2015 年 12 月 第 12 期
Vol. 36 No. 12 2015

基于半监督隐马尔科夫模型的汉语词性标注研究

韩 霞,黄德根

(大连理工大学 计算机学院,辽宁 大连 116024)
E-mail:huangdg@dlut.edu.cn

摘 要: 提出一种基于词语相似度计算的半监督隐马尔科夫词性标注方法. 首先,利用小规模的训练语料进行半监督隐马尔科夫学习,通过反复迭代不断扩充语料,增强隐马尔科夫的标注效果;然后,通过计算词语相似度的方法,给测试语料中每个未登录词都标上候选词性;最后,在隐马尔科夫标注时,不是选取一条最佳路径,而是选取两条最佳路径,通过二次选择,以此得到标注结果. 实验结果证明,该方法与传统的隐马尔科夫标注方法相比提高了约 2.60%,汉语词性标注准确率达到了 95.65%.

关 键 词: 词性标注;词向量;词语相似度;迭代训练
中图分类号: TP391 **文献标识码**: A **文 章 编 号**: 1000-1220(2015)12-2813-04

Research on Chinese Part-of-speech Tagging Based on Semi Hidden Markov Model

HAN Xia, HUANG De-gen
(School of Computer Science&Technology, Dalian University of Technology, Dalian 116024, China)

Abstract: We present in this paper methods to improve semi HMM (Hidden Markov Model) based POS (part-of-speech) tagging of Chinese utilizing word similarity. First, we iteratively expand corpus beginning with the small size of training corpus to make semi supervised HMM learning. Thus, the effect of HMM tagging enhanced. Then, through the method of word similarity computation, each unknown word in the test corpus has candidate POS. At the same time, in order to get the annotated results, we select top two paths to conduct the second choice rather than just the optimal one. Experiments show that, this method has a 2.60% increase compared to simple HMM. Our model achieves an accuracy of 95.65%.

Key words: part-of-speech tagging; word vector; word similarity; iterative training

1 引 言

汉语词性标注是自然语言处理领域的一项基础工作. 汉语中存在很多兼类词,这类词具有多个词性. 词性的选择取决于不同的上下文环境. 词性标注是指给句子中的每个词都标上它在这个句子中的词性,即它是动词、名词、连词还是副词等其他词性[1]. 正确的词性标注能为后续的工作,如名词短语识别、句法分析、机器翻译等提供良好的前提[2]. 目前,英文词性标注已取得 97% 左右的正确率[3],而汉语由于缺乏能够提供明显句法特征的形态学线索,标注的任务比较困难[4],标注正确率只有 93%-95%.

词性标注方法主要有两大类:基于规则的方法和基于统计的方法. 基于统计的方法相对比较主流,主要有最大熵模型(ME)、隐马尔科夫模型(HMM)、条件随机域(CRFs)等. 其中,根据训练语料库的选择,又可分为有监督的,半监督的和无监督的训练方式. 有监督的训练需要大量的标记语料,耗费的人力和时间比较大,但是简单、易于操作,能够保证较高的精确率;无监督的训练不需要训练语料,灵活性大,精确率较低;半监督的训练,利用部分训练语料,得到的精确率低于有监督的方式但高于无监督方式.

文献 [5] 分别选取近距离和远距离的特征,即上下文词及词性特征和触发词特征,进行 CRF 训练及标注;文献 [6] 基于最大熵模型,先利用 Beam-Search 算法求出最优标记序列和次优标记序列,再对兼类词的词性进行选择. 文献 [7] 利用 Rankboost 算法,对 HMM 选出来的若干条最佳路径进行等级排位,挑选出最好的一条路径. CRF 的方法主要存在训练时间长的缺点[8],EM 对语料库依赖性强,同时,这两种方法都需要对特征进行选择,不同的特征对实验效果影响也较大,HMM 则不存在这些问题.

近年,HMM 被广泛的应用于词性标注任务中[9]. 从现有公开发表的文献来看,目前基于 HMM 的汉语词性标注的最好正确率是 95.11%[7]. 传统的 HMM 通常假设当前词只与前面若干词有关,即只利用了单向的词性依赖关系,不能利用丰富的前后向信息[10]. 为了克服这一缺点,本文进行了双向的标注,选出最优的两条路径,之后通过迭代地进行规则处理和改进的词性转移概率的计算,使得词性的选择达到收敛状态,这时也得到了一条最优路径.

2 Word2vec 及词语相似度

2.1 Word2vec 模型简介

Word2vec 是 Google 公司在 2013 年发布的一个开源项目. 其基本原理是是利用一个简单的三层神经网络把词转换

收稿日期:2014-09-09 收修改稿日期:2014-11-21 基金项目:国家自然科学基金项目(61173100,61173101,61272375)资助. 作者简介:**韩 霞**,女,1990 年生,硕士研究生,研究方向为汉语词性标注;**黄德根**,男,1965 年生,教授,博士生导师,研究方向为自然语言处理与机器翻译.

成一个 K 维向量[11]. 该工具可以用来进行词与词之间相似度的计算、词语聚类、短语的自动识别等任务. 例如,搜索词语“中国”的近似词的时候,系统会给出“英国”、“意大利”、“俄罗斯”“加利福尼亚”等词,这些词都是系统根据它们各自的特征向量所计算出来的. 另外,有如下式的一个规律:

$$V(\text{国王}) - V(\text{男}) + V(\text{女人}) \approx V(\text{王后}) \quad (1)$$

其中 V(国王)表示“国王”这个词对应的向量,V(男人)表示“男人”这个词对应的向量,V(女人)表示“女人”这个词对应的向量,V(王后)表示“王后”这个词对应的向量[12]. 公式 1 也证明了词语的向量能正确反映词语间的相似性.

2.2 利用 Word2vec 进行词语相似度计算

从网上下载搜狗实验室的新闻语料,经过去标签等预处理工作,得到 500M 纯文本语料. 用 Nihao 分词工具对文本进行分词,将分好词的汉语语料作为输入,可得到语料中每个词的特征向量,本文取 M 为 200,即一个汉语词语对应一个 200 维的向量.

实验开始时,先根据词典给词典内已有的词语标上词性,这样测试语料中的一部分词语已经标有词性;利用词语向量(以下简称词向量)之间的距离来进行词语相似度计算,词向量之间的距离越小,则词越相似,它们的词性相同的可能性越大,通过计算每个已有词性的词与没有词性的词的相似度,得到每个没有词性的词的最相似的 N 个词,然后将这些词的重复词性作为该词的候选词性.

3 HMM 模型用于词性标注

3.1 HMM 模型的基本原理

HMM 是以朴素贝叶斯(Naïve Bayesian)为基础的[13]. 在隐马尔科夫模型中,不知道模型所经历的状态序列,而只知道输出序列,即状态序列的一个随机函数. 可以用图 1 来解释隐马尔科夫的基本原理.

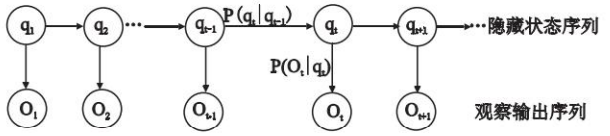

图 1 HMM 图解

Fig. 1 HMM illustration

HMM 一般记为一个五元组 $\mu = (S, K, A, B, \pi)$. 其中,S 为所有的状态的集合,K 为输出符号的集合,π 表示初始状态的概率分布,A 表示状态的转移概率,B 表示符号的发射概率. 在本任务中,S 是指语料中所有的词,K 指词性集合,π 指句首的词性,A 是词性到词性之间的转移概率矩阵,B 为词性到词的发射概率矩阵.

3.2 维特比算法

维特比(Viterbi)算法运用了动态规划的搜索算法求解最佳序列. 维特比变量 $\delta_t(i)$ 是指在时间 t 时,HMM 沿着某条路径达到状态 s_i,并输出观察序列 $O_1O_2\cdots O_t$ 的最大概率:

$$\delta_t(i) = \max_{q_1,q_2,\cdots q_{t-1}} P(q_1, q_2, \cdots, q_t = s_i, O_1O_2\cdots O_t \mid \mu) \quad (2)$$

$\delta_t(i)$ 有如下递归关系:

$$\delta_{t+1}(i) = \max_j [\delta_t(i) \cdot a_{ji}] \cdot b_i(O_{t+1}) \quad (3)$$

其中,a_{ji} 表示从 t 时刻词的第 j 个词性转换到 t + 1 时刻词的第 i 个词性的概率,$b_i(O_{t+1})$ 表示由 t + 1 时刻词的第 i 个词性到当前词的发射概率.

3.3 HMM 模型训练和测试

使用的正确标注语料是一个小规模的语料,大小为 5.25M. 语料的大小直接影响训练结果的好坏,然而大规模的标注语料又要花费大量的人力和时间. 希望能通过某种方式,即节省人力和时间,又能获得较好的实验效果.

从搜狗实验室下载的语料中随机获取一部分(约 50M),分成 10 组,每组 5M. 分词之后用训练语料对第一组语料进行标注,再定义一个评价指标,剔除结果中认为不可靠的部分,把通过筛选的语料加入到训练语料中,用该新生成的训练语料进行第二组语料的标注,将结果中可靠的部分再加入到当前的训练语料中,由此进行后面几组的训练和测试过程,反复迭代,最终获得一个较大规模的训练语料,对本实验的测试语料进行标注.

HMM 标注时,利用维特比算法,进行向前和向后两种方式标注,分别选出最优路径. 以一个完整的句子为单位进行评价,如果该句子的前后两种标注结果相同,则认为是可靠的标注,并将该句子添加到训练语料中,反之,只要有一个词的前后向词性不同,则剔除该句子.

3.4 HMM 二次计算

进行最后测试语料的标注时,依旧选出两条最佳路径,这两条路径分别是通过 HMM 向前 Viterbi 和向后 Viterbi 选出的词性序列,把两条路径中相同的标注当作是可靠标注,不同标注为不可靠部分.

由于 HMM 计算过程只计算了词性到词的发射概率,而忽略了词本身到词性的发射概率,所以会有一些词因为发射概率太小而被标错. 例如,“一”有“COM-NUM”、“ADV”两种词性,由于词性“ADV”到词“一”发射概率较小,为“0.00668”,而“COM-NUM”到“一”发射概率较大,为“0.19585”,在应该被标成“ADV”的情况下,很容易被标成“COM-NUM”. 因此,需利用一些通用规则来处理这类问题,规则修改过的词性也被认为是可靠的.

在一个句子里面,如果某不可靠的词性的前后位置都为可靠词性,则对该词再次进行词性转移概率的计算,作为二次 HMM 标注. 在这次的词性转移概率计算时,不仅计算前词词性到当前词词性的转移概率,同时计算当前词词性到后词词性的转移概率,把它叫做三元词性转移概率. 如式(4),t 时刻为当前词,其前后向标注的词性不同,t - 1 和 t + 1 时刻的词均有可靠词性,分别是其第 i 个词性和第 l 个词性,即 t - 1 时刻的词选择的是其第 i 个词性,t + 1 时刻的词选择的是其第 l 个词性.

$$P = \max_j (a_{ij} \cdot a_{jl}) \quad (4)$$

利用规则和 HMM 标注进行反复迭代,使得标注趋于稳定. 即,句子中的词性的可靠性都已稳定,再进行规则也不会有词性由不可靠变为可靠. 不可靠的词性均选择向后 Viterbi 选出的词性.

HMM 二次计算算法的具体流程如下:

输入:前向词性序列 forwardPOS,后向词性序列 backwardPOS.

Step 1. 初始化:

依次遍历每个词的两种词性,若 forwardPOS [i] 跟 backwardPOS [i] 不同,则将 forwardPOS [i] 变为空,0 < i < n,n 为该句文本中词的个数;

Step 2. HMM 再计算:

遍历每个词的前向词性,若 forwardPOS [i - 1] 和 forwardPOS [i + 1] 均为空,而 forwardPOS [i] 不为空,计算 $P = \max_j (a_{ij} \cdot a_{ji})$,选出当前词的词性赋值给 backwardPOS [i],同时设置 forwardPOS [i] 为空,1 < i < n - 1,n 为该句文本中词的个数;

Step 3. 规则:

如果 backwardPOS [i] 符合规则库中某条词性处理规则,backwardPOS [i] 变为规则中对应的词性,并设置 forwardPOS [i] 为空,0 < i < n,n 为该句文本中词的个数;

Step 4. 迭代:

若 Step 3 过程有被修改的词性,则重新返回到 Step 2 进行迭代,否则,结束.

4　实验结果和分析

4.1　实验语料及工具

测试语料是采用 2000 年的人民日报语料,分词采用 Nihao 分词工具,词性标注集共有 85 个词性,表 1 给出了实验语料的基本信息.

表 1　实验语料信息表

Table 1　Simple statistics of experimental corpus

语料信息	大小
测试语料大小	2.5M
测试语料词语个数(含重复词)	239671
Word2vec 训练语料大小	500M
词典词个数	38407
测试语料中未登录词个数	1531
测试语料中词典词个数	20754

4.2　各阶段效果对比

测试语料中存在较多的未登录词,这类词没有候选词性.为了给未登录词标上可靠的候选词性,先利用规则给一些具有共性的,而且比较容易被识别的词标上词性.表 2 列举了部分规则处理词.这类词在统计时不包含在未登录词范围内.

表 2　规则处理的部分词性表

Table 2　A part of POS dealed by rules

词　性	举　例
时间词 TIME	"1999 年"、"6 点半"、"10 月份"
数字 COM-NUM	"104.8"、"103:90"、"3.3%"、"1069.6 万"
字符串 STRING	"27A1"、"CA930"、"Flex-ATX"

初始训练语料规模较小,需要进行训练语料的扩充.先把从搜狗实验室下载的 50M 语料分 10 组,每组 5M,作为辅助训练语料.利用初始训练语料对第一组辅助训练语料进行标注,将可靠的标注结果放入训练语料,组成新的训练语料,并用该语料对第二组的辅助训练语料进行标注,同样将标注结果中可靠的部分放入训练语料,由此不断地迭代进行 HMM 训练和测试.由图 2 可以看出,随着训练语料的越来越大,词性标注准确率的增幅在不断的降低,汉语标注正确率趋于稳定,若想再提高标注效果,依赖增大训练语料是不太有效果的,必须通过其他方式去提高.

规则处理能有效的标出语料中的大部分的数字,时间词和字符串等词,但是由于语料本身存在的一些问题,例如全角半角、书写错误等,还是会导致一些错误的发生.例如:"——1.77",本身其实是一个负数,应该被标成"COM-NUM",但是由于前面的负号被写成了"——"而非"-",在规则处理时未能正确识别.

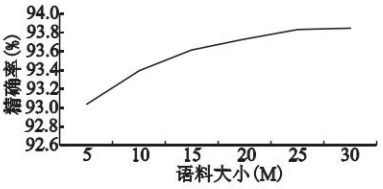

图 2　不同训练语料大小对应的实验结果

Fig. 2　Accuracy on different size of corpus

通过 Google 的 Word2vec 工具,利用相似度计算的方法,把已有词性的词跟没有词性的词做相似度计算,并将相似度最大的若干词的重复词性当作是未知词的候选词性.这种获取候选词性的方法相比于把所有词性都当成是候选词性或者把概率最高的若干个词性当作是未登录词的候选词性要更加可靠,有效.

由于 HMM 本身只考虑单向的词性转移,会失掉很多词性间的信息,所以要考虑更多的词性间的关系.对测试语料中的每句话都双向计算,选出其中相同的标注作为正确的标注,剩下的再利用规则跟三元词性转移选出最大概率的词性.这样不仅解决了 HMM 单向依赖的问题,而且通过选取两条最佳路径进而二次选择,提高了结果的可靠性.表 3 给出了不同改进方法下,词性标注的正确率.

表 3　不同改进方法下的标注正确率

Table 3　Accuracy of different methods

标号	方法介绍	词性标注正确率
1.	baseline	93.03%
2.	1 + 迭代训练语料	93.84%
3.	2 + 词语相似度计算	94.58%
4.	3 + 不加规则的二次计算	95.13%
5.	3 + 有规则的二次计算	95.65%

方法 1 是 baseline,仅利用初始 5.25M 的训练语料进行的传统 HMM 标注.表中的"+"表示当前的方法是在前一个方法的基础上,加上后面的改进部分,如:"1 + 迭代训练语料"是指将方法 1 中的语料进行迭代训练后进行的实验;"2 + 词语相似度计算"是指方法 3 是在方法 2 的基础上,加入词语相似度计算的方法.从该表中可以看出,迭代的扩大训练语料,计算词语相似度及 HMM 二次计算都能有效的提高词性标注的正确率.

4.3　不同实验方法对比

为了方便对比,本文把宾州中文树库 5.0 的部分语料作为测试语料进行了相同的实验,对于同样基于 HMM 的汉语词性标注方法,进行了对比,见下页表 4.

文献 [7] 在未登录词处理上使用的是基于单字的未登录词的词性选择方法,考虑词中每个字的作用,但未对每个字的

重要程度加以区分. 文献[14]利用 EM 算法处理时收敛速度慢,且一些参数的选择对测试语料有很强的依赖性. 本文利用基于词语的相似度计算,有效地选择了未登录词的词性,并且不存在参数的选择过程.

表 4 不同实验比较

Table 4 Comparation of different experiments

实验方法	训练方式	测试语料	精确率
文献[7]	有监督	宾州中文树库 5.2	95.11%
文献[14]	半监督	宾州中文树库 6.0	94.78%
本文	半监督	宾州中文树库 5.0	95.48%

由于不同实验的训练语料,标注集,及一些辅助工具,如词典等都不尽相同,以上的对比不能完全体现不同方法的优劣性,但还是可以作为部分参考.

4.4 部分实验结果分析

表 5 给出了各类词性的正确率. 数词、量词、由于存在词形 比较容易辨认,搭配固定等优势,标注正确率相对比较高.

表 5 不同词性的词的正确率

Table 5 Distribution of tagging errors on different types

词性	普通名词	动词	形容词
正确率	95.11%	93.26%	96.74%
词性	连词	数词	专有名词
正确率	94.98%	98.46%	98.04%
词性	副词	介词	量词
正确率	96.68%	93.04%	96.46%

规则处理阶段加入了常见人名词典,专有名词的识别率也较高. 名词和动词识别率较低,主要是因为未登录词中,这类词本身占有很高的比例;同时,汉语本身的语言特性使得这两类词性兼类情况也较为严重.

5 结 论

提出了一种基于半监督隐马尔科夫的汉语词性标注方法. 通过迭代训练方式不断扩大训练语料,把辅助训练语料分成 10 组,先通过初始训练语料对第一组辅助训练语料进行标注,将可靠的标注结果放入训练语料中形成新的训练语料,并用该新生成的训练语料作为下一步的训练语料,每组标注后都将结果中可靠的部分加入训练语料,由此不断扩大训练语料. 实验结果表明,该方法比起传统 HMM 标注,汉语词性标注的正确率提高了 0.8%,并节省了人工标注语料的时间. 本文还首次将 Word2vec 引入词性标注任务中,用于计算词语相似度,以便获取未登录词的候选词性,不仅减少了未登录词候选词性的个数,也使得候选词性的选择更加可靠,使用了词语相似度的标注结果比没有使用词语相似度时的结果提高了约 0.7%. 在 HMM 标注时,通过向前向后两种方式去标注词性,克服了 HMM 只是单向依赖的缺点,又利用语法规则,确保了词性选择的正确性,通过反复迭代进行二次计算,使词性标注的正确率又提高了 1%. 通过上述方法,汉语词性标注的正确率提高了 2.60%.

但是 HMM 只是一个二元的模型,很难利用长距离的词语关系;同时,HMM 的参数中又只利用到了词性到词的发射概率,弱化了词对词性的分布情况. 在后续的研究中,我们拟考虑三元 HMM 模型,并且利用特定词语之间的固定搭配或共现情况,以进一步提高标注效果.

References:

[1] Sun Jing, Li Jun-hui, Zhou Guo-dong. An unsupervised Chinese part-of-speech tagging approach using conditional random fields. [J]. Computer Applications and Software, 2011, 28(4): 21-24.

[2] Song H J, Son J W, Noh T G, et al. A cost sensitive part-of-speech tagging: differentiating serious errors from minor errors [C]. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, Association for Computational Linguistics, 2012: 1025-1034.

[3] Garrette D, Baldridge J. Type-supervised hidden markov models for part-of-speech tagging with incomplete tag dictionaries [C]. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012: 821-831.

[4] Sun W, Uszkoreit H. Capturing paradigmatic and syntagmatic lexical relations: towards accurate Chinese part-of-speech tagging [C]. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, Association for Computational Linguistics, 2012: 242-252.

[5] Jiang Wei, Guan Yi, Wang Xiao-long. Conditional random fields based pos tagging. [J]. Computer Engineering and Applications, 2006, 42(21): 13-16.

[6] Li Ze-zhong, Huang De-gen. Chinese pos tagging based on maximum entropy model with CRFs [C]. CIPS-ParsEval-2009, Beijing, 2009: 1-6.

[7] Huang Z, Harper M P, Wang W. Mandarin part-of-speech tagging and discriminative reranking [C]. Conference on Empirical Methods in Natural Language Processing Conference on Computational Natural Language Learning (EMNLP-CoNLL), 2007: 1093-1102.

[8] Li Ze-zhong. Chinese pos tagging employing maxent and word clustering [D]. Dalian: Dalian University of Technology, 2010.

[9] Cheng A, Xia F, Gao J. A comparison of unsupervised methods for part-of-speech tagging in Chinese [C]. Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010: 135-143.

[10] Hong Ming-cai, Zhang Kuo, Tang Jie, et al. A Chinese part-op-speech tagging approach using conditional random fields [J]. Computer Science, 2006, 33(10): 148-151.

[11] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. ArXiv Preprint ArXiv: 1301.3781, 2013.

[12] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C]. Advances in Neural Information Processing Systems. 2013: 3111-3119.

[13] Zong Cheng-qing. Statistical natural language processing (The second edition) [M]. Beijing: Tsinghua University Press, 2008.

[14] Huang Z, Eidelman V, Harper M. Improving a simple bigram HMM part-of-speech tagger by latent annotation and self-training [C]. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. Association for Computational Linguistics, 2009: 213-216.

附中文参考文献:

[1] 孙 静,李军辉,周国栋. 基于条件随机场的无监督中文词性标注 [J]. 计算机应用与软件,2011,28(4):21-24.

[5] 姜 维,关 毅,王晓龙. 基于条件随机域的词性标注模型 [J]. 计算机工程与应用,2006,42(21):13-16.

[6] 李泽中,黄德根. 基于最大熵模型结合 CRFs 的汉语词性标注 [C]. 第一届汉语句法分析评测研讨会,北京,2009:1-6.

[8] 李泽中. 最大熵结合词语聚类的中文词性标注研究 [D]. 大连:大连理工大学,2010.

[10] 洪铭材,张 阔,唐 杰,等. 基于条件随机场(CRFs)的中文词性标注方法 [J]. 计算机科学,2006,33(10):148-151.

[13] 宗成庆. 统计自然语言处理(第 2 版) [M]. 北京:清华大学出版社,2008.