doi:10.11887/j.cn.201404015 http://journal.nudt.edu.cn

主动学习与自学习的中文命名实体识别*

钟志农,刘方驰,吴 烨,伍江江
(国防科技大学 电子科学与工程学院,湖南 长沙 410073)

摘 要:命名实体识别是信息抽取中的一项基础性任务,如何利用丰富的未标注语料来提高实体识别的指标是该领域一个重要的研究方向。基于条件随机场提出一种将主动学习与自学习相结合的方法——SACRF,通过设置置信度函数和2-Gram频度阈值来选取样本,并采用人工与自动相结合的方式进行标注来扩展训练语料。实验表明,该方法在提高实体识别的精确率和召回率的同时,能够显著地降低人工标注的工作量。

关键词:主动学习;自学习;条件随机场;命名实体识别
中图分类号:TP316 **文献标志码**:A **文章编号**:1001-2486(2014)04-0082-07

Chinese named entity recognition combined active learning with self-training

ZHONG Zhinong, *LIU Fangchi*, *WU Ye*, *WU Jiangjiang*
(College of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073, China)

Abstract: Named Entity Recognition (NER) is a basic task in information extraction, and it is an important research direction in this domain to use the abundant unlabeled corpus to improve the performance of NER system. An approach combining self-training with active learning based on CRF (SACRF) is proposed. It selected samples by setting the threshold of confidence and 2-Gram frequency, and expanded the training set by annotating the unlabeled corpus manually and automatically. The experiments revealed that this approach can not only improve the precision and recall of NER system, but also reduce the manually annotation efforts greatly.

Key words: active learning; self-training; conditional random fields; named entity recognition

命名实体识别是信息抽取的一项关键子任务,是指在非结构化文本中提取出代表现实世界中具体或抽象的实体(人名、地名、组织机构名……)的词或者词串。命名实体识别的主要方法有基于规则和统计两类,其中基于统计的方法由于其在处理大规模语料中表现出的优异性能而成为主流方法。在所有统计方法中,有监督方法使用人工标注的训练语料训练模型来抽取命名实体,性能最好,但这类方法过分依赖于标注语料,而大规模标注语料的获取是非常困难的。与此相对应,未标注语料是充足的,规模巨大的。如何有效地利用未标注语料对于提高命名实体识别系统的实用性具有巨大的意义。

常见的利用未标注语料来提高学习性能的方式有两种:半监督学习和主动学习策略。在半监督学习中,学习器自行利用未标记实例改善训练模型的泛化能力,整个学习过程不需人工干预。自1994年Shahshahani[1]使用未标注语料对半监督学习进行研究以来,已有大量的学者对相关技术进行了研究,如Olivier[2]提出的自学习方法,Blum等[3]提出的协同训练方法以及Tri-Training[4]、CoFroest[5]、CoTrade[6]等提出的协同学习改进方法,Blum[7]和Zhu[8]等提出的图的分割方法等。

主动学习的思想是分类器自动从未标注语料中选取最有训练价值的实例提交人工标注并加入到训练集中改善分类器的性能。主动学习可以避免花费代价去标注对进一步学习帮助不大的样本,从而可以缩减分类模型的规模和训练时间。主动学习的思想已经成功地应用于自然语言处理领域,如构建语料库[9],文本分块[10]等。

这两种方法尽管在有效利用未标注语料进行实体识别方面取得良好的效果,但各自存在问题。半监督方法不需要人工参与,在不断扩展训练实例的同时也会将错误累积,使最终的分类器精度降低。主动学习依赖于人工标注,当面对大规模

* 收稿日期:2013-10-11
基金项目:国家高技术研究发展计划主题项目(2011AA120300);湖南省自然科学基金资助项目(11JJ4028)
作者简介:钟志农(1975—),男,湖南新邵人,副教授,博士,硕士生导师,E-mail:nongnudt@gmail.com

语料时,其代价是不可接受的。一种有意义的思路是将两种思想结合,在英文处理领域,已有部分学者开始这方面的尝试[11-12],但在中文的命名实体识别领域相关研究还很少。本文提出的 SACRF(Self-training with Active learning based on CRF)方法以条件随机场(Conditional Random Fields,CRF)作为基础分类器,将主动学习与自学习方法结合,在保证识别精度的同时极大地降低了人工标注的代价。

1　预备知识

1.1　基于条件随机场地命名实体识别

命名实体识别可以看作是一个标注问题。条件随机场[13]是一种优秀的解决序列标注问题的模型,并且可以避免最大熵等模型会产生的标记偏置问题。

使用条件随机场进行命名实体识别有以下几个过程:确定标注序列集,获得特征函数集,制定特征模版,模型参数估计,数据预测。标注集一般采用{B,I,E,O}符号集来标注实体的首字、中间部分、最后一个字符以及非实体部分。特征函数是一个二值函数,将上下文特征、词性特征、外部特征等进行数字化表示。特征模版为特征函数的生成提供一个统一的模式,可以方便选择所需要的特征以及特征间的组合。后续步骤通过选定的特征对训练集训练生成模型进行实体的预测。

1.2　主动学习

主动学习是一种样本选择技术,能够通过筛选最有标注价值的样本来减少人工标注工作量,同时可以减小分类器学习时的计算规模。其理论依据是通过降低统计学习的期望误差率来对训练数据优化选择[14]。未知数据的期望错误率表示为

$$\int_X \mathrm{E}[(\hat{y}(x,D) - y(x))^2 \mid x]p(x)\mathrm{d}x \quad (1)$$

其中 D 表示标注的训练样本集合,x 是一个未标注的样本,$y(x)$ 是它的标注结果。如果将期望错误率切分,可以分为三部分:噪声产生的偏差,学习器的偏差以及学习器的方差。可以形式化的证明,主动学习可以降低后两项偏差,因此可以降低期望错误率。

1.3　自主学习

在自主学习中,首先利用标记数据集训练出初始分类器,使用该分类器对一些未标记数据进行标记,将可信度最高的一些标记新示例放入到标记数据集中再在新标记数据集上进行下一次训练直到满足截止条件为止。自学习同主动学习的区别在于样本的选择,前者是选出当前模型所能够区分的具有最大可信度的样本作为标注样本进行下一轮迭代,后者则是选出当前模型不确定性最大的样本来进行人工标注扩展。相比较而言,自学习属于半监督学习,主动学习则属于有监督学习。由于自学习的初始分类器是一个弱分类器,因此不可避免地会在迭代过程中不断地累积错误,但是自学习是最简单的半监督方法,可以方便地与其他方法组合使用。

2　主动学习与自学习结合的实体识别

主动学习方法可以充分利用人工标注的结果扩充训练集,并且大大减少了人工工作量,但是该方法返回的句子中并不是每个字符对于训练都有着积极的意义。在每个句子中有相当大的部分的字符(词语)是非实体词,同时有很多字分类器已经给出了正确的分类结果,将所有字符进行重新标记是对人力的严重浪费。基于此,我们结合主动学习与自学习的优势,提出了基于 CRF 的实体识别算法 SACRF,来减少人工标注的工作量。该算法的框架如图 1 所示。算法以主动学习框架为基础,以 CRF 作为分类器,迭代地从未标注扩展语料库中抽取标注价值大的样本,并采用自学习与人工标注混合的标注方式将样本加入训练集中,扩展分类模型。

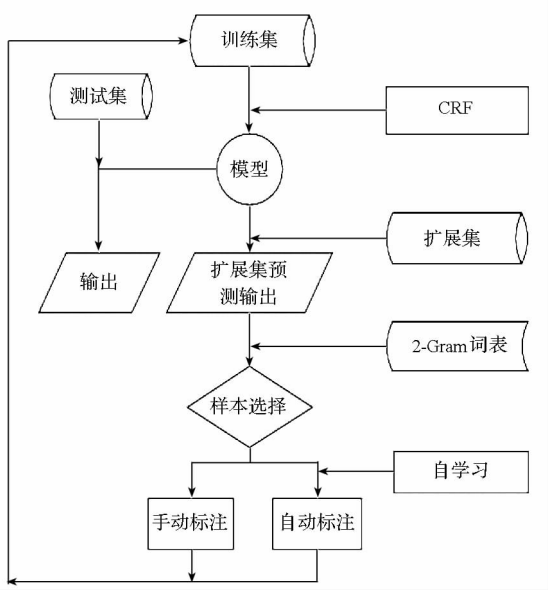

图 1　SACRF 算法框架

Fig. 1　Framework of SACRF

2.1 基于置信度的主动学习

在实体识别阶段,为了避免分词工具的错误带来的影响,基于字的识别策略,我们使用标注{PER-B, PER-I, LOC-B, LOC-I, ORG-B, ORG-I, N},分别表示人名、地名、组织机构名的第一个字和后面的字符以及不属于上三类实体的字。使用CRF对标注语料训练,并对未标注语料的每个字进行预测,根据式(1)计算字 w 属于7个不同标注的概率 $p_i(w)$ $(i=1,2,3)$。在主动学习过程中,我们使用基于置信度的学习策略,使用字 w 的最大预测概率 $p_0=\max p_i(w)$ 作为置信度。如果 p_0 小于某个阈值,则认为当前模型没有足够的信心来判断当前字的类别,因此该字具有较大的标注价值。

一般的基于置信度的主动学习先计算每个字的置信度,然后对一个句子的所有字进行加权求和来计算句子的置信度[15]。但是由于一个句子中大部分字符都是非实体字,因此对所有字进行加权求和来返回整个句子并不能凸显出有可能是实体词的字符的特殊性。我们采取的策略是首先设定一个置信度阈值 ε, 当语句L中有某个字 $conf(w_i)<\varepsilon$ 时,则认为L具有较高的训练效用,将L作为候选样本加入候选池。

2.2 样本选择

并不是所有候选池中的样本都会作为最终的标注样本输出,因为每个字都是在一定的上下文环境中出现的,如果某个字的置信度很低但是在所有的语料中出现的频率非常低,就没有必要将其进行标注。因此,我们在系统中引入二元语法模型,首先获取了所有语料的2-Gram统计词表,对于L中每个满足 $conf(w_i)<\varepsilon$ 候选字 w_i,查询其前后两个字与它本身((w_{i-1}, w_i), (w_i, w_{i+1}))在语料库中出现的频度之和 f,如果 f 大于设定的阈值 f_0,则认为 (w_{i-1}, w_i, w_{i+1}) 这三个字符具有较大的标注价值,从池中选择L作为标注样本。综合以上所述,扩展集中语句L的判别函数可以表示为

$$f(L)=\sum_{w_i\in L}\operatorname{sgn}(conf(w_i)-\varepsilon)\times\operatorname{sgn}(fre(w_{i-1},w_i)+fre(w_i,w_{i+1})-f_0) \quad (2)$$

其中

$$\operatorname{sgn}(x)=\begin{cases}1 & x>0\\ 0 & \text{其他}\end{cases}$$

如果 $f(L)>0$, 则选择L作为标注样本。

2.3 自学习与人工混合的样本标注

对于某个样本,很大一部分子序列使用当前的模型已经能够准确地预测,这时完全可以采用自学习的方法对这部分序列进行自动标注。具体来说,对于样本L,将集合 $W=\sum(w_{i-1},w_i,w_{i+1})$ 提交人工标注,剩余的字符集合为

$$V=\sum_{w_j\in L\&w_j\notin W}w_j \quad (3)$$

依次计算 $conf(w_j)$,如果 $conf(w_j)>\mu$(μ 为设定的置信度阈值),使用模型对 w_j 的预测自动标注,否则提交人工标注。

算法的流程如下:

输入:少量的标注语料L,大量的未标注语料U,CRF特征模版T,2-Gram词表G,置信度阈值 ε,$\mu(\mu>\varepsilon)$,频度阈值 f_0,集合 H,W。

过程:

1. 选取适当的特征,使用模版T,将L在CRF上进行训练,获得模型M;
2. 将U以句子为单位进行切分,并使用M对U进行预测,获得三元组 $<w, tag, p>$ 集合,其中,w 是字符,tag 是预测结果,p 是预测概率;
3. 对U中的每个句子L,根据式(2)计算 $f(L)$,若 $f(L)>0$,将 L 加入 H,将 (w_{i-1}, w_i, w_{i+1}) 加入集合 W;
4. 对H中每个样本L中的字 v 进行判别,如果 $conf(v)>\mu$,并且 $v\notin W$,则将 v 的 tag 作为 v 最终的标注,否则提交人工标注;
5. 将H从U中抽取出来,加入到L中;
6. 迭代上述过程直到U为空或性能指标收敛。

输出:训练模型M。

---

3 实验与分析

实验采用1998年1月的《人民日报》(约3万句)作为语料库,该语料库由北大富士通标注,是中文处理常用的标注语料。实验中将其分为三部分:训练集,测试集,扩展集。其中,初始训练集和测试集都是1000句语料,根据实验内容的不同,扩展集规模会有变化。实验首先对语料做简单的预处理,主要是将原来的标注改为{PER-B, PER-I, LOC-B, LOC-I, ORG-B, ORG-I, N}标注集。由于扩展集应为未标注集,在实验中没有使用扩展集的标注特性。

实验中条件随机场的训练和测试采用工具包CRF++,为避免分词错误对实验结果的影响,实验基于单个字来进行训练和测试。选取的上下文窗口大小为2,实验中采用的模版及其对应的意义如表1所示。

表 1　模版及其意义

Tab. 1　Templates and meanings

模　版	意　义	模　版	意　义
U00:%x[-2,0]	当前字前第二个字	U05:%x[-2,0]/%x[-1,0]/%x[0,0]	前两个字与当前字组合
U01:%x[-1,0]	当前字前第一个字	U06:%x[-1,0]/%x[0,0]/%x[1,0]	前后两字与当前字组合
U02:%x[0,0]	当前字	U07:%x[0,0]/%x[1,0]/%x[2,0]	后两个字与当前字组合
U03:%x[1,0]	当前字后第一个字	U08:%x[-1,0]/%x[0,0]	前一字与当前字组合
U04:%x[2,0]	当前字前第二个字	U09:%x[0,0]/%x[1,0]	后一字与当前字组合

模版的意义是选取当前字的上下文窗口内的字以及这些字的组合作为特征。实验中采用的性能指标包括常用的准确率 P、召回率 R、F 值以及总的标注量。具体定义为

$$\begin{cases} P = N_1/N_2 \\ R = N_1/N_3 \\ F = \dfrac{(\beta^2+1)\times P\times R}{R+\beta^2\times P}(\beta=1) \end{cases} \tag{4}$$

式中,N_1 是标注正确的实体数,N_2 是标注的实体数,N_3 是实际的实体数。我们进行了多组对比实验来证明方法的有效性,实验中当输出的总的 F 值收敛时(前后两次 F 值之差小于 0.01)停止迭代。

实验一

第一组对比实验,将 SACRF 方法(方法一)、随机主动学习方法(方法二)、完全自学习方法(方法三)、扩展集完全标注训练预测(方法四)进行对比,扩展集规模为 5000 句,以直接用训练集模型对测试集进行预测的结果作为基准。方法一的参数设置为 $\varepsilon = 0.5$, $\mu = 0.99$, $f_0 = 30$,方法二每次迭代过程中随机选取与方法一相同数量的样本,通过主动学习的方式进行标注扩展到训练集中,方法三每次选择置信度大于 0.99 的样本直接加入到训练集中,方法四则将所有的扩展集标注,训练生成模型并进行预测。实验结果如表 2,其中 P、R、F 以及总的 F 值 F_{all} 都是取百分数,W 代表万。

表 2　实验一数据

Tab. 2　Data of experiment 1

方法	P			R			F			F_{all}	选择样本	标注量(字)
	PER	LOC	ORG	PER	LOC	ORG	PER	LOC	ORG			
一	91.83	87.26	85.26	44.29	53.73	62.67	59.76	66.51	72.24	66.87	806	1.77W
二	89.67	87.54	84.66	37.77	42.66	53.41	53.15	57.36	65.50	58.90	806	5.63W
三	84.25	88.88	73.98	26.90	29.66	30.81	40.78	44.48	43.50	43.45	1362	0
四	90.14	88.58	91.38	57.20	61.74	64.67	69.99	72.76	75.73	73.01	5000	35.12W
BL	83.47	88.65	73.86	27.44	28.96	32.38	41.30	43.66	45.02	43.57	0	0

图 2 是三种实体总的 F 值与迭代次数的关系,横坐标为迭代次数,纵坐标为 F 值。

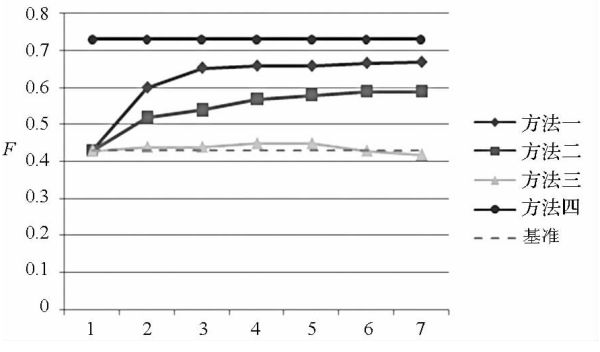

图 2　各种方法 F 值对比

Fig. 2　F-measure of different methods

从以上的实验数据可以看出,简单的自学习方法实验指标基本与基线相同,甚至有些指标比基线还要低,这是因为自学习仅依据训练集扩展,所增加的样本的训练效用与训练集类似,同时由于训练集本身的覆盖率有限,在扩展过程中不可避免地会引入错误,因此反而会使 F 值降低。方法二同基线相比,性能有了大幅提高,这是因为加入了主动学习这一过程,但是随机选择样本会遗漏一些标注效用大的样本,以及重复选择与训练集有相似训练效用的样本,因此方法二选择的 806 条样本的"含金量"并不高。SACRF 克服了随机选取的两个弊端,并使用了自学习来进行自动标注,因此在相同样本数量的前提下 F 值比方法二高出 8 个百分点,标注

量却仅是方法二的1/3。方法四与基线相比 F 值提高了30%,SACRF 比基线提高了24%,但是后者的标注样本仅是前者的17%,标注工作量仅是前者的5%。也就是说,通过 SACRF 方法选择出的17%的标注样本对于提高预测性能贡献率为80%,而需要人工标注的字符仅是所有字符中的5%,这突出反映出 SACRF 的有效性。

实验二

第二组对比实验,对比不同置信度阈值 μ 对于实验性能和标注量的影响。参数设置为 $\varepsilon = 0.5$, $f_0 = 30$,扩展集的规模为7000,实验结果如表2所示。图3表示的是总的 F 值和标注量与迭代次数的关系。

表3 实验二数据

Tab. 3 Data of Experiment 2

μ	P			R			F			F_{all}	选择样本	标注量(字)
	PER	LOC	ORG	PER	LOC	ORG	PER	LOC	ORG			
1.0	91.48	87.90	85.26	46.73	58.19	59.62	61.87	70.02	70.17	68.51	1375	76782
0.99	93.00	85.74	86.47	43.34	55.61	65.19	59.12	67.47	74.34	67.89	1236	24583
0.98	92.81	85.03	84.43	42.11	53.14	53.62	57.94	65.41	65.59	64.04	1261	20078
0.95	93.01	85.92	85.23	37.91	47.87	50.99	53.86	61.49	63.81	60.68	1083	13317

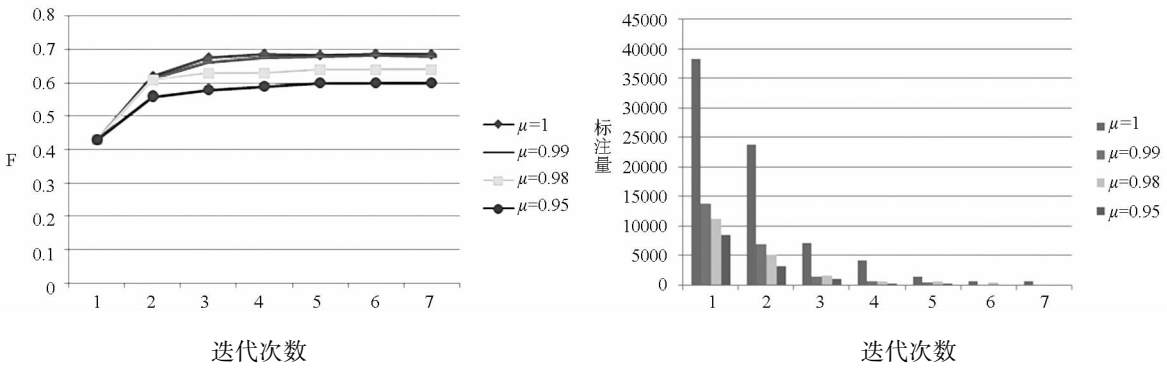

图3 不同 μ 值对应的 F 值与标注量

Fig. 3 F-measure and annotation number of different μ

μ 的作用在于控制系统通过自学习进行自动标注的比例,如果使用传统的主动学习方法,即不自动标注而完全使用手工标注,可以设置 $\mu = 1$。通过观察发现,通过自学习进行自动标注可以大大降低人工标注量。当阈值为0.99时,F 值曲线与完全标注基本重合,而标注量仅为完全标注的1/3。阈值为0.98时,标注量仅为完全标注的四分之一左右,而进一步降低阈值所降低的标注量则并不明显。但是减少标注量会造成 F 值的降低,阈值为0.98时同完全标注相比,F 值降低了3个百分点,这与它减少的巨大工作量相比是可以容忍的。相比较而言,阈值0.95会使 F 值降低较多(约8个百分点),同时标注量的减少并不是很可观,因此选择阈值 μ 并不是越小越好,通过实验发现,根据数据量的大小,μ 在0.97~0.99取值较合适。

实验三

第三组对比实验验证参数 ε 对于试验性能的影响,实验中,$\mu = 1$, $f_0 = 30$,扩展集规模为5000句,实验结果如表4所示。

表4 实验三数据

Tab. 4 Data of experiment 3

ε	P			R			F			F_{all}	选择样本
	PER	LOC	ORG	PER	LOC	ORG	PER	LOC	ORG		
0.3	91.76	86.93	85.38	33.28	41.10	47.31	48.85	55.81	60.89	55.91	180
0.4	93.41	87.59	85.46	42.39	50.83	61.82	58.31	64.33	71.75	65.45	485
0.5	88.23	87.45	84.72	46.87	55.07	57.72	61.22	67.58	68.66	67.04	857
0.6	92.11	88.84	85.27	50.81	59.91	63.93	65.49	71.56	73.07	70.65	1269
0.7	91.88	88.61	89.03	56.92	61.47	64.03	70.30	72.58	74.49	72.65	1587

参数 ε 可以控制需要标注的样本数量,ε 要满足 $\varepsilon > 1/7$,因为 CRF 选择的标注是 7 个概率值中最大的一个。如果 $\varepsilon = 1$,相当于将扩展集中所有的句子完全标注,就退化为实验二中的方法四。从以上数据中可以分析出,随着 ε 的不断扩大,标注的样本数基本是线性增长的,而 F 值的变化近似地服从对数函数曲线,增长越来越缓慢。根据实验二可知,当 ε 趋近于 1 时,其将会收敛于 73% 附近。在具体应用中可以根据指标要求和数据规模综合考虑 ε 的大小,如果结果要求精度较高且未标注积集规模不大,可以选择 0.6 或 0.7 等较大的值,但最好不要超过 0.7,因为继续增大的话,F 值提升的空间已经非常有限了。如果扩展集的规模较大,就应该首先考量人工标注量,此时宜选择略小的 ε 值。

实验四

第四组对比实验验证参数 f_0 对于试验性能的影响,实验中,$\varepsilon = 0.5$,$\mu = 1$,扩展集规模为 5000 句,实验结果如表 5 所示。

表 5　实验四数据

Tab. 5　Data of experiment 4

ε	P			R			F			F_{all}	选择样本
	PER	LOC	ORG	PER	LOC	ORG	PER	LOC	ORG		
0	91.46	87.57	89.87	50.95	57.92	61.61	65.44	69.72	73.11	69.81	1064
10	90.84	87.68	85.94	49.86	58.14	61.09	64.38	69.91	71.41	69.25	983
20	87.87	88.31	88.01	48.23	56.85	58.67	62.28	69.17	70.41	68.56	911
40	91.48	87.78	86.76	46.73	55.23	60.67	61.87	67.81	71.41	67.67	808
60	93.20	88.13	85.71	46.60	53.89	60.56	62.51	66.88	70.09	66.94	743
100	95.05	86.85	85.11	44.42	51.47	59.51	60.55	64.64	70.04	65.40	660
200	93.15	85.64	84.41	36.95	45.51	55.83	52.91	59.43	67.21	60.46	464

参数 f_0 的作用在于甄选出在整个语料库中经常出现并且标注效用高的样本,我们认为,出现频率不同的样本对于提高 F 值的贡献率是不同的。如果模型对于某个样本信心非常低,即通常意义上的“标注效用”高,但该样本只出现为数不多的几次(我们称之为“一次性样本”),那么对其进行标注将对后面的预测没有太大的帮助。为了验证该假设,我们可以计算相应频度范围内的样本 w 对于提高的 F 值的平均贡献度 γ_w。γ_w 的计算公式为

$$\gamma_w = \frac{F_i - F_j}{C_i - C_j} \times 1000\% \tag{5}$$

其中 F_i 是 $f_0 = i$ 时的 F 值,C_i 是 $f_0 = i$ 时标注的样本数,w 是频度位于 $[i, j)$ 内的样本,当 $f_0 \to \infty$ 时,选择的标注样本为 0,根据实验一,$F_\infty = 43.57\%$。由以上数据可以得到表 6。

表 6　不同区间对应的贡献度

Tab. 6　Contribution of different frequency interzone

区间	[0 - 10)	[10 - 20)	[20 - 40)	[40 - 60)	[60 - 100)	[100 - 200)	[200 - ∞)
γ_w	0.069	0.095	0.086	0.112	0.185	0.252	0.364

从上表可以看出一个整体的趋势,出现频度越高的样本对于提高 F 值的贡献度越高,从而验证了假设的正确性。可以预见,规模越大的语料,这个规律体现得越明显。因此通过选取适当的 f_0,可以在保持 F 值稳定的同时进一步降低标注工作量。在本实验中 f_0 的数值是根据《人民日报》语料库设置的,f_0 取经验值 20 ~ 40 比较适合,但在实际应用中应根据语料库的规模灵活调整。

通过以上 4 个实验可以发现,当迭代次数达到 4 次的时候,F 值已经趋于收敛,当面对的数据规模巨大时,训练模型的时间是很长的,但实际过程中没有必要等到 F 值完全收敛,可以根据经验值只进行适当次数的迭代即可。

4　结论

本文提出一种将主动学习与自学习相结合的中文命名实体识别方法:SACRF,该方法以 CRF 作为基础分类器,迭代地从未标注语料中提取样本,标注后扩展到训练集中进行训练。在样本选

择中,采用置信度函数和2-Gram频度结合的方式,筛选出最具有标注价值的样本。主动学习过程中,将自学习的自动标注与人工标注相结合,极大地降低了手工标注的工作量,使得较大规模数据的主动学习具有了可操作性。通过实验验证了方法的有效性,并对方法中使用的三个参数进行了分析讨论。

由于本文是为了验证算法的有效性,选取的特征仅是上下文的字特征,因此预测的准确率、召回率都相对偏低,为了进一步提高指标,可以采用加入词性、外部词典等特征或者扩充特征模版等方法。同时,后续工作还要在不同的数据集上进一步验证和改进算法,使得算法具有更强的普适性和健壮性。

参考文献(References)

[1] Shahshahani B, Landgrebe D. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon[J]. IEEE Transactions on Geoscience and Remote Sensing, 1994, 32(5):1087 - 1095.

[2] Chapelle O, Schölkopf B, Zieneds A. Semi-supervised learning[M]. Cambridge: The MIT Press, 2006.

[3] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the Eleventh Annual Conference on Computational Learning Theory. New York, NY: ACM, 1998: 92 - 100.

[4] Pise N N, Kulkarni P. A survey of semi-supervised learning methods[C]//International Conference on Computational Intelligence and Security. Washington, DC: IEEE Computer Society, 2008: 30 - 34.

[5] Li M, Zhou Z H. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples[J]. IEEE Transactions on Systems, Man and Cybernetics, 2007, 37(6):1088 - 1098.

[6] Zhang M L, Zhou Z H. CoTRADE: confident co-Training with data editing[J]. IEEE Transactions on Systems Man and Cybernetics Part B (Cybernetics), 2007, 7(3):753 - 760.

[7] Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts[C]//Proceedings of the 18th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers, 2001:19 - 26.

[8] Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using gaussian fields and harmonic functions[C]//Proceedings of the 20th International Conference on Machine Learning, Washington, DC, 2003:912 - 919.

[9] Engelson S P, Dagan I. Minimizing manual annotation cost in supervised training from corpora[C]//Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 1996:319 - 326.

[10] Ngai G, Yarowsky D. Rule writing or annotation: Cost-efficient resource usage for base noun phrase chunking[C]//Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2000:117 - 125

[11] Yu D, Varadarajan B, Deng L, et al. Active learning and semi-supervised learning for speech recognition: a unified framework using the global entropy reduction maximization criterion[J]. Computer Speech and Language, 2010, 24(3):433 - 444.

[12] Tomanek K, Hahn U. Semi-supervised active learning for sequence labeling[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Stroudsburg, PA: Association for Computational Linguistics, 2009:1039 - 1047.

[13] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.
LI Hang. Statistical learning methods[M]. Beijing: Tsinghua University Press, 2012. (in Chinese)

[14] Bikel D M, Miller S, Schwartz R, et al. Nymble: a high performance learning name-finder[C]//The 5th Conference on Applied Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 1997:194 - 201.

[15] 陈霄. 基于支持向量机的中文组织机构名识别[D]. 上海: 上海交通大学,2007.
CHEN Xiao. Chinese organization names recognition based on support vector machine[D]. Shanghai: Shanghai Jiao Tong University, 2007. (in Chinese)