文章编号：1003-0077（2011）00-0000-00

基于λ-主动学习方法的中文微博分词*

张婧，黄德根，黄锴宇，刘壮，孟祥主

（大连理工大学 计算机科学与技术学院，辽宁省 大连市 116024）

摘要：由于面向中文微博的分词标注语料相对较少，导致基于传统方法和深度学习方法的中文分词系统在微博语料上的表现效果很差。针对此问题，本文提出一种新的主动学习方法从大规模未标注语料中挑选更具标注价值的微博分词语料。该方法根据微博语料的特点，在主动学习迭代过程中引入参数λ来控制所选的重复样例的个数，确保了所选样例的多样性；同时，根据样例中字标注结果的不确定性和上下文的多样性，采用 Max、Avg 和 AvgMax 三种策略衡量样例整体的标注价值；此外，用于主动学习的初始分词器除了使用当前字的上下文作为特征外，还利用字向量自动计算当前字成为停用字的可能性作为模型的特征。实验使用 NLPCC 2015 公开的训练语料和测试语料，结果表明，本文提出的基于主动学习的分词方法，其 F 值较基线系统提高了 0.84%~1.49%，与目前最优的 WBA 主动学习方法相比提升效果更加显著。

关键词：中文分词；主动学习；样例多样性；微博语料

中图分类号：TP391　　**文献标识码：**A

Enhancing Microblog-oriented Chinese Word Segmentation with λ-Active Learning Method

ZHANG Jing, HUANG Degen, HUANG Kaiyu, LIU Zhuang, MENG Xiangzhu

(School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116024, China)

Abstract: The manual segmented microblog-oriented corpora are inadequate, which is the reason that the performance of both the conventional Chinese word segmentation (CWS) systems and the deep learning based CWS systems is still unsatisfactory. To address this problem, we propose a novel active learning method to effectively select samples with high annotation value from unlabelled tweets for microblog-oriented CWS. Considering to the characteristics of microblog data, the parameter λ is introduced in the procedure of measure the context diversity of the characters to control the number of the repeatedly selected samples. Furthermore, three strategies (Max, Avg and AvgMax) are also used to evaluate the overall values of a sample taking advantages of the uncertainty confidence and the context diversity of the characters in the sample. In addition, for the feature construction of the initial segmenter, both the basic context of the current character and the probability of the current character being a stop character which is calculated via character embeddings are taken into consideration. Our experiments are conducted on the benchmark datasets released by NLPCC 2015 for the shared task of CWS for microblog text. The results demonstrate that the λ-active learning method we proposed obviously outperforms the baseline system with a gain of 0.84%~1.49% as well as the-state-of-the-art active learning method Word Boundary Annotation (WBA).

Key words: Chinese word segmentation; active learning; diversity of samples; microblog-oriented data

1 引言

微博等社交媒体数据承载着大量舆情信息及商业信息，其传播的实时性高，速度快，影响力深远。近年来，面向微博等社交媒体语料的自然语言处理任务受到广泛关注，例如面向微博的情感分析[1]、微博领域命名实体识别[2,3]、热点事件抽取[4]等，中文分词是上述任务的

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金项目(61672127，61672126）

首要步骤。目前，很多优秀的分词系统在传统语料（例如新闻和专利）上的分词效果取得了很大进展[5-7]。然而，与传统语料上的分词效果相比，面向微博语料的分词效果显著下降，有的 F 值甚至会下降 10 个百分点[8,9]。为了促进面向微博领域中文分词的研究，自然语言处理领域会议增加了面向微博的中文分词评测任务，例如 NLPCC[8,9]，COAE 等会议。

面向微博语料的分词任务的难点在于：微博语料用语和句式都比较随意，且经常包含表情符号、URL 等噪音；微博语料涵盖的内容比较宽泛，通常包含较多的不相关的主题，即微博语料本身具有跨领域性。因此，面向微博语料的中文分词任务与面向传统语料的中文分词任务相比，分词器模型对训练语料的依赖程度更大。通过构建充足的涵盖范围广的训练语料，可以有效提高面向微博的中文分词效果。在扩充训练语料时，为了尽可能减少人工标注的工作量，主动学习方法被广泛用于挑选更具有标注价值的语料[10-13]，其主要思路为：通过现有的训练语料训练得到初始分类器，之后利用初始分类器对未标注语料进行标注，根据初始分类器的标注结果计算样例的标注价值，选择标注价值较高的语料进行人工修正，将修正后的语料加入到现有的训练语料中重新训练分类器，经过多次迭代，直到训练语料达到预期的规模时，停止迭代，获得最终的分类器。

为了获得更具标注价值的面向微博领域的分词语料，本文采用主动学习方法进行语料挑选。目前用于中文分词任务的主动学习方法有：文献[13]提出一种词边界标注模型 Word Boundary Annotation（WBA），根据 CRFs 标注结果的边缘概率对字符标注结果的不确定性进行评价，文献根据字符是否是词语右边界，将 CRFs 标注集 B、M、E、S（分别表示词首、词中间、词尾、单字词）划分为两类：B、M 为一类，表示该字符不是词的右边界，不需要进行切分，记为 N；E、S 为一类，表示该字符为某个词的右边界，需要进行切分，记为 Y；之后，分别计算这两类标签的边缘概率与阈值（阈值=0.5）的差值，选择两个差值中的较高者作为该字符标注结果的不确定性，其计算公式如下：

$$H(c)=\max_{x\in\{N,Y\}}P_x(c)-0.5 \qquad (1)$$

其中，$P_x(c)$ 表示字符 c 被标注为 x 的后验概率。$H(c)$ 的值越低，该字符的边界信息越不确定，该字符的标注价值越高。

文献[14]提出了一种基于最邻近规则的主动学习方法（Active Learning based on Nearest Neighbor, ALNN），使用基于最邻近规则衡量字标签的不确定性，该方法在进行样本选择时，除了考虑标注样例的最近邻集合熵值外，还通过计算每个未标识样例同训练集合的欧氏距离来增加样本集合的多样性。

文献[15]提出了一种基于置信度的主动学习分词算法，将所有置信度高于设定阈值的样本放入样本池中，然后从样本池中选择一定数量的样本进行人工标注。

主动学习方法在选取语料时，需要考虑所选样例的多样性，上述主动学习方法虽然在面向传统语料的中文分词任务中取得了显著成果，但无法合理有效地处理微博语料，因为微博语料中存在大量局部内容相同但整体不同的句子，如：

样例 a：#微评#弱势群体怎能被“弱视”？

样例 b：#微评#拍饭有风险！手滑需谨慎！

假设样例 a 中的“微”或“评”作为主动学习方法选择语料时的考察对象，那么样例 b 中考查对象的上下文和样例 a 中的上下文相同，没有差异，现有的主动学习方法对样例的差异性要求过于严苛，进行样例选择时，样例 b 通常会被过滤掉，导致样例 b 中的其他有价值的信息无法被考察。

为此，本文提出一种 λ-主动学习方法，该方法在主动学习选择语料的过程中，引入参数 λ 来控制所选的重复样例的个数，确保所选样例的多样性，并采用多种策略对样例整体的标注价值进行衡量。实验表明，本文新提出的方法既能有效避免重复性标注工作，又可以提高分词效果，比基线系统和目前最佳的主动学习方法之一 WBA 方法都有显著提高。

2 基于半监督方法的初始分词器

基于主动学习的中文分词方法首先需要构建初始分词器。由于大规模未标注语料中存在很多同现信息和字边界信息（例如标点符号通常是前词的右边界，是后词的左边界），因此，我们从大规模未标注语料中抽取统计信息，构建基于半监督方法的初始分词器，半监督方法能够有效利用未标注语料来提高分词效果[16,17]。实验中，我们使用条件随机场 CRFs 模型作为初始分词器。在训练 CRFs 模型时，除了使用窗口为 5 的上下文特征和是否是标点的特征外，还使用了从未标注语料中自动提取的点互信息 PMI 和停用字相似度信息作为半监督学习的特征。

2.1 点互信息

点互信息（Pointwise Mutual Information, PMI）是一种用来度量关联性的统计量，在本文任务中，使用 PMI 来衡量字与字之间的共现程度，PMI 计算公式如下：

$$PMI(x,y)=\log\frac{P(x,y)}{P(x)P(y)} \qquad (2)$$

其中，x、y 表示语料中的字，$P(x,y)$ 表示 x 和 y 作为相邻二元字串出现的概率，$P(x)$、$P(y)$ 分别表示 x 和 y 在整个语料中出现的概率。

根据上述公式，从未标注语料中统计所有二元字串的点互信息。构造训练语料时，考虑到 CRFs 模型学习的特征是离散的，我们将二元字串的点互信息进行向上取整，使用 $\left(C_0,\lceil PMI(C_{-1},C_0)\rceil\right)$ 和 $\left(C_0,\lceil PMI(C_0,C_1)\rceil\right)$ 作为当前字的特征。C_0 表示当前字，C_{-1} 表示当前字的前一个字，C_1 表示当前字的后一个字。

2.2 停用字相似度

由于词向量的提出，使得在无监督的条件下获得语料中词或字的语义信息成为可能[18,19]。在训练 CRFs 模型时，除了使用 PMI 这个统计量，我们还利用 word2vec 模型在未标注语料上训练得到字向量，利用字向量构建了停用字集合，进而得到了当前字与停用字集合的相似程度。下面分别介绍停用字集合构建方法和相似度计算方法。

构建停用字集合时，我们人工收集了高频停用字作为种子集合。根据实验效果，种子集合中保留了 11 个常用的标点和单字词：{“我”，“是”，“的”，“了”，“在”，“。”，“，”，“、”，“；”，“！”，“？”}。我们提出按照公式（3）的方法，利用词向量计算当前字与种子集合的相似度，进而扩展种子集合。

$$AvgSim\left(token,SCset^N\right)=\frac{1}{N}\sum_{i=1}^{N}sim(token,chara_i) \qquad (3)$$

其中，N 是集合 $SCse$ 中元素的个数，$chara_i$ 是集合 $SCse$ 中的第 i 个字串；$sim(token,chara)=\frac{\boldsymbol{CE}(token)\cdot\boldsymbol{CE}(chara)}{\left|\boldsymbol{CE}(token)\right|\left|\boldsymbol{CE}(chara)\right|}$；CE($c$)是 c 的字向量，$\left|\boldsymbol{vector}\right|$ 是计算向量 $\boldsymbol{vector}$ 的模。

训练字向量时，本文使用基于 softmax 方法的神经网络模型 Skip-gram 训练得到字向量。实验中我们收集了 30 万条未标注的微博语料，将未标注的语料按字切分，将字作为神经网络模型的训练单位，训练得到字向量。训练参数为：维度=200，窗口=9，最低词频=1。

利用字向量和公式（3）计算得到当前字与种子集合相似度后，按照相似度值从高到低进行排序，选择相似度较高的前 M 个字加入到种子集合中。经过 T（本文实验中 $T=3$，此为经验值）次迭代进而得到最终的停用字集合 $SCset\_Final$。得到停用字集合后，利用字向量计算语料中的每个字 $token$ 与停用字集合 $SCset\_Final$ 中的字的相似度，取其中的最大值作为该字成为停用字的可能性 $stopProb$。考虑到 CRFs 模型学习的特征是离散的，我们对获得的数值进行了离散化处理，处理方法如公式（4）所示。

$$simCRF=\begin{cases}-1, stopProb<0\\-2, stopProb=0\\\lceil stopProb*10\rceil, stopProb>0\end{cases} \quad (4)$$

3 λ-主动学习方法

3.1 基于主动学习的分词算法

在主动学习过程中，首先使用初始分词器对未标注语料进行标注。然后采用一定策略计算每个样例整体的标注价值，按照样例的标注价值从高到低进行排序。根据排序结果，选择前 N 个样例进行人工修正。将修正后的样例加入到训练语料中，重新训练分词器（CRFs 模型）。利用新得到的 CRFs 模型对未选中的语料重新标注，根据标注结果进行新一轮的迭代。经过多次迭代直到达到终止条件，获得最终的分词器。基于主动学习的中文分词算法如下：

算法 1：基于主动学习的分词算法

输入：初始训练语料 L_0，未标注样本集 U_0，初始分词器（CRFs 模型），每次选择的样例个数 M，终止条件 D

输出：最终的训练语料和分词器

L1. **while True do**
L2. 用 L_i 训练 CRFs 模型，得到 $Model_i$
L3. **if** L_i 中的样本数达到终止条件 D **do**
L4. break
L5. **end if**
L6. 用 $Model_i$ 对 U_i 中的样本进行序列标注
L7. **for** S **in** U_i do
L8. 计算样例 S 中所有字的边界差异性和标注结果的不确定性
L9. 计算样例 S 的所有字的标注价值 $\varphi(c)$
L10. 根据样例 S 中字的标注价值计算样例 S 整体的标注价值 $\varphi(S)$
L11. **end for**
L12. 根据 $\varphi(S)$ 对 U_i 中所有样本进行排序
L13. 选择标注价值最高的 M 个样本，放入样本池 P_i 中
L14. 由人工对 P_i 中的 M 个样本中的不确定性较高的字符的标注结果进行修正
L15. 从 U_i 中删除 P_i 中包含的样本，获得 U_{i+1}
L16. 将修正后的 M 个样本加入到标注样本集 L_i 中，获得 L_{i+1}
L17.**end while**

3.2 字边界的差异性

由于微博语料中存在大量局部相同而整体不同的句子（如样例 a 和样例 b 所示），在利用主动学习方法选取语料的过程中，如果对样例的差异性要求过于宽松，则会选入很多重复的语料，增加人工标注的工作量；如果对样例的差异性要求过于严苛，则会漏选有价值的语料。

确保样例的多样性是所有主动学习方法选择语料时需要解决的问题，但现有的主动学习方法对于局部相同但整体不同的语料的选择过于严苛，选取语料时无法合理控制此类语料的数量。为了更加有效地控制样例的差异性，本文在主动学习过程中引入参数 λ。由于字边界的差异性与其上下文字符密切相关，因此本文利用字边界的上下文一元字符来考察字边界的差异性，并构造了公式（5），根据该方法计算字边界的差异性。

$$F(c)=-\left(d^t(c)/\lambda\right)^3 \qquad (5)$$

其中，c 表示当前被考察的字符，$d^t(c)$ 表示 c 的上下文 t 出现的频度，初始时为 0，当 c 第 $i+1$ 次遇到上下文 t 时，$d_{i+1}^t(c)=d_i^t(c)+1$。该方法通过参数 λ 对抽取的重复样例的个数进行控制。当 $d^t(c)<\lambda$ 时，其差异性取值为(-1, 0)，对字的标注价值影响较小；而随着 $d^t(c)$ 的增加，$F(c)$ 下降的速度会增大，字边界的差异性对字的标注价值的影响越来越大。当 λ 取不同值时，字边界的差异性的变化曲线如下图所示，图中横坐标为 $d^t(c)$，纵坐标为 $F(c)$。

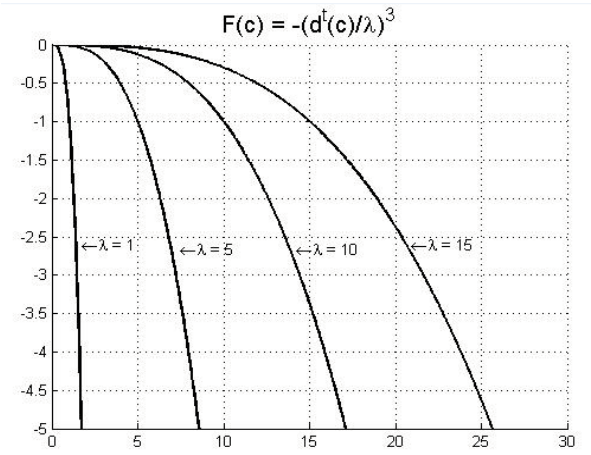


图 1 参数 λ 对字上下文差异性的影响

3.3 字标注结果的不确定性

在主动学习迭代的过程中，除了确保语料的多样性，还要保证所选取的样例，能够对重新训练模型提供有价值的信息，即，所选样例的标注结果应具有不确定性。为此，本文提出新的计算公式，以衡量字标注结果的不确定性：

$$H_{category}(c)=-\sum_{i=N,Y}\left(P_i(c)+\gamma\right)\log\left(P_i(c)+\gamma\right) \qquad (6)$$

其中，$P_N(c)=P_B(c)+P_M(c)$；$P_Y(c)=P_E(c)+P_S(c)$，$P_B(c)$ 表示根据 CRFs 模型的标注结果，字符 c 被标注为 B 的后验概率，$P_M(c)$、$P_E(c)$、$P_S(c)$ 的意义与 $P_B(c)$ 的意义类似；c 表示当前被考察的字符；γ 是为了解决计算信息熵时产生的光滑性问题。$H_{category}(c)$ 越大，该字符的不确定性越高，对重新训练模型越有价值。

上述方法将 CRFs 的标注集划分为两类（即 N 和 Y），可能在一定程度上模糊了标注结果的分布信息。为了考察将 4 个标签合并为 2 类后是否会掩盖有用信息，本文还提出一种基于 4 个标签的字标注结果不确定性计算方法，即利用信息熵全面考虑 CRFs 标注结果中 4 种标签（B、M、E、S）的分布情况，根据标签的边缘概率计算字标注结果的不确定性：

$$H_{label}(c)=-\sum_{i=B,E,M,S}\left(P_i(c)+\gamma\right)\log\left(P_i(c)+\gamma\right) \qquad (7)$$

其中，各个符号的意义与公式（6）相同。

3.4 样例整体标注价值的评价方法

在主动学习过程中，综合考虑样例中所包含的字的标注价值，进而获得每个样例整体的标注价值。判断字标注价值时，主要从两个方面进行衡量：字标注结果的不确定性和字边界的差异性，计算方法如下：

$$\varphi_c(c)=\alpha H(c)+\beta F(c) \qquad (8)$$

其中，$H(c)$表示字标注结果的不确定性，本文采用两种计算方法，分别为$H_{category}(c)$和$H_{label}(c)$；$F(c)$表示字边界的差异性；α和β表示字标注结果的不确定性和字边界的差异性在计算字的标注价值时的权重，$\alpha+\beta$=1。

然而，中文分词任务的训练语料是基于句子的，因此，在选取需要进行人工修正的样例时，样例整体的标注价值比字的标注价值更具有参考性。本文根据样例中包含的所有字的标注价值来确定样例整体的标注价值。本文提出三种策略来计算样例S整体的标注价值。

策略一：计算样例中所有字的标注价值的平均值作为样例S的标注价值$\varphi(S)$。

策略二：选择样例中所有字的标注价值中的最大值作为样例S的标注价值$\varphi(S)$。

策略三：将样例中包含的所有字的标注价值的平均值和最大值相加，作为样例S的标注价值$\varphi(S)$。

本文采用上述三种策略挑选样例进行人工修正和迭代训练，在实验分析部分会给出采用上述三种方法的实验效果的对比分析。

4 实验设计及结果分析

4.1 实验语料

实验使用的训练语料和测试语料是 NLPCC 2015 年公布的面向微博的中文分词评测任务的训练语料和测试语料[9]（http://nlp.fudan.edu.cn/nlpcc2015），详细信息如表 1 所示。此外，我们还使用了未标注的背景语料（包括 300,000 条微博，约 20,000,000 个字），用于训练基于半监督方法的初始分词器（http://www.sina.com.cn/）。

表 1 训练语料和测试语料的统计信息

Dataset	Sentences	Words	Characters
Training	10,000	215,027	347,984
Test	5,000	106,327	171,652
Total	15,000	322,410	520,555

4.2 评价方法

实验采用精确率、召回率和 F 值对分词结果进行评价，三个评价指标的计算方法为：

精确率=分词结果中正确的词语个数/分词结果中的总词数；

召回率=分词结果中正确的词语个数/标准答案中词语的总数；

F 值=（2*召回率*精确率）/（召回率+精确率）。

4.3 实验结果

4.3.1 初始分词器

本文用于主动学习方法的初始分词器是基于字的 CRFs 序列标注模型。由于大规模未标注语料中存在很多共现信息和字边界信息，我们除了提取窗口为 5 的上下文和是否是标点作为 CRFs 模型的特征外，还利用大量未标注语料获得连续二字串的 PMI 值和当前字成为停用字的可能性，作为训练 CRFs 模型的特征。实验结果如表 2 所示：

Baseline：是利用基础特征训练 CRFs 模型得到的分词器，基础特征包括窗口为 5 的上下文特征和是否是标点的特征；

Baseline$_{PMI}$：表示在训练 CRFs 模型时，除了使用 Baseline 的基础特征外，还使用了语料中连续二字串的 PMI 值作为特征；

Baseline$_{PMI+CE}$：表示在训练 CRFs 模型时，除了使用 Baseline 的基础特征和 PMI 统计量外，还使用了利用当前字的字向量和停用字集合计算得到的停用字相似度 *stopProb* 作为特征。

表 2 初始分词器的分词结果

	精确率	召回率	F 值
Baseline	93.46	92.99	93.22
$Baseline_{PMI}$	93.71	93.35	93.53
$Baseline_{PMI+CE}$	94.02	93.71	**93.87**

由表中数据可知，从大规模未标注语料中提取的 PMI 有效提高了初始分词器的分词效果；利用字向量和停用字集合计算得到的停用字相似度这个特征的加入又进一步提高了分词结果。为了获得最优的分词结果，最终，我们选择 $Baseline_{PMI+CE}$ 作为 λ-主动学习的初始分词器。

4.3.2 λ-主动学习方法分词结果

为了更加有效地控制样例的差异性，本文在主动学习过程中引入参数 λ，通过参数 λ 对字边界的差异性进行衡量，从而限制选取的具有相同上下文的样例的数量。实验过程中我们对参数 λ 的取值进行了多次调整， λ 取不同值时对主动学习方法的影响如图 2 所示，图中的横坐标为主动学习的迭代次数，纵坐标为分词结果的 F 值。

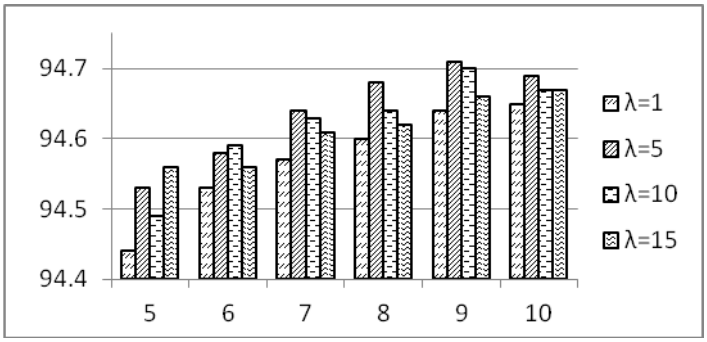


图 2 参数 λ 对主动学习分词结果的影响

数据表明，当 $\lambda=5$ 时，分词结果最佳。该结果意味着，每次主动学习迭代的过程中，对于局部相同但整体不同的样例的选取数量控制在 5 时，模型重新训练后的效果最佳，分词结果提升效果最为显著，此时，如果字边界相同的样例的个数超过 5，则其字边界的差异性会显著降低（低于-1），并且随着字边界相同的样例个数的增加，其差异性的降低幅度会越来越大，对样例整体的标注价值的影响也越来越大，因此，被过滤掉的几率也随之增大。最终，我们将 λ-主动学习方法的参数 λ 的值设定为 5。

在主动学习选择样例的过程中，除了引入参数 λ 外，还提出了三种不同的策略对样例整体的标注价值进行衡量。为了验证提出的选择策略是否有效，我们进行了多组对比实验。在主动学习过程中，随着迭代次数的增加，训练语料的规模逐渐扩大，分词结果也会随之变化。为了保证实验对比的公平性，每次迭代过程中，所有策略所选取的语料的数量是相同的（都是 500 条微博）。图 3 展示了基于不同选择策略的主动学习方法的分词结果的 F 值随着迭代次数的增加而发生的变化。实验中，计算字的标注价值时，字标注结果的不确定性的权重 α 和字边界的差异性的权重 β 分别取经验值 0.5。

Avg：代表根据**策略一**对每个样例整体的标注价值进行评价，然后进行排序，根据排序结果选择需要进行人工修正的语料；

Max：代表根据**策略二**对每个样例整体的标注价值进行评价，然后进行排序和选择；

AvgMax：代表根据**策略三**对每个样例整体的标注价值进行衡量，然后进行排序和选择；

WBA：是对目前用于中文分词任务中效果最佳的主动学习方法 WBA 的重现[13]；

Random：是采用随机选择的方法在每次迭代过程中随机选择一定数量的语料进行人工修正。

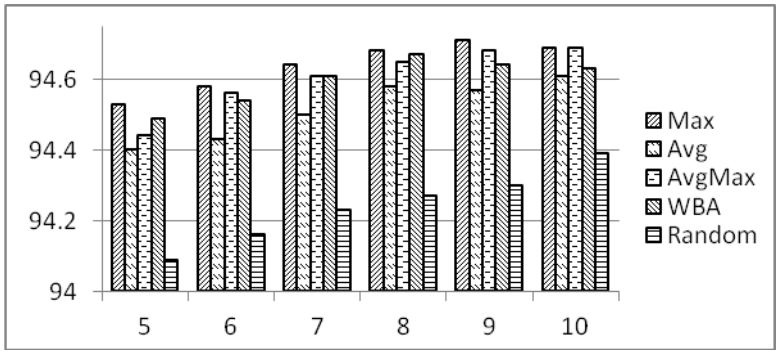


图 3 不同选择策略的分词结果

图 3 数据表明，每次迭代过程中，Max 方法的 F 值都显著高于其他选择策略，说明 Max 方法在选择语料时，能有效选取更具标注价值的语料，为重新训练分词器更有帮助；此外，随着迭代次数的增加，所有方法的分词效果处于上升趋势，说明语料的增加有助于提高分词效果；但是随着语料的不断增加，分词效果的上升速度逐渐趋于平稳，甚至有所下降，该现象的一个可能的原因是，当语料增加到一定规模时，模型出现过拟合现象；图中数据显示，本文提出的三种方法和 WBA 方法都明显优于随机选择方法，进一步说明了主动学习方法能够有效选择具有标注价值的语料。

表 3 不同方法的分词结果的最佳 F 值

	迭代次数	精确率	召回率	F 值
$\text{Baseline}_{\text{PMI+CE}}$	--	94.02	93.71	93.87
Max	9	**95.06**	**94.36**	**94.71**
Avg	10	94.97	94.25	94.61
AvgMax	10	95.05	94.33	94.69
WBA	8	95.01	94.34	94.67
Random	10	94.74	94.03	94.39

表 3 为迭代过程中每种方法所达到的最高 F 值。从表中数据可以看出，基于 Max 方法的分词结果的 F 值高于其他方法，且该方法与初始分词器的结果相比，有显著提高（提高了 0.84%），比文献[13]提出的 WBA 方法也有所提高，再次说明本文提出的 Max 方法在主动学习过程中能够有效选择具有较高标注价值的样例。

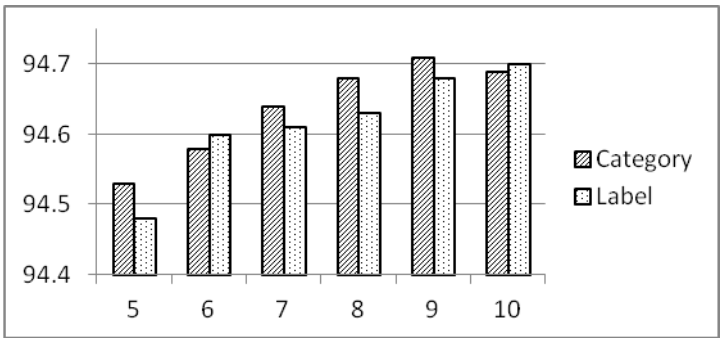


图 4 字的标注结果的不确定性的衡量方法

此外，我们还对本文提出的两种衡量字的不确定性的方法 $H_{category}(c)$ 和 $H_{label}(c)$ 进行了比较，实验结果如图 4 所示，两种方法的实验结果比较接近，但是 $H_{category}(c)$ 方法更加稳定，且该方法取得的最佳的 F 值比 $H_{label}(c)$ 方法更高。该结果说明将四字标注集划分为两类后，并不会模糊四种标签之间的信息，反而为选取有效样例提供了帮助。因此，本文最

终的系统采用 $H_{category}(c)$ 方法进行字的不确定性的判断。

5 总结和展望

本文根据微博语料中存在大量局部相同而整体不同的句子的特点，提出了 λ-主动学习方法。该方法的初始分词器采用基于半监督方法的 CRFs 模型，在迭代学习的过程中，通过参数 λ 对字边界的差异性进行衡量和控制，结合字的标注结果的不确定性获得字的标注价值，进而根据字的标注价值采用 Max、Avg、AvgMax 三种策略评价每个样例整体的标注价值。实验结果表明，基于 Max 的衡量方法优于其它两种方法，并且该方法优于目前该领域最佳的主动学习方法 WBA，说明本文提出的主动学习方法所选取的语料具有更高的标注价值。

本文的主要贡献如下：

1) 针对微博语料的特点，提出了一种新的 λ-主动学习方法，在迭代过程中，引入参数 λ 来控制所选的局部相同但整体不同的样例的个数，确保所选样例的多样性；
2) 根据样例中字的标注价值，采用多种策略考察样例整体的标注价值，实验结果显示我们提出的基于 Max 的主动学习方法可以有效选择具有较高标注价值的语料；
3) 用于主动学习的初始分词器除了使用当前字的上下文作为特征外，还利用字向量自动计算当前字成为停用字的可能性作为 CRFs 模型的特征。

在未来的研究中，我们希望能利用本文所提方法进行训练语料的自动获取，尽量减少或避免人工标注工作，并利用大规模自动获取的训练语料提高基于深度学习模型的面向微博语料的中文分词效果。

参考文献

[1] Nguyen T H, Shirai K. Topic modeling based sentiment analysis on social media for stock market prediction[C]//Proc of the 53rd Annual Meeting of the ACL. New York: ACL, 2015: 1354-1364

[2] Liu Xiaohua, Zhou Ming, Wei Furu, et al. Joint inference of named entity recognition and normalization for tweets[C]//Proc of the 50th Annual Meeting of the ACL. New York: ACL, 2012: 526-535

[3] Li Cheng, Liu Yang. Improving Named Entity Recognition in Tweets via Detecting Non-Standard Words[C]//Proc of the 53rd Annural Meeting of the ACL. New York: ACL, 2015: 929-938

[4] Dong Guozhong, Li Ruiguang, Yang Wu, et al. Microblog burst keywords detection based on social trust and dynamics model[J]. Chinese Journal of Electronics, 2014, 23(4): 695-700

[5] Tseng H, Chang P, Andrew G, et al. A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005[C]//Proc of the fourth SIGHAN workshop on Chinese Language Processing. New York: ACL, 2005: 168-171

[6] Zhang Huaping, Yu Hongkui, Xiong Deyi, et al. HHMM-based Chinese lexical analyzer ICTCLAS[C]//Sighan Workshop on Chinese Language Processing. New York: ACL, 2003: 184-187

[7] 黄德根, 焦世斗, 周惠巍. 基于子词的双层CRFs中文分词[J]. 计算机研究与发展, 2010, 47(5):962-968

[8] Qiu Xipeng, Qian Peng, Yin Liusong, et al. Overview of the NLPCC 2015 Shared Task: Chinese Word Segmentation and POS Tagging for Micro-blog Texts[C]//Natural Language Processing and Chinese Computing. Berlin: Springer, 2015: 541-549

[9] Qiu Xipeng, Qian Peng, Shi Zhan. Overview of the NLPCC-ICCPOL 2016 shared task: Chinese word segmentation for micro-blog texts[C]// Natural Language Understanding and Intelligent Applications. Berlin: Springer, 2016: 901-906

[10] Tang Min, Luo Xiaoqiang, and Roukos S. Active learning for statistical natural language parsing[C]//Proc of the 40th Annual Meeting on ACL. New York: ACL, 2002: 120-127

[11] Li Shoushan, Xue Yunxia, Wang Zhongqing, et al. Active Learning for Cross-domain Sentiment Classification[C]//Proc of the Twenty-Third IJCAI. Palo Alto: AAAI, 2013: 2127-2133

[12] Chen Yukun, Lasko TA, Mei Qiaozhu, et al. A study of active learning methods for named entity recognition in clinical text. Journal of biomedical informatics[J]. 2015, 58(C): 11-18

[13] Li Shoushan, Zhou Guodong, Huang Chu-Ren. Active learning for Chinese word segmentation[C]//Proc of COLING 2012: Posters. New York: ACM, 2012: 683-692

[14] 梁喜涛, 顾磊. 基于最近邻的主动学习分词方法[J]. 计算机科学 42.6 (2015): 228-232

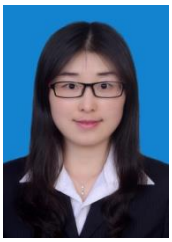
[15] 冯冲, 陈肇雄, 黄河燕, 等等. 基于Multigram语言模型的主动学习中文分词[J]. 中文信息学报, 2006, 20(1): 52-60

[16] Sun Weiwei, and Xu Jia. Enhancing Chinese word segmentation using unlabeled data[C]//Proc of the Conf on EMNLP of the ACL. New York: ACL, 2011: 970-979

[17] Zhao Hai, and Kit Chunyu. Exploiting unlabeled text with different unsupervised segmentation criteria for Chinese word segmentation[J]. Research in Computing Science 33, 2008: 93-104

[18] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations[C]//Proc of NAACL-HLT 2013. New York: ACL, 2013: 746–751

[19] Chen Xinxiong, Xu Lei, Liu Zhiyuan, et al. Joint learning of character and word embeddings[C]//Proc of IJCAI. Palo Alto: AAAI, 2015: 1236-1242



张婧（1987——），博士研究生，主要研究领域为自然语言处理。

E-mail：zhangjingqf@mail.dlut.edu.cn

黄德根（1965——），通讯作者，教授，博士生导师，主要研究领域为自然语言处理、机器翻译、人工智能等。

E-mail:huangdg@dlut.edu.cn

黄锴宇（1992——）硕士研究生，主要研究领域是自然语言处理。

E-mail：huangkaiyucs@foxmail.com