基于最近邻的主动学习分词方法

梁喜涛 顾 磊
(南京邮电大学计算机学院 南京 210003)

摘 要 分词是中文自然语言处理中的一项关键基础技术。为了解决训练样本不足以及获取大量标注样本费时费力的问题,提出了一种基于最近邻规则的主动学习分词方法。使用新提出的选择策略从大量无标注样本中选择最有价值的样本进行标注,再把标注好的样本加入到训练集中,接着使用该集合来训练分词器。最后在PKU数据集、MSR数据集和山西大学数据集上进行测试,并与传统的基于不确定性的选择策略进行比较。实验结果表明,提出的最近邻主动学习方法在进行样本选择时能够选出更有价值的样本,有效降低了人工标注的代价,同时还提高了分词结果的准确率。

关键词 中文分词,主动学习,不确定性取样,最近邻规则
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.6.048

Active Learning in Chinese Word Segmentation Based on Nearest Neighbor

LIANG Xi-tao GU Lei
(School of Computer Science & Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract As the basis of Chinese information processing, Chinese word segmentation (CWS) plays a very important role. To solve the problems of lacking of training samples and accessing a large number of labeled samples laboriously, a fresh active learning method based on nearest neighbor was proposed. The method adopts CRFs as the basic framework and uses the proposed active learning sampling strategy to select the most useful instances to annotate from a large number of unlabeled samples. Next the annotated are put instances into the labeled set and then the segmenter is trained by using the labeled set. Finally the method was tested in PKU corpora, MSR corpora and shanxi university corpora, and compared with the uncertainty sampling strategy. The experiment result shows that the fresh active learning selection strategy can select more valuable samples, reduce the cost of manual annotation effectively, and improve the accuracy of segmentation.

Keywords Chinese word segmentation, Active learning, Uncertainty sampling, Nearest neighbor rule

中文分词是中文信息处理中的重要基础问题,在机器翻译、信息检索、汉字识别、语音合成等诸多领域有着广泛应用。传统方法多基于人工词典和需要大规模标注语料的统计模型,虽然已经取得了很大的成绩,但无论编写词典,还是标注语料库,都需要大量人工劳动[1]。尽管表现较好的分词系统准确率能达到95%~97%,但一般都需要标注大规模语料来对系统模型进行训练,而获取大量标注样本是一件非常费时费力的工作。因此要达到正确率的要求,传统的监督学习方法(即被动学习)需要付出很大的人工代价。机器学习中的主动学习方法应运而生,用于解决这类标注样本较少的情况[2]。

主动学习方法已被用于词性标注、文本分类、句法分析等诸多自然语言处理任务中,但把主动学习方法应用在中文分词领域还不多见[3]。2002年,Sassano M[4]提出了一种基于支持向量机的主动学习分词算法,该算法是针对日语进行分词,把整个句子当作基本的标注单元,需要人工标注的量比较大,对SVM模型进行训练也需要大量时间。2004年,张健沛[5]提出了一种基于到超平面距离最近策略的ASVM算法,每次迭代选取一个离SVM分类超平面最近的样本点,认为它的类别最不确定,也最有可能被分错,信息量最大,所以它最有可能改变分类超平面的位置,而远离超平面的样本点对其位置的改善影响不大[6]。把主动学习与SVM相结合可以高效利用未标注样本,建立最有价值的训练集,由此得到的分词器也能很好地继承SVM较强的泛化性能。2012年,Shoushan Li等[7]提出了一种最小化数据获取代价的中文分词方法,该方法以CRF为基本框架,采用基于不确定性取样的主动学习选择策略(Active learning Based on Uncertain Sampling, ALUS),仅仅注释最不确定的字符边界,人工标注的代价得到大大降低。

本文第1节主要讲述了文献[7]的ALUS分词方法;第2节介绍了本文使用的基于最近邻规则的主动学习分词方法

到稿日期:2014-07-06 返修日期:2014-10-09 本文受国家自然科学基金(61302157),教育部人文社会科学研究青年基金(12YJC870008),江苏省教育厅高校哲学社会科学基金(2013SJB870004),江苏省社科研究文化精品课题(12SWC-030)资助。
梁喜涛(1989-),男,硕士生,主要研究方向为中文信息处理,E-mail:centaolang@163.com;顾 磊(1978-),男,副教授,硕士生导师,主要研究方向为中文信息处理、机器学习,E-mail:gulei@njupt.edu.cn。

(Active learning Based on Nearest Neighbor, ALNN)；第 3 节分别采用 ALUS 和 ALNN 策略进行分词实验，比较得到的分词结果并对结果进行分析；最后总结了主动学习方法的优点，并提出了今后需要进一步研究的方向。

1 基于 ALUS 的选择策略

主动学习过程的关键是如何从大量的未标注样本中挑选最有价值的样本点进行标注，不确定性取样(ALUS)和委员会查询[8]是其中两种主要的选择方法。由于分词器是根据少量已标注样本进行训练得到的，学习到的知识是不充分的，判断未标注样本会存在不确定性，因此常用的主动学习样本选择方法是利用这种不确定性来进行取样，选择当前分词器最无法确定其类别的样本进行标注。

基于不确定取样策略(ALUS)由 Lewis[9]等人提出，算法选取当前分词器最不确定类别的样本交给专家进行标注。由于中文分词可看作序列标注任务，可以根据句子中每个位置的字符标注来判断何处应该被切分，从而把分词问题当作一种二分类问题，通过查询句子中每个边界的后验概率来计算边界的不确定值[10]，在此二分类问题中，ALUS 会挑选后验概率接近 0.5 的样本[11]。从几何意义上来看，这是选择接近分类边界的样本。

ALUS 策略对每个未标注样本点给出一个评价值来表示其不确定性，对这些样本的评价值进行排序，挑选不确定值较高的样本进行标注。标注后的样本点会加入到已标注样本集中，继续训练分词器，直到人工标注的数量达到给定阈值。定义

$$f(x_i)=\max_{y\in\{0,1\}} p(y|I_i)-0.5 \tag{1}$$

来计算每个边界的不确定值，其中 x_i 代表第 i 位置的字符，I_i 代表字符 x_i 的右边界，$y=0$ 表示 x_i 后不应被切分，$y=1$ 表示 x_i 后应被切分。$f(x_i)$ 值越低表示 $p(y|I_i)$ 的值越接近 0.5，系统越不确定 x_i 右边界是否应被切分。文献[7]同时还利用了差异性测量方法 $d(x_i)=f_{c_ic_{i+1}}$，该方法利用边界的上下文二元字符来判断字符间的差异性。在训练过程中，检查所有未标注数据的字符边界，c_ic_{i+1} 代表边界 I_i 的上下文字符，$f_{c_ic_{i+1}}$ 表示二元字符 c_ic_{i+1} 出现的概率，初始化为 0，当遇到相同的 c_ic_{i+1} 时，$f_{c_ic_{i+1}}=f_{c_ic_{i+1}}+1$。把不确定选择策略 $f(x_i)$ 和差异性测量方法 $d(x_i)$ 组合在一起来进行样本选择，即通过公式

$$S(x_i)=f(x_i)d(x_i) \tag{2}$$

对未标记集合 U 中所有字符计算 $S(x_i)$，集合 U 中每个字符边界根据该值进行排序，然后再从 U 中挑出一些最不确定的边界来进行人工标记。

2 基于 ALNN 的分词方法

主动学习分词方法一般可分为两部分：学习部分[12]和选择部分[13]。ALNN 方法的选择部分首先对字符进行预处理，用字符间的相关属性特征来表示每一个字符。然后使用最近邻选择策略通过对字符特征进行处理来从未标记集合 U 中选择样本进行标记，把标记后的样本加入到 L。学习部分就是一个基本的分词器，使用训练集合 L 来训练分词器。接下来对这几部分进行详细介绍。

2.1 字符预处理

在对文本进行分词之前，首先对待训练和测试文本进行一系列的预处理，将其转化为机器学习方法易于处理的形式。把训练样本和测试样本用一系列刻画其相关属性的特征来表示，用 3 种统计量特征来表示每个训练和测试样本。

第一类特征是词位统计特征，根据训练语料中已标注好的词位标记统计出其中每个字符可以作为边界字符出现的概率大小，以此来判断相邻字符的结合程度。第二类特征是互信息特征，它是计算当前字符与后面下一个字符同时出现的频率大小，将这个频率与当前字符单独出现频率进行比较，根据比值大小来判断当前字符与后一字符是否组成词语[14]。例如有序字符串 xy 之间的互信息定义为：

$$I(xy)=\log_2\frac{p(xy)}{p(x)p(y)}=\log_2\frac{n(xy)*n}{n(x)*n(y)}$$

其中，$p(xy)$ 为字符串 xy 出现的概率，$p(x)$ 为字符 x 出现的概率，$p(y)$ 为字符 y 出现的概率，$n(xy)$ 表示字符串 xy 的出现次数，$n(x)$ 表示字符 x 的出现次数，$n(y)$ 表示字符 y 的出现次数，n 表示字符的总数目[15]。第三类特征是 t 测试值，通过公式计算来比较当前字符与它前面一个字符的结合能力以及与后面一个字符的结合能力，以此来判断它到底与哪个字符结合得更紧密，更能组成一个词语[16]。对于有序字符串 xyz，字符 y 相对于 x 及 z 的 t 测试定义为

$$t_{x,z}(y)=\frac{p(z|y)-p(y|x)}{\sqrt{\sigma^2(p(z|y))+\sigma^2(p(y|x))}}$$

其中，$p(z|y)$ 表示字符 z 对字符 y 的条件概率，$p(y|x)$ 是字符 y 对字符 x 的条件概率，$\sigma^2(p(z|y))$ 和 $\sigma^2(p(y|x))$ 则代表各自的方差。

特征表示中互信息体现了相邻字符结合关系的紧密程度[17]。如上所述，如果 $I(xy)>0$，则表示 xy 间是正相关的，随着 $I(xy)$ 增加，x 和 y 的相关度增加，当 $I(xy)$ 大于给定的阈值，也就是 x 和 y 的紧密程度高于某一阈值时，便可认为此字符组合 xy 构成了一个词；如果 $I(xy)=0$，则 xy 间是不相关的；如果 $I(xy)<0$，则 xy 间是互斥的，表示 xy 间基本不会结合成词。t 测试值体现了当前字符与前后字符相连的趋势，从 t 测试的定义可知当 $t_{x,z}(y)>0$ 时，字符 y 与后继字符 z 有相连的趋势，$t_{x,z}(y)$ 值越大，相连趋势越强；当 $t_{x,z}(y)=0$ 时，不反映任何趋势；当 $t_{x,z}(y)<0$ 时，字符 y 与它的前一个字符 x 有相连的趋势，$t_{x,z}(y)$ 值越小，相连趋势越强。互信息和 t 测试值各有特点，分别反映了字符与字符之间的静态结合能力和动态结合能力，具有一定程度的互补性。因此把这两个特征结合起来，构造一组互补的特征组合，可以形成更趋合理的统计数据[18]。利用上述的 3 个特征来刻画训练语料和测试语料中的字符表示，然后通过对这 3 个字符特征进行处理来进行样本选择。

2.2 基于最近邻规则的选择策略

最近邻(Nearest Neighbor, NN)规则[19]是机器学习中比较常用的一种方法，主要应用在文本分类、模式识别、图像及空间分类等领域，在中文分词领域还不多见。它假设一个未知样本的类别与它附近样本的类别保持一致，故可通过近邻所带来的局部信息进行学习。其中近邻是指样本一定距离范围内的邻域样本集合。

本文将 NN 规则用于中文分词领域，先利用初始分词器

对未标记语料 U 中样本进行初步切分，根据切分好的结果对其样本点进行特征表示。为了消除样本点特征数值单位不一致而影响比较的情况，在对其进行处理之前先对它们进行归一化处理，使之处在指定的范围内。假设样本点 e 的特征表示为 $e=\{e_1,e_2,e_3\}$，归一化后的特征 $e_0=e/|e|=\{e_1/|e|,e_2/|e|,e_3/|e|\}$，其中 $|e|=\sqrt{e_1^2+e_2^2+e_3^2}$ 表示向量 e 的模。以 U 中每个样本 x_i 为球心、r 为球半径，定义一个球体子空间(Sphere)[20]。然后计算 L 中的每个样本点是否在这个子空间里，在子空间里面的样本点都已经标注了类别，这些样本点就构成了一个近邻集合，用这些样本点的类别信息来表示 x_i 的不确定性。$x_i\in U$ 的近邻集合为

$$S(x_i)=\{x_j\mid x_j\in L,d(x_i,x_j)\leqslant r\} \tag{3}$$

其中，$d(x_i,x_j)=(\sum_{m=1}^{3}(|f_{im}-f_{jm}|^2))^{\frac{1}{2}}$ 表示样本点 x_i 和样本点 x_j 之间的欧氏距离，f_{im} 和 f_{jm} 分别表示样本 x_i 和样本 x_j 对应的词位、互信息和 t 测试值特征。r 越大，则越多与 x_i 接近的样本就会被归到 x_i 的邻域集合中，但这会降低未标记样本之间的差异，不利于找到合适样本点。如果 r 值太小，即邻域范围太小，则搜索过程会陷入局部搜索。所以半径 r 的定义十分关键，在此将其定义为 $r=ave\_dis(x_i)\times\theta$，其中 $ave\_dis(x_i)$ 表示 U 中所有样本点欧氏距离的平均值，θ 表示一个比例因子[21]。后文将采用实验对 θ 值进行分析。

信息熵是对事物状态有序性的一种度量，某一事物状态不确定性的大小与该事物可能出现的状态数目以及各状态出现的概率有关。熵值越大，说明系统中的数据越无序，系统越杂乱；熵值越小，则说明系统中的数据越有序，系统越纯净[22]。因此利用信息熵 $Entropy(S(x_i))=-\sum_{j=1}^{m}p_j\log_2 p_j$ 来衡量上述每个邻域集合内样本点概率分布的不确定性程度，其中 m 表示分类数目，p_j 表示集合 $S(x_i)$ 中的样本属于第 j 类的比例。由以上分析可知，分词问题可以表示成一种二分类问题，即对每个字符标记为 1 或者 -1，所以此处 m 取值为 2，即

$$\begin{aligned}Entropy(S(x_i))&=-\sum_{j=1}^{2}p_j\log_2 p_j\\&=-(p_1\log_2 p_1+p_2\log_2 p_2)\end{aligned} \tag{4}$$

$Entropy(S(x_i))$ 的值越大，表示样本集合的不确定性程度越大，样本 x_i 被选择进行人工标记的可能性越大。

假设样本点 x_a 的邻域集合内存在 1 个正例和 1 个负例，而样本点 x_b 的邻域内只存在 1 个正例。利用式(4)计算 x_a 和 x_b 的近邻集合熵值，分别为 1 和 0，这表示 x_a 比 x_b 的不确定性要大，样本挑选算法更倾向于选择 x_a 进行人工标记，即选择不确定性大的进行人工标记。把标记后的样本加入到 L 中来降低局部区域的不确定性。从几何上来看，熵值越大越靠近类边界。

此外，本文在进行样本选择时，除了考虑标注样本的最近邻集合熵值外，还通过计算每个未标识样本同训练集合的欧氏距离平均值

$$d(x_i)=\sum_{j=1}^{n}(\sqrt{\sum_{m=1}^{3}(|f_{im}-f_{jm}|^2)})/n,x_i\in U,x_j\in L \tag{5}$$

来增加样本集合的多样性，以缓解或者消除重复标注对分词器的影响，其中 n 表示训练集合 L 中样本的个数，m 表示词位、互信息和 t 测试值这 3 类特征，f_{im} 和 f_{jm} 分别表示样本 x_i 和样本 x_j 对应的词位、互信息和 t 测试值特征。$d(x_i)$ 衡量了样本 $x_i\in U$ 被选为样本后样本集 L 的多样性，$d(x_i)$ 值越大表示样本 x_i 与训练集合 L 差异性越大，样本的冗余度越小，越能避免重复标注问题[23]。

如果仅仅把样本的最近邻集合熵值作为样本选择的唯一依据，在进行样本选择时则很容易出现冗余。通过利用未标识样本同训练样本的欧氏距离平均值 $d(x_i)$ 可以有效避免甚至消除样本选择时产生的冗余。为了更好地发挥两者的性能，本文引入了参数 $\lambda(0\leqslant\lambda\leqslant1)$ 来平衡这两部分，利用公式

$$f(x_i)=\lambda\times Entropy(S(x_i))+(1-\lambda)d(x_i),x_i\in U \tag{6}$$

来进行样本选择。通过此方法可以在选择最有价值样本的同时避免冗余问题，有效克服样本冗余对分词器的影响。

2.3 ALNN 分词方法的学习部分

本文 ALNN 分词方法中的学习部分选用 CRF++[24]作为基本的分词器。CRF 是一种专门用于对序列数据进行切分和标注的概率框架，它克服了最大熵模型和 HMM 模型的标记偏置问题，可以更好地模拟现实世界的数据[25]。把 CRF 应用在中文分词系统中，最关键的就是选择标记集和特征模板。本文使用的是 BMES 4 位标记集，利用该标记集标注后分词问题就可以转化成序列标注问题[26]，这就使 CRF 进行中文分词变成了可能。句子中每个字符根据其在词中出现的位置给予不同的标记，其中 B(begin)代表词的开始部分，M(middle)代表词的中间部分，E(end)代表词的结尾部分，S(single)代表单字词。其中标点作为单独成词的字符来处理，以 S 标记其词位。

例如：迈向/充满/希望/的/新/世纪。

标注后为：BE BE BE S S BE S。

另外可以按照字符后是否应该被切分开，把上面的 4 分类问题看作一种二分类问题，字符标记为 B 和 M 的表示该字符后不应被切分(标记为 -1)，标记为 E 和 S 的表示应当被切分(标记为 1)。

在 CRF 模型中选取特征模板是最为关键的部分，它决定了识别的正确率。在特征模板的选取上，一般选择字符本身及字符的上下文特征信息，所选取的特征要尽可能体现所识别对象的特点[27]。本次实验中采用的 CRFs 特征模板来自 CRF++包，如表 1 所列。

表 1 CRFs 特征模板

# Unigram	# Bigram
U00:%x[-2,0]	B
U01:%x[-1,0]	
U02:%x[0,0]	
U03:%x[1,0]	
U04:%x[2,0]	
U05:%x[-2,0]/%x[-1,0]/%x[0,0]	
U06:%x[-1,0]/%x[0,0]/%x[1,0]	
U07:%x[0,0]/%x[1,0]/%x[2,0]	
U08:%x[-1,0]/%x[0,0]	
U09:%x[0,0]/%x[1,0]	

该特征模板中的特征分为 Unigram 和 Bigram 两类。Unigram 模板：第一个字符为“U”。其中，“U01:”或“U02:”是为了区分这些特征而加的标识符；模板中每一行代表一个

模板，在每个模板里，特定的宏%x[row,col]用来描述输入数据的片段(文字片段，如词等)，row 表示当前片段的相对位置，col 则表示列的绝对位置。Bigram 模板：第一个字符为“B”，这个模板用来描述双连词的特征[28]。通过这个模板，当前输出片段和前一个输出片段会自动结合在一起。对于例句“迈向充满希望的新世纪”，假设当前的字符为“希”，则每个模板对应的特征表示如表 2 所列。

表 2　特征模板及其对应的特征表示

Template	expanded feature
%x[−2,0]	充
%x[−1,0]	满
%x[0,0]	希
%x[1,0]	望
%x[2,0]	的
%x[−2,0]/%x[−1,0]/%x[0,0]	充/满/希
%x[−1,0]/%x[0,0]/%x[1,0]	满/希/望
%x[0,0]/%x[1,0]/%x[2,0]	希/望/的
%x[−1,0]/%x[0,0]	满/希
%x[0,0]/%x[1,0]	希/望

2.4　ALNN 分词方法过程

基于最近邻规则的主动学习分词方法的详细过程如下：

已标注样本集 L，未标注样本集 U，CRF＋＋分词器，每次采样的样本个数 M，终止条件 S。

(1)用 L 训练 CRF＋＋分词器，对 U 中样本点进行初始切分。

(2)反复执行(a)－(d)，直到满足终止条件 S。

(a)对每一个 $x_i \in U$，利用式(6)计算其不确定性值 $f(x_i)$；

(b)根据 $f(x_i)$ 在 U 中选择包含 M 个样本的子集 A，A 中样本具有最大的不确定性值；

(c)由人工对 A 中的样本进行标注，并从 U 中删除子集 A 包含的样本，再将标注后的样本加入到 L 中；

(d)用 L 继续训练 CRF＋＋分词器，并对 U 中样本进行切分。

(3)输出从最终样本集 L 上训练得到的 CRF＋＋分词器。

因为文献[7]中没有详细说明算法的迭代结束条件，所以本文以挑选进行人工标注的样本个数达到事先设定的比例作为算法的终止条件 S，即人工标注的样本满足一定数目时停止下一次的样本挑选。M 值的大小为 200，即迭代过程中每次选择 200 个样本进行人工标注并加入到标注集合 L 中。ALUS 方法的流程也与上述过程一致，只是采用式(2)来计算不确定性值。

3　实验

3.1　实验准备

为了验证 θ 取值对分词准确率的影响，分别在 PKU、MSR 和 SX(山西大学)数据集上进行了验证。分别设定比例因子 θ 为 0.1，0.2，0.3，0.4，0.5，0.6，0.7，0.8，0.9，1.0 进行熵值计算和实验，实验过程如 2.4 节所述，但仅通过式(4)进行不确定取样，在每个 θ 值迭代过程中每次选取 200 个样本进行人工标注，人工选取的样本总数为 1000。实验结果如图 1 所示。

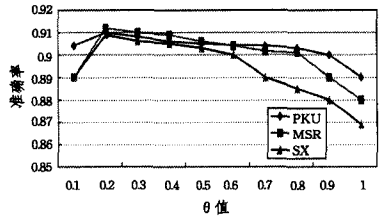

图 1　不同 θ 值下的分词准确率

由图 1 分析可知，在 θ 取 0.2 左右时分词准确率最大。另外还可看出随着 θ 值增大，分词准确率逐渐下降，分析原因主要是随着 θ 值增大，样本近邻集合中的元素增多，降低了未标记样本之间的差异，不利于找到合适样本点。

另外对于式(6)中权数 λ 的取值情况，本文在 PKU、MSR 和 SX(山西大学)数据集上也采用同样的方法进行验证，如图 2 所示。先设定 θ 值为 0.2，然后分别设定 λ 为 0.1，0.2，0.3，0.4，0.5，0.6，0.7，0.8，0.9，1.0 进行式(6)的计算，选择最有价值的样例。人工选取和标注的样本总数是每个 λ 值对应数据集样本总数的 2%，迭代过程每次选取 200 个样本进行人工标注。可以看出，在其它参数相同的情况下，当 λ=0.5时，分词效果较好。

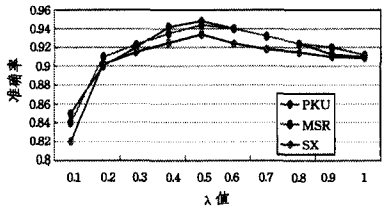

图 2　不同 λ 值下的分词准确率

在下面的实验中分别采用 ALUS 和 ALNN 两种策略从未标注语料集合 U 中选择最不确定的样本进行标注，采用 PKU、MSR 和 SX 数据集作为实验测试数据集。其中 ALNN 策略中 θ 取 0.2，λ 取 0.5。利用 ALUS 和 ALNN 两种策略对这 3 个测试集中的未标注字符边界按照不确定性高低进行排序，挑选不确定值最大的样本，对选出的样本进行人工标注后，把已标注的样本加入到标注集合 L 中，再利用 L 对 CRF＋＋进行训练，依次进行直到人工标注的个数达到设定的比例。PKU 数据集、MSR 数据集和 SX 数据集设定的人工标注的比例均为：2%，5%，8%，11%和 14%。对每个数据集中设定的比例分别进行 10 次重复实验，$F1$ 值取这 10 次分词结果的平均值。

3.2　实验结果

选用系统 $F1$ 值作为衡量系统性能的指标，$F1$ 定义为 $F1=2PR/(P+R)$，其中 P=(正确切分的词组数/系统切分的词组数)×100%，表示分词系统的整体准确率；R=(正确切分的词组数/切分文本中的词组数)×100%，表示系统的整体召回率。为了公平比较 ALUS 和 ALNN 这两种选择取样策略的效果，这两种策略从大量未标注集合 U 中挑选的样本数目始终保持一致，选择最有价值的样本进行人工标注的数目也始终保持一致。两者在 PKU、MSR 和 SX 数据集上的分词结果如表 3 所列。

表 3　不同标注比例下 ALUS 方法和 ALNN 方法在测试集上的 $F1$ 值

		2%	5%	8%	11%	14%
PKU	ALUS	0.933	0.945	0.955	0.962	0.967
	ALNN	0.948	0.955	0.961	0.967	0.969
MSR	ALUS	0.935	0.948	0.959	0.966	0.969
	ALNN	0.944	0.959	0.964	0.968	0.971
SX	ALUS	0.918	0.935	0.950	0.959	0.966
	ALNN	0.934	0.945	0.956	0.965	0.969

将表 3 绘制成折线图能直观看出这两种策略的差异，如图 3－图 5 所示，其中横坐标表示人工标记边界的字符所占的比例，纵坐标表示 $F1$ 值。

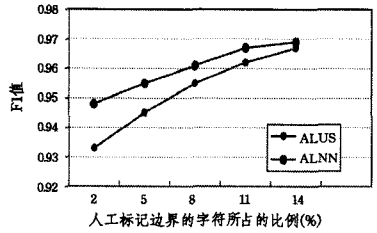

图 3　PKU 数据集上的分词结果

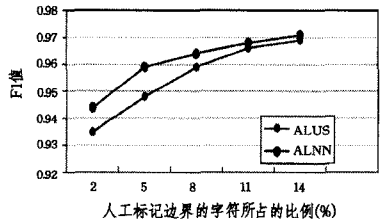

图 4　MSR 数据集上的分词结果

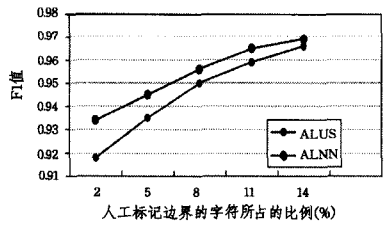

图 5　SX 数据集上的分词结果

由表 3 和图 3－图 5 可以看出人工标注的字符比例越小，ALNN 策略与 ALUS 策略两者在这 3 个数据集上的 $F1$ 值差距越大。在字符比例为 2%时，PKU 数据集和 SX 数据集上 $F1$ 值差距达到了 1.5 个百分点，MSR 数据集上 $F1$ 值差距也将近 1 个百分点。这恰能说明 ALNN 策略在选择有价值样本方面的优越性，在进行相同比例样本选择时能够选出更有价值的样本来提高分词器的性能。尤其在人工标记字符比例较小时，两者从未标注集合 U 中挑选相同数目样本进行人工标注所花费的人工标记代价相同，但 ALNN 方法可以达到更好的效果。随着人工标注字符比例的增加，从未标注样本 U 中选择出的有价值样本越来越多，U 中剩余的有价值样本也就越来越少，这两种策略在这 3 个数据集上的 $F1$ 值增长幅度都开始逐渐减缓，在达到特定字符比例时，U 中有价值的样本已经所剩无几，两者的 $F1$ 值差距也越来越小。但从图 3－图 5 仍然可以看出在字符比例为 2%，5%，8%，11%和 14%时，ALNN 策略的效果都要优于 ALUS 方法，因此可以说明本文提出的 ALNN 选择策略进行样本选择的表现要优于 ALUS 策略。分析原因，主要是本文使用的最近邻规则更加关注其近邻的类别程度，也就是更趋于找到类边界附近的样本，从而提高最近邻分词器的性能，这比 ALUS 方法仅仅考虑词位频率特征来计算不确定性更加符合字符本身的特点。

结束语　本文把主动学习方法成功应用在中文分词领域，首先对文本字符进行特征表示，然后通过最近邻规则来计算字符的不确定性，同时利用未标注字符与训练集 L 的欧氏距离来消除冗余。文中提出的 ALNN 方法利用未标注样本 U 内部的信息，在每次的迭代中，对未标注样本循序渐进地进行学习和标注，通过对训练集的有效扩充，逐渐提升了 CRF＋＋分词器的性能。在 PKU 数据集、MSR 数据集和 SX 数据集上的实验结果表明，提出的 ALNN 方法能用较少的人工和时间代价达到较好的分词效果。接下来将继续优化本文中的主动学习选择策略，使其能很好地应用在微博这种训练语料少的短文本领域中；另外还准备把主动学习和半监督学习方法结合起来，把工作重点转向中文分词领域中姓名、地名、组织机构名的识别以及复合名词的识别领域。

参 考 文 献

[1] Lai Si-wei, Xu Li-heng, Chen Yu-bo. Chinese Word Segment Based on Character Representation Learning[J]. Journal of Chinese Information Processing, 2013, 27(5): 8-14

[2] Liu Kang, Qian Xu, Wang Zi-qiang. Survey on Active Learning Algorithms[J]. Computer Engineering and Applications, 2013, 48(34): 1-4

[3] Feng Chong, Chen Zhao-xiong, Huang He-yan. Active Learning in Chinese Word Segmentation Based on Multigram Language Model[J]. Journal of Chinese Information Processing, 2006, 20(1): 50-58

[4] Ranganathan K, Iamnitchi A, Foster I. Improving data availability through dynamic model-driven replication in large peer-to-peer communities[C]// 2002 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid. IEEE, 2002: 376-376

[5] Zhang Jian-pei, Xu Hua. Study and Application of Active Learning with SVM [J]. Computer Applications, 2004, 24(1): 1-3

[6] Sassano M. An Empirical Study of Active Learning with Support Vector Machines for Japanese Word Segmentation [C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002: 505-512

[7] Li Shou-shan, Zhou Guo-dong, Huang Chu-ren. Active Learning for Chinese Word Segmentation[C]// Proceedings of COLING 2012. Posters, COLING, Mumbai, December 2012: 683-692

[8] Settles B. Active learning literature survey[R]. Computer Sciences Technical Report 1648. University of Wisconsin-Madison, 2009

[9] Lewis D D, Gale W A. A sequential algorithm for training text classifiers[C]// Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval Springer-Verlag. New York, Inc., 1994: 3-12

[10] Huang Chu-ren, Yo Ting-shuo, Šimon P, et al. A Realistic and Robust Model for Chinese Word Segmentation[C]// Proceedings of the Conference of Computational Linguistics and Speech Processing(ROCLING-08). 2008

（下转第 261 页）

Electronic Science and Technology Photonic Sensors, 2009, 38 (5):537-543

[13] Newman M E J. Random graphs with clustering [J]. Phys Rev Lett, 2009, 103:058701

[14] Newman M E J. Detecting community structure in networks [J]. Eur Phys J B, 2004, 38(2):321

[15] Sales-Pardo M R, Guimera A, Moreira A, et al. Extracting the hierarchical organization of complex systems [J]. Proc Natl Acad Sci USA, 2007, 104:15224

[16] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of community hierarchies in large networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 10:10008

[17] Shen H, Cheng X, Cai K, et al. Detect overlapping and hierarchical community structure in networks [J]. Physica A, 2009, 388: 1706-1712

[18] Zhang S, Wang R S, Zhang X S. Identification of overlapping community structure in complex networks using fuzzy c-means clustering [J]. Physica A, 2007, 374:483-490

[19] Zhang S, Wang R S, Zhang X S. Uncovering fuzzy community structure in complex networks [J]. Phys Rev E, 2007, 76(4): 046103

[20] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Phys Rev E, 2004, 69(2):026113

[21] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Phys Rev EstatNonlin Soft Matter Phys, 2004, 69(6):06133

[22] Kernighan B W, Lin S. A efficient heuristic procedure for partitioning graphs [J]. Bell System Technical Journal, 1970, 49(2): 291-307

[23] Fiedler M. Algebraic connectivity of graphs [J]. Czech Math J, 1973, 23(98):298-305

[24] Pothen A, Simon H, Liou K-P. Partitioning sparse matrices with eigenvectors of graphs [J]. SIAM J Matrix Anal Appl, 1990, 11 (3):430-452

[25] Zachary W W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research, 1977, 33:452-473

[26] Michael J H. Labor dispute reconciliation in a forest products manufacturing facility[J]. Forest Products Journal, 1997, 47:41-45

[27] Lusseau D. The emergent properties of a dolphin social network [J]. The Royal Society, 2003, 270(Suppl):186-188

(上接第 232 页)

[11] Ju Sheng-feng, Wang Zhong-qing, Li Shou-shan, et al. A Comparative Study on Different Active Learning Strategies for Sentiment Classification[C]// Advances of Computational Linguistics in China(2009-2011). 2011:506-511

[12] Song H, Yao T. Active Learning Based Corpus Annotation [C] // IPS-SIGHAN Joint Conference on Chinese Language Processing. Beijing, China, 2010:28-29

[13] Long Jun, Yin Jian-ping, Zhu En, et al. An Active Learning Algorithm by Selecting the Most Possibly Wrong-Predicted Instances[J]. Journal of Computer Research and Development, 2008, 4(3):472-478

[14] Church K W, Hanks P. Word association norms, mutual information and lexicography[J]. Computational linguistics, 1990, 16 (1):22-29

[15] Liu Bin, Huang Tie-jun, Cheng Jun, et al. A New Statistical-based Method in Automatic Text Classification[J]. Journal of Chinese Information Processing, 2002, 16(6):18-24

[16] Zhu xiao-juan. The Researeh on Chinese Word Segmentation System Based on SVM[D]. Central South University, 2007

[17] Yan Hui, Zhang Xue-gong, Li Yan-da. Kenal-based maximal-margin clustering algorithm[J]. Journal of Tsinghua University (Natural Sciences), 2002, 42(1):36-38

[18] Dai Y, Loh T E, Khoo C S G. A new statistical formula for Chinese text segmentation incorporating contextual information[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1999:82-89

[19] Cover T, Hart P. Nearest Neighbor Pattern Classification[J]. IEEE Trans on Information Theory, 1967, 13(1):21-27

[20] Zhao Ying, Liu Hong-xing, Wang Zhong-yu, et al. An Improved Nearest Neighbor Searching Method for Classification Problems [J]. Journal of Nanjing University: Natural Sciences, 2009, 45 (4):455-462

[21] Wang Zhen-yu, Wang Xi-zhao. Active Learning Algorithm Based on Neighborhood Entropy[J]. Pattern Recognition and Artificial Intelligence, 2011, 24(1):97-102

[22] Carter T. An introduction to information theory and entropy [EB/OL]. http://astarte. csustan-edu/tom/ SFICSSS

[23] Bai Long-fei, Wang Wen-jian, Guo Hu-sheng. A noval Support Vector maching Active Learning strategy[J]. Journal of Nanjing University: Natural Sciences, 2012, 48(2):182-189

[24] http://crfpp. sourceforge. net/

[25] Qiu Sha, Wang Fu-yan, Shen Hao-ru, et al. Chinese Named Entity Recognition Based on Part of Speech Feature with Edges [J]. Computer Engineering, 2012, 38(13):128-130

[26] Li Shou-shan, Huang Hu-ren. Word Boundary Decision with CRF for Chinese Word Segmentation [C]// 23th PACLIC. 2009: 726-732

[27] Xiao Qin, Liang Zong, Wu Yu-qian, et al. CRF-based Experiments for Cross-Domain Chinese Word Segmentation at CIPS-SIGHAN-2010 [C]// Proceedings of CLP 2010. 2010

[28] Zhong Ke-li, Zhou Xue, Li Hang-yu, et al. Cascaded Chinese Weibo Segmentation Based on CRFs[C] // Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing. Tianjin, China, 2012:69-73