
1000
MIRA doesn't support multi-thrading. use thread_num=1
CRF++: Yet Another CRF Tool Kit
Copyright (C) 2005-2013 Taku Kudo, All rights reserved.

reading training data: 100.. 200.. 300.. 400.. 500.. 600.. 700.. 800.. 900.. 1000..
1100.. 1200.. 1300.. 1400.. 1500.. 1600.. 1700.. 1800.. 1900.. 2000.. 2100.. 2200..
2300.. 2400.. 2500.. 2600.. 2700.. 2800.. 2900.. 3000.. 3100.. 3200.. 3300.. 3400..
3500.. 3600.. 3700.. 3800.. 3900..
Done!0.82 s

Number of sentences: 3945
Number of features:  2658721
Number of thread(s): 8
Freq:                1
eta:                 0.00010
C:                   1.00000
shrinking size:      20
iter=0 terr=0.12255 serr=0.31128 act=3945 uact=0 obj=115.40146 kkt=75.97336
iter=1 terr=0.03511 serr=0.11787 act=3945 uact=0 obj=157.12915 kkt=99.13901
iter=2 terr=0.01766 serr=0.06489 act=3945 uact=0 obj=181.46084 kkt=43.78797
iter=3 terr=0.00992 serr=0.04005 act=3945 uact=0 obj=197.43734 kkt=13.01529
iter=4 terr=0.00703 serr=0.02839 act=3945 uact=0 obj=208.87132 kkt=10.56902
iter=5 terr=0.00305 serr=0.01546 act=3945 uact=0 obj=214.16387 kkt=7.44148
iter=6 terr=0.00263 serr=0.01318 act=3945 uact=0 obj=218.99803 kkt=6.50943
iter=7 terr=0.00352 serr=0.01343 act=3945 uact=0 obj=225.07030 kkt=13.63841
iter=8 terr=0.00161 serr=0.00532 act=3945 uact=0 obj=227.62229 kkt=8.66654
iter=9 terr=0.00119 serr=0.00558 act=3945 uact=0 obj=229.83357 kkt=5.90349
iter=10 terr=0.00158 serr=0.00811 act=3945 uact=0 obj=232.86389 kkt=7.68019
iter=11 terr=0.00142 serr=0.00532 act=3945 uact=0 obj=235.27308 kkt=18.86842
iter=12 terr=0.00098 serr=0.00304 act=3945 uact=0 obj=237.02414 kkt=8.71288
iter=13 terr=0.00107 serr=0.00456 act=3945 uact=0 obj=238.94518 kkt=11.66935
iter=14 terr=0.00098 serr=0.00406 act=3945 uact=0 obj=240.68047 kkt=6.90341
iter=15 terr=0.00074 serr=0.00330 act=3945 uact=0 obj=242.08902 kkt=9.99894
iter=16 terr=0.00025 serr=0.00152 act=3945 uact=0 obj=242.61780 kkt=2.98955
iter=17 terr=0.00000 serr=0.00000 act=3945 uact=0 obj=242.61780 kkt=0.00000
iter=18 terr=0.00000 serr=0.00000 act=3945 uact=0 obj=242.61780 kkt=0.00000

Done!244.51 s

准确率:  0.863649522339091
```

```
召回率：  0.6804013988140489
f值：  0.761151507420164
```

实验代码

五、参考

如何用简单易懂的例子解释条件随机场（CRF）模型？

实现汉语自动命名实体自动识别系统