北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2016.002

中文分词中基于主动学习的领域自适应方法

许华婷　张玉洁†　杨晓晖　单华　徐金安　陈钰枫

北京交通大学计算机与信息技术学院，北京　100044; †通信作者, E-mail: yjzhang@bjtu.edu.cn

摘要　在新闻领域标注语料上训练的中文分词系统在跨领域时性能会有明显下降。针对目标领域的大规模标注语料难以获取的问题，提出 Active learning 算法与 n-gram 统计特征相结合的领域自适应方法。该方法通过对目标领域文本与已有标注语料的差异进行统计分析，选择含有最多未标记过的语言现象的小规模语料优先进行人工标注，然后再结合大规模文本中的 n-gram 统计特征训练目标领域的分词系统。采用 CRF 训练模型，在 100 万句的科技文献领域上，验证了所提方法的有效性，评测数据为人工标注的 300 句科技文献语料。实验结果显示，在科技文献测试语料上，基于 Active learning 训练的分词系统在各项评测指标上均有提高。

关键词　中文分词；领域自适应；主动学习
中图分类号　TP391

Domain Adaptive Method Based on Active Learning in Chinese Word Segmentation

XU Huating, ZHANG Yujie†, YANG Xiaohui, SHAN Hua, XU Jin'an, CHEN Yufeng

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044;
† Corresponding author, E-mail: yjzhang@bjtu.edu.cn

Abstract　Chinese word segmentation systems, trained on annotated corpus of newspaper, often obviously decrease in performance when faced with a new domain. Since there is no large scale annotated corpus on the target domain, statistics based methods could not work well. The authors attack domain adaptation of Chinese word segmentation by combining active learning with the statistical features of n-gram. The idea is to select such a small amount of data for annotation that the gap from the target domain to the News will be overcome. The word segmentation model is trained again by the corpus added with newly annotated data and the statistical features of n-gram from the raw corpus. The authors use the CRF model for training and a raw corpus of one million sentences on patent description to verify the proposed approach. For test data, 300 sentences are randomly selected and manually annotated. The experimental results show that the performances of the Chinese word segmentation system based on the approach are improved on each evaluation metrics.

Key words　Chinese word segmentation; domain adaptation; active learning

传统的中文分词方法是基于词典的方法，主要有正向最大匹配算法、逆向最大匹配算法、N-最短路径分词算法等。随着标注语料库的建立和统计机器学习的发展，基于统计的中文分词方法成为主流方法。常用的统计机器方法包括：基于隐马尔可夫模型(Hidden Markov Model, HMM)中文分词方法[1]、基于最大熵模型(Maximum Entropy, ME)的中文分词方法[2]和基于条件随机场模型(Conditional Random Fields, CRF)的中文分词方法[3–4]等。在基于字符序列标注的分词方法上，又有研究人员提出

国家国际科技合作专项(2014DFA11350)和国家自然科学基金(61370130)资助
收稿日期: 2015–06–19; 修回日期: 2015–08–15; 网络出版时间: 2015-09-29 15:45:10

结合单词信息的混合模型和解决高复杂度训练的方法[5–6]。

当中文分词任务的领域发生变化时，未登录词的比例会上升，导致中文分词系统的精度大幅下降。为了解决分词系统领域自适应的问题，近些年，研究者提出很多方法，主要有数据加权算法和半监督学习算法。张梅山等[7]采用领域词典与统计方法相结合的方法，分词系统针对不同领域的文本进行分词时，通过加载相关领域的词典辅助分词系统进行分词。但是，以上这些方法都受到标注语料或特定词典的限制，相关资源不易获得。针对这一问题，有研究者提出可以通过从大规模生语料中抽取 n-gram 统计特征的方法，改善由于领域变化导致的分词性能下降问题[8]。但这种方法仅利用计算机的统计方法，由于未考虑到领域专有词也具有一定的中文构词规律，造成在一些专有词上分词不准确，影响了分词精度。为了进一步提高领域变化后中文分词系统的分词精度，有研究者提出在利用 n-gram 统计特征的基础上，增加平行语料语言知识(一般来讲是中英文平行语料)，通过英文单词的边界辅助对应的中文字符串划定词语界线[9]。但是，对于大多数中文语料来讲，并不是都有对应的英文译文，所以这种方法不适合推广。但它为中文分词提供了一个新的思路，可以利用不同资源的叠加来提高分词精度。在对比前人研究结果的基础上，我们考虑利用大规模生语料中的统计特征与少量人工标注相结合的方法，以提高中文分词领域自适应能力。

本文工作围绕中文分词领域自适应的问题，针对大规模人工分词标注语料难以获取的现状，提出基于 Active learning 的中文分词领域自适应方法。本文方法通过对目标领域文本与已有标注语料之间差异性的统计分析，选出小规模的包含目标领域中特有语言现象的语句，进行人工标注；然后结合大规模生语料中的 n-gram 统计特征，调整已有分词模型的领域适应性，从而达到通过标注少量语料，改善分词精度的目的。

1 Active learning 算法介绍

Active learning 算法由耶鲁大学 Angluin 教授提出[10]。它选择部分未标记样例进行标记，然后把它们放入之前已有的标记样例集合，重新训练分类器，利用分类器再次选择未标记样例。通过有选择地扩大有标记样例集合和循环训练，使得分类器逐步获得更强的泛化能力。与以往的算法相比，它具有模拟人的学习过程的特点，因此受到广泛关注，近年来被大量地应用于信息检索和文本分类等自然语言处理领域，成为机器学习领域中最重要的方向之一。

为了更加直观地展示 Active learning 算法的有效性和它对分类器训练精度的提高程度，用一个对二维空间中的点进行分类的问题[11]为例来介绍。假设有一个布满红绿两种共 400 个点的平面，欲找到红绿两种点的分界线。已知红绿两种点在坐标 $x=0$ 附近产生分界。分别利用被动学习和主动学习两种方法，各自选择和标记 30 个点，并利用标记的点找到分界线。用常规的被动学习方法，随机选择并标记点，通常标记的点比较分散，只有很少的点分布在 $x=0$ 附近，很难找到正确的分界线。用这种方法分类的精度较低，据统计结果显示，正 确率只有 70%。Active learning 方法是通过选择运算，最终选择红绿两种点混合分布比较紧密的位置，也就是位于 $x=0$ 附近的点进行标记，这就为找到正确的分界线提供了有效的信息。利用这种方法训练得到的分类器精度较高，正确率可达到 90%。这个例子充分说明，Active learning 选择的样本点比盲目选择的样本点更有利于高精度分类器的训练，在同等的标注代价下能够得到更多的区分信息，有利于提高分类模型的精确度。

2 基于 Active learning 的中文分词领域自适应方法

假设有一个在分词标注语料(原领域)上训练得到的中文分词模型，要对一个不同领域(目标领域)的文本进行分词处理，为此需要将中文分词模型从原领域调整到目标领域。

为了更好地处理领域内专有名词及特殊句式的分词问题，本文提出基于 Active learning 的中文分词领域自适应方法，借助 Active learning 算法，选择目标领域中最具有领域特点的小规模语句进行人工标注，再与该领域大规模生语料 n-gram 统计特征相融合，从而实现分词精度更高的领域自适应的中文分词系统。系统的总体框图如图 1 所示，虚线框部分表示实现领域自适应的核心部分。

相对于原领域的词语分布，目标领域的词语分布中出现的汉字以及构词模式会有很大差异。如果将含有差异多的句子筛选出来进行人工标注，将为

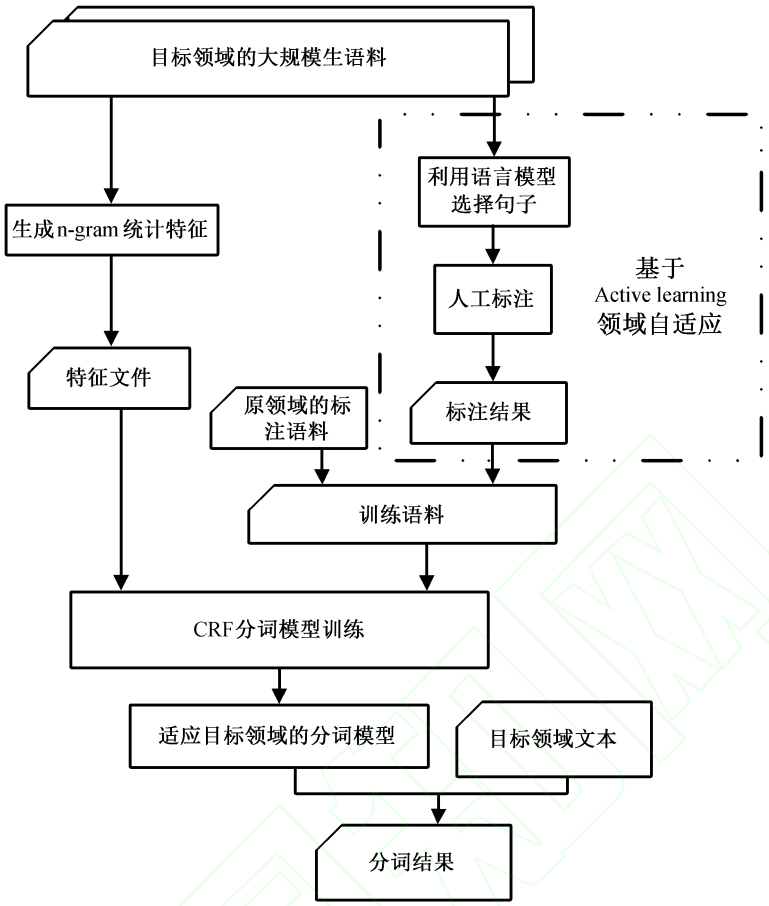

图 1　基于 Active learning 中文分词领域自适应的整体框架

Fig. 1　Framework of domain adaption of Chinese word segmentation based on active learning

分词模型的重新训练优先提供目标领域特有的训练语料，使得分词模型可以快速获取目标领域特有的分词知识，从而有效提高在目标领域的分词精度。因此，如何筛选出含有差异多的句子成为关键。

为了从大规模目标领域的生语料中抽取在构词规律和词汇分布上具有目标领域特征的语句进行人工标注，本文采用基于 n-gram 加权统计的方法来计算每个句子相对于原领域在 n-gram 上的分布差异性，具体计算如式(1)所示。

$$\phi^N(s)=\sum_{n=1}^{N}\frac{w_n}{X_s^n}\sum_{x\in X_s^n}\log\frac{P(x\,|\,U,n)}{P(x\,|\,L,n)},\qquad(1)$$

L 表示原领域的语句集合，U 表示目标领域的语句集合。X_s^n 表示句子 S 中的 n-gram 集合。$P(x|D, n)$ 表示语句集合 D 的 n-gram 集合中 x 的概率。w_n 表示调整不同长度的 n-gram 的重要性的权重。本文使用 N=4 的语言模型。$\phi^N(S)$是对句子 S 的评分。评分越高，说明该语句中含有更多的未被原领域覆盖的字符串，人工标注后，可以为分词模型的重新训练提供目标领域特有的训练语料，使得分词模型可以获取目标领域特有的分词知识；相反地，评分越低，说明该语句的词汇分布与原领域接近，对这样语句进行人工标注，不会对构建目标领域分词模型提供更多的新增的分词知识。

利用上述方法对目标领域生语料的所有语句进行评分计算后，按评分对语句进行排序。根据在人工标注上的投入预算或者需要达到的精度要求，选择小规模的高位语句，按照目标领域分词标注标准进行人工标注。标注好的语料与原领域的标注语料构成新的训练语料，然后采用 CRF 模型在新的语料进行训练，构建适应目标领域的分词模型。

3　领域自适应方法在科技领域的应用

以科技领域为例，利用上面提出的方法，对构

建科技领域上的中文分词自适应系统做详细介绍。

3.1 科技领域分词系统的建立

已有的中文分词模型是在宾州中文树库(Penn Chinese Treebank, CTB)上训练获得的，原领域为新闻领域。目标领域的语料是 NTCIR-10 中的 100 万句中文科技文献语句。我们从中随机选出一部分语句作为测试数据。

为了获取科技领域的分词特征，一方面利用语言模型对科技领域生语料的所有语句进行评分排序，筛选出一小部分得分高的语句，依据科技领域分词标注标准进行人工分词。科技领域分词标注标准的建立将在第 3.2 节详细介绍。标注结果将加入新闻领域的标注语料形成新的训练数据。另一方面，从科技领域的大规模生语料中抽取 n-gram 统计特征生成特征文件。然后采用 CRF 模型在这两方面生成的训练数据和特征文件上进行训练，得到适用于科技领域的中文分词模型。基于 Active learning 实现中文分词在科技领域上适应的总体框图如图 2 所示。

3.2 科技领域分词标注标准的制定

“词是什么(词的抽象定义)”，“什么是词(词的具体界定)”，这两个基本问题至今没有一个公认的、具有权威性的定义。同时，对于中文“词”的认定，普通人的标准和语言学家的标准在认定上也有比较大的差异。调查结果表明，在母语是中文的被测试者之间，对中文文本中出现的词语的认同率大约只有 70%。研究人员邀请 258 名文理科大学生对同一篇约 300 字的短文进行手工分词，结果表明，在其中的 45 个中文双音节和三音节结构的词语上，分词的结果与专家给出的标准分词结果相同的人甚少[12]。1992 年国家标准局颁布了作为国家标准的《信息处理用现代汉语分词规范》[13]，其中，大部分是通过举例和定性描述来定义分词规范的。例如，规范 4.2 规定：“二字或三字词，以及结合紧密、使用稳定的二字或三字词组，一律为分词单位。”但是，没有明确定义如何判断“紧密”以及如何判断“稳定”，这样的形容在很多规定中都有出现。这样规定的判断准则极易受到主观因素的影响，造成具体判断非常困难，因此，建立一个易于操作、能够保证标注者之间较高一致性的分词标注标准，需要解决很多问题。

目前具有广泛影响的中文分词标注标准中，有 CTB 的中文分词规范[14]和北京大学的中文分词规范[15]。前者是针对新闻领域的分词规范，后者是面

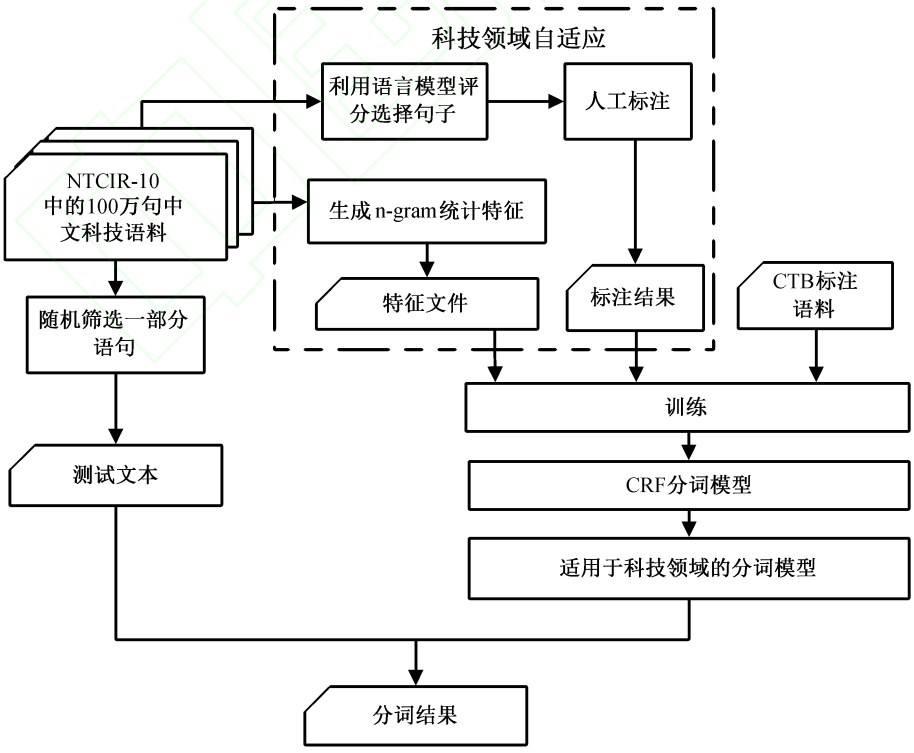

图 2　基于 Active learning 中文分词方法在科技领域上的应用框架

Fig. 2　Framework of the Chinese word segmentation adapted to the patent description text

对一般领域的分词规范，它们对科技领域的专业词汇没有十分明确详细的标注标准。科技文本中的词语和构词的汉字与新闻或一般领域有很大不同。考虑到这一特点，本文制定了面向科技领域的中文分词标注标准，其中，一般词汇的分词标准与 CTB 中文分词规范保持一致。针对专业词语，我们分析了 CTB 中文分词规范中各种类别的汉字构成词语的模式，制定相应的标注标准。以化学、药物的中文文本为例，对增添的分词标注规则举例说明，规则在表 1 中列出。其中，“+”表示“任何非空汉字字符串”，“*”表示“任何汉字字符串，包括空串”，“|”表示“或者”，“\”表示“词语的边界”。下面对化学类词汇的标注规则进行解释:

1) 当遇到“+基*酸*酯”时，规定切分为“+基*酸*酯”；当遇到“+酸*酯”时，规定切分为“+酸*酯”;

2) 当遇到“+菌霉”时，如果“+”中是形容词的时候，规定切分成“+\菌\霉”，否则切分为“+菌\霉”;

3) 当遇到“+剂”时，如果“+”只是一个汉字，且“+剂”是出现在新闻领域的词语，规定“+剂”当作一个词语来切分；如果“+”是两个或以上的汉字，那么规定切分为“+\剂”。

3.3　科技领域 n-gram 统计特征

n-gram 是指文本中连续出现的 *n* 个连续汉字组成的串。从形式上看，词是稳定的字串，即组成词的字之间凝固度较高。当训练语料足够丰富时，词的出现次数一般高于不成词的 n 元字串。从直观的角度考虑，词一般是高频 n 元字串，但是高频 n 元字串并非一定是词。例如: “巧克力”在未标注的语料中出现 *m* 次，那么“巧克”出现的次数一定不小于 m，但“巧克”并不是一个词。一个完整的词单元应能适应多样的上下文，如“吃巧克力”、“黑/白巧克力”、“巧克力糖”、“精致的巧克力键盘”等等，“巧克力”作为词单元有丰富的上下文，而“巧克”在多数情况下与“力”搭配，它的下文环境单一。可以直观地认为成词(包括未登录词)的字串应当同时具备出现次数多和上下文环境丰富的特点。

在不同领域的语料中，字与字连在一起构成词的情况是不一样的。我们希望通过统计大规模生语料中 n 元字串的一些特征供统计模型学习，以达到分词系统领域自适应的目的。基于词单元在未标注语料中所体现的特性，所采用的统计特征包括 n-gram 频度特征和 n-gram AV 特征。

n-gram 频度特征　n-gram 的频度值即 n 元字串在语料中的出现次数。本文统计了目标领域生语料中所有 2 元、3 元、4 元和 5 元字串的频度，其中频度小于 5 的字串被过滤。由于 n 元字串的频度值取值范围从 5 到几千甚至几万，为了避免数据稀

表 1　科技领域人工分词标注标准举例

Table 1　Examples of artificial tagging standards in field of science and technology

化学名词	药物类名词	+于	+剂	最|更+	所+
丙烯酸	抗\微生物\剂	约\等于	溶剂	最佳	所\谓
甲基\丙烯酸	抗\寄生虫\剂	基于	抑制\剂	最有用	所\述
聚\甲基\丙烯酸\甲酯	抗\病毒\剂	由于	药物\制剂	最上\部分	示例\所\示
高\密度\聚乙烯	抗\糖尿病\剂	不\限于	复合\催化\剂	更适合于	
二酰\亚胺\键	抗\止血\剂	取决于	活性\剂		
枯草杆菌\霉	神经肌肉\阻断\药	滴注于	药剂		
阳\离子		有利于	洗涤\剂		
碳\原子		用于			
否定表示	其他	其他	其他	其他	其他
并\不	本文	给药\途径	远程控制	抑郁\症	高速空气射流
而\不\是	制得	温度\传感器	电阻\传感器	侧视\图	重\放
反\之\亦然	化学式	光学\扫描\设备	熔融\聚合物\树脂	目的\在于	化学式
	各\个\方面	信息\层	受测\对象	示意\图	
	各式各样	电光\调制器	聚烯烃\聚合物	组合物	
	示例\性				

疏影响 CRF 学习的效果，本文采用离散化的方法将 n 元字串的频度归为 3 类：高频(H)、中频(M)、低频(L)。n 元字串按照频度值从高到低排序，前 5%的 n 元字串归为高频，表示为 H；排名低于 5%但高于 20%的 n 元字串归为中频，表示为 M，最后 80%的 n 元字串的频度值归为低频，用 L 表示。

字串的特征只有转化为字的特征才能供 CRF 模型学习。在给定句子中的当前汉字产生 n-gram 频度特征时，依次考察句子中包含当前汉字的所有候选词。该字在词中的位置信息在前，频度信息在后，用“-”把它们连接起来。最后，按照当前汉字所处候选词中的位置从前到后的顺序(即 B, B_1, B_2, M, E 的顺序)把前面记录的信息用“|”连接起来作为当前汉字最终的 n-gram 频度特征。

n-gram AV 特征 AV (Accessor Variety)是从生语料中提取词语判断一个字串是否是词的统计标准。与 n-gram 频度值不同的是，n-gram AV 值对频度值进行了筛选。AV 的主要思想是：若一个字串在多种语境下出现，那么该字串成为词的可能性就高。AV 的定义如下：

$$AV(s) = \min\{L_{av}(s), R_{av}(s)\}, \quad (2)$$

$L_{av}(s)$和 $R_{av}(s)$分别表示字串 s 的不同前驱和后继的数量。

与 n-gram 频度特征的使用类似，首先统计目标领域生语料中的所有 2 元、3 元、4 元和 5 元字串的 AV 值，过滤掉 AV 值小于 5 的 n 元字串；然后采用与 n-gram 频度值相同的分类标准，将 n 元字串按照 AV 值分成 3 个频档：H, M 和 L；最后将字串的特征转化为字的特征供 CRF 训练和解码。

4 实验评测与分析

为了评测在科技领域上适应的中文分词模型的性能，验证本文方法的有效性，我们在 NTCIR-10 的英中科技专利数据上设计了一组实验。通过对大规模中文语料的分词处理，从中文分词精度方面进行评测，并分析人工标注数据规模对基于 Active learning 的中文分词系统的影响。

4.1 实验数据

NTCIR-10 英中科技专利数据包括 100 万句中文语句，我们把这个语料作为科技领域的大规模生语料。为了制作测试集(TS)，随机选出 300 句，利用 3.2 节制定的科技领域分词标注标准进行人工分词标注，作为原领域的标注语料，利用新闻领域上 CTB5.0 中的第 1~270 篇、400~931 篇和 1001~1151 篇的标注数据。

从除去 TS 的语料中，利用 3.1 节描述的方法对所有语句计算与原领域的标注语料的差异性，并进行评分排序，选出高位的前 300 个句子(AS)，并根据第 3.2 节制定的分词标注标准进行人工标注。为了考察标注语料的规模对分词系统的影响，按如下方式构成 4 个标注语料集：前 50 句记作 AS1，前 100 句记作 AS2，前 200 句记作 AS3，前 300 句记作 AS4。

另外，为了对比基于 Active learning 的语句筛选方法，在除去 TS 与 AS 的语料中，随机抽取 300 句(RS)，同样地进行人工分词标注，并以同样方式构建 4 个标注语料集，分别记为 RS1，RS2，RS3 和 RS4。

4.2 实验设置

为了验证本文所提方法的有效性，首先利用基于 Active learning 方法制作小规模标注语料，并从科技领域的大规模生语料中抽取 n-gram 统计特征，将科技领域小规模标注语料与抽取的 n-gram 特征加入原有新闻领域的标注语料，训练出科技领域上的中文分词模型；然后利用该模型在测试集上进行评测。该系统记为 Our (Active learning + n-gram + 原领域语料)。为了考察目标领域上标注语料的规模对分词系统的影响，利用 4 个语料集(AS1，AS2，AS3 和 AS4)分别进行模型训练和评测。

为了考察基于 Active learning 的中文分词系统的优越性，进行了与上述相同的实验和评测，但加入的小规模标注语料是随机选取的 RS1，RS2，RS3 和 RS4。该系统记为 Baseline (随机+ n-gram +原领域语料)。

为了与其他领域自适应方法进行对比，我们重现苏晨[9]的方法训练了中文分词模型，同样利用生语料的 n-gram 特征和原有新闻领域的标注语料，另外利用 NTCIR 英中专利平行语料中的英文部分和英中对齐处理抽取新的特征。该系统记为 Su (英文+ n-gram + 原领域语料)。

为了对比没有进行领域自适应的分词系统，我们选取了利用原有新闻领域的标注语料训练的分词模型，该系统记为无领域自适应分词模型(原有新闻领域)，并进行相同测试集上的实验和评测。

利用公开的斯坦福中文分词系统在相同的测试集上进行评测，该系统记为无领域自适应分词模型

表 2　不同分词系统的评价结果

Table 2　Results of different word segmentation systems

中文分词系统		*P*	*R*	F1
Our (Active learning + n-gram + 原领域语料)	50	0.87750	0.89272	0.88504
	100	0.88206	0.89910	0.89050
	200	0.88251	0.90088	0.89160
	300	0.88366	0.90462	0.89402
Baseline (随机+ n-gram + 原领域语料)	50	0.87591	0.88831	0.88207
	100	0.87725	0.89066	0.88390
	200	0.87989	0.89441	0.88708
	300	0.88012	0.89620	0.88809
Su (英文+ n-gram + 原领域语料)		0.87975	0.88530	0.88252
无领域自适应分词系统(原有新闻领域)		0.69731	0.78949	0.74054
无领域自适应分词系统(斯坦福)		0.86355	0.88482	0.87406

(斯坦福)。

4.3　结果与分析

评测实验采用准确率(*P*)、召回率(*R*)、综合性能指标(F1 值)对中文分词系统进行测评。各个系统的评测结果如表 2 所示。

通过对比表 2 中的评测结果，分析得到如下结论。

1) 通过对比表中有领域自适应的 3 个分词系统和无领域自适应的两个分词系统的评测结果，可以看出，有领域自适应的分词系统的各项评测结果均高于无领域自适应的分词系统，说明领域自适应对改进中文分词系统性能的重要性。

2) 在领域自适应的 3 个分词系统中，通过对比加入目标领域的 300 句标注语料的分词系统和未加入标注语料的 Su 的分词系统的评测结果，可以看出，前者的各项评测结果均高于后者，说明目标领域人工标注语料对分词模型领域自适应有重要帮助，少量的 300 句语料就有明显效果。

3) 在加入人工标注的领域自适应的两个分词系统中，通过对比利用 Active learning 方法筛选语料的分词系统和随机筛选语料的分词系统，可以看出，在加入数量相同的标注语料的前提下，前者的各项指标的评测结果均高于后者，甚至前者添加 100 句标注语料得到的分词系统的评测结果高于后者添加 300 句标注语料得到的分词系统，表明本文提出的基于 Active learning 的中文领域自适应方法具有明显效果。

4) 在本文系统(Our)中，通过对比在不同规模的标注语料上的分词模型的评测结果，可以看出，随着标注语料的增大，系统的分词精度一直有所提升，300 句的标注语料还未达到饱和状态。我们预测，随着经过筛选的标注语料的加入，系统的性能还有提升空间。

5　结语

本文工作围绕中文分词领域自适应方法的探索，针对大规模人工分词训练语料难以获取的问题，提出基于 Active learning 的中文分词领域自适应方法，并应用于科技领域自适应的任务中，制定科技领域的人工标注标准，进行小规模的人工标注。在 NTCIR-10 的专利数据上，充分验证了该方法可以快速获取领域特有的分词知识，有效提高分词系统的领域适应性能。

本文针对中文分词领域自适应的任务，做出一些探索，并取得初步的研究成果。然而，中文分词还面临很多问题，例如在面对不同领域时，分词粒度的确定。今后将选取其他代表性的领域，对中文分词领域适应方法进行更深入的探索。

参考文献

[1] Rabiner L, Juang B. An introduction to hidden Markov models. ASSP Magazine, 1986, 3(1): 4–16

[2] Berger A, Della Pietra V, Della Pietra S. A maximum entropy approach to natural language processing.

Computational linguistics, 1996, 22(1): 39–71

[3] Lafferty J. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Proc International Conference on Machine Learning Williamstown USA, 2001, 3(2): 282–289

[4] Andrew G. A hybrid markov/semi-markov conditional random field for sequence segmentation // Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney: Association for Computational Linguistics, 2006: 465–472

[5] Sun X, Wang H, Li H. Fast online training with frequency-adaptive learning rates for Chinese word segmentation and new word detection // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island: Association for Computational Linguistics, 2012: 253–262

[6] Sun X, Zhang Y, Matsuzaki T, et al. A discriminative latent variable Chinese segmented with hybrid word/character information // Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Singapore: Association for Computational Linguistics, 2009: 56–64

[7] 张梅山，邓知龙，车万翔，等. 统计与词典相结合的领域自适应中文分词. 中文信息学报, 2012(2): 8–12

[8] Guo Z, Zhang Y, Su C, et al. Exploration of N-gram features for the domain adaptation of Chinese word segmentation // Natural Language Processing and Chinese Computing. Berlin: Springer, 2012: 121–131

[9] 苏晨，张玉洁，郭振，等. 适用于特定领域机器翻译的汉语分词方法. 中文信息学报，2013，27(5): 184–190

[10] Angluin D. Queries and concept learning. Machine Learning, 1988, 2(4): 319–342

[11] Settles B. Active learning literature survey [D]. Madison, WI: University of Wisconsin Madison, 2009: 127–131

[12] 宗成庆. 统计自然语言处理. 北京: 清华大学出版社, 2008

[13] GB/T 13715–1992 信息处理用现代汉语分词规范. 北京: 中国标准出版社,1992

[14] Xia F. The segmentation guidelines for the Penn Chinese Treebank (3.0) [R]. Philadelphia: University of Pennsylvania, 2000

[15] 段慧明，松井久人於，徐国伟，等. 大规模汉语标注语料库的制作与使用. 语言文字应用，2000(2): 72–77