中文分词模型的领域适应性方法

韩冬煦 常宝宝

(北京大学信息科学技术学院计算语言学研究所计算语言学教育部重点实验室 北京 100871)

摘 要 字标注分词方法是当前中文分词领域中一种较为有效的分词方法.但因为受制于训练语料的领域和规模,该方法在领域适应性方面效果不佳,影响了该方法在应用系统中的实际应用.在文中,作者提出使用卡方统计量以及边界熵提升未登录词的处理能力,并结合自学习和协同学习策略进一步改善字标注分词方法在领域适应性方面的性能.实验结果证实,文中提出的这些方法有效改善了分词方法的领域适应性.

关键词 卡方统计量;边界熵;领域适应性;自举算法;中文分词;社会计算;社交网络

中图法分类号 TP391 **DOI号** 10.3724/SP.J.1016.2015.00272

Approaches to Domain Adaptive Chinese Segmentation Model

HAN Dong-Xu CHANG Bao-Bao

(*Key Laboratory of Computational Linguistics, Ministry of Education, Institute of Computational Linguistics, School of Information Science and Technology, Peking University, Beijing* 100871)

Abstract Character-based tagging method is currently one of effective methods in Chinese Word Segmentation (CWS). Constrained by domain and size of the training corpus, this method doesn't work well in domain adaptability, affecting its use in practical application. This paper puts forward using chi-square statistics and boundary entropy to enhance the segmentation method in handling the Out-Of-Vocabulary words. Combined with self-training and co-training strategies, we further improve the performance of domain adaptability in CWS. Experiments show that with the use of these proposed methods, the domain adaptability of CWS is effectively improved.

Keywords chi-square statistics; boundary entropy; domain adaptability; bootstrapping method; Chinese Word Segmentation (CWS); social computing; social networks

1 引 言

中文分词是中文文本处理的基础,具有十分重要的理论和应用意义.它是指将组成句子的汉字序列用分隔符加以区分,切分成一个个单独的词.在过去的三十多年里,经过学者们的研究和探索,中文分词已取得了长足的进步,准确度获得了提升.特别是在使用了机器学习和基于统计的方法后,中文分词效果有了显著的进步[1].基于统计的中文分词方法不仅考虑了句子中词语出现的频率信息,同时也考虑到词语与上下文的关系,具备较好的学习能力,对歧义词和未登录词(Out-Of-Vocabulary,OOV)的识别有良好的效果,使得中文分词准确度比之前其他方法有了很大的提升.

中文分词方法中,有指导的字标注分词方法(Character-based tagging approach)[2]具有较好的分词效果.该方法本身需要标注语料,一般在处理与训练语料同领域的语料时,分词效果较好.根据ACL SIGHAN的评测数据,使用同一领域的测试

收稿日期:2012-11-27;最终修改稿收到日期:2014-11-22.本课题得到国家自然科学基金(60975054、61273318)和国家社会科学基金(06BYY048)资助.**韩冬煦**,男,1988年生,硕士研究生,主要研究方向为自然语言处理.E-mail:handx@pku.edu.cn;hweibo@126.com.
常宝宝,男,1971年生,博士,副教授,主要研究方向为自然语言处理.

语料，有指导的分词方法的 F 值可以高达 0.95①. 然而经验表明，换至其他不同领域的语料时，受标注语料领域和规模的限制，会产生领域的不适应性，分词的准确度常会大幅度下降. 文本领域涉及方方面面，为各个领域都制作标注好的语料极难实现. 所以，分词方法的领域适应性问题，成为一个值得关注的课题.

在跨领域中文分词中，由于文章领域内容的变化，不可避免地带来了诸多训练语料中未出现的领域词汇，使得未登录词识别问题成为跨领域分词的一个关键问题. 同时由于领域改变导致的上下文变化也会引起已登录词(In-Vocabulary，IV)处理性能的下降. 但综合而言，未登录词仍是引起跨领域分词性能下降的主要因素.

随着有指导中文分词方法的逐渐完善，跨领域中文分词的研究工作在近些年来才逐渐引起关注，研究人员不断探究新的方法. 常见的方法是基于字标注分词方法，通过引入新的特征来训练中文分词模型，加强中文分词模型的领域适应性. 如提出将 AV(Accessor Variety)统计量[3-5]、卡方统计量[6]作为特征加入模型等. 也有结合使用词典[7]、词表[8]或者汉语拼音表[9]的方法来改进字标注分词方法，或者使用自学习训练[10-11]的方法来加强字标注分词方法的领域适应性. 此外，也有学者尝试不使用字标注分词方法，将无指导中文分词的方法引入，比如使用词频[12]或者 N 元互信息(NGMI)[13]的方法，探索使用新的分词训练模型，使用结构化 SVM[14]、PA 算法(Passive-Aggressive)[15]、CMM[16]用作训练分词模型.

基于这些工作，我们认为，跨领域中文分词的难点在于如何有效提升领域变化带来的大量未登录词的识别效果，同时还应尽量避免对已登录词处理能力的下降. 在本文中，我们首先提出使用边界熵特征以及改进卡方统计量特征的使用方式，通过提取未标注语料中统计特性的方式改善分词模型的未登录词处理效果，改善中文分词的领域适应性. 在此基础上，我们也将自学习和协同学习策略引入分词模型训练，以迭代自举的方式进一步提高分词模型的领域适应性. 实验结果证实，边界熵特征、以及卡方统计量特征的引入和改进，并结合自举学习技术的使用，有效改善了跨领域中文分词的效果.

本文第 2 节简介字标注分词方法和相关工作；第 3 节介绍边界熵特征；第 4 节介绍卡方统计量特征及其使用方法的改进；第 5 节介绍自学习和协同学习在跨领域中文分词中的使用；第 6 节介绍实验步骤，列举结果以及相关的分析；第 7 节是结论.

2　字标注分词方法及相关工作

在有指导的中文分词方法中，字标注分词方法具有较好的分词效果. 该方法主要着眼于中文句子中的每个字(或符号)，它们在构造每一个词时占据一个确定的构词位置，以此作为分词的标记. 比如，规定一个字在构词时占有如下 4 种可能的位置：词的开端(begin)、多字词的中间部分(middle)、词的末尾(end)以及该字本身构成一个单字词(single). 这样就构成了一个四元标记集：$\{B,M,E,S\}$，通过它们标记每一个字以达到分词的目的.

字标注分词方法实施简单，准确度高，近年来很多分词工作沿着该方法展开. 比如论证使用 CRF 模型训练比使用最大熵模型更具有可靠性[17]，六词位标记集比四词位标记集能带来更高的准确度[18]. 在特征建模方面，除了广泛使用的基本特征外，为了进一步改善字标注分词方法的准确度，研究人员也提出了许多新的特征.

2.1　基本特征

字标注分词方法通常会采用下列特征模板，细微变化是特征元数目不同，我们称之为分词的基本特征：

(1) $C_n(n=-2,-1,0,1,2)$

(2) $C_nC_{n+1}(n=-2,-1,0,1)$

(3) $C_{n-1}C_{n+1}(n=-1,0,1)$

这里 C 代表一个字，n 代表当前字的相对位置.

基本特征在字标注分词方法中统计每个字在上下文中的语境关系，可以整体提升中文分词的准确度.

2.2　类型特征

为了有效地处理数字、字母等问题，前人提出使用类型特征[19]. 本文也使用类型特征，将文本中的字符按照汉字、数字、字母、标点符号以及其他类共 5 种类型加以标记区分. 其中，其他类是指不同于前 4 种任何一种的类型，如一些特殊数学符号、俄文法文等非英文的字母等. 模板如下：

(4) $T_n(n=-2,-1,0,1,2)$

(5) $T_nT_{n+1}(n=-2,-1,0,1)$

(6) $T_{n-1}T_{n+1}(n=-1,0,1)$

① http://www.sighan.org/bakeoff2005/data/results.php.htm

这里 T 代表一个字对应的类型特征，n 代表当前字的相对位置.

2.3 AV 统计量特征

AV 统计量特征[3-4]是近几年来被广泛使用的一个特征.它通过引入词表①，评估一个字串是否具有独立性，可否单独作为词语，在未登录词提取方面有良好的效果.该方法定义为

$$\mathrm{AV}(s)=\min\{\mathrm{Lav}(s),\mathrm{Rav}(s)\},$$

其中 s 是预期可能成为词的字串，$\mathrm{Lav}(s)$ 定义为 s 左边出现的不同词语的个数和 s 作为句首的次数的总和；$\mathrm{Rav}(s)$ 定义为 s 右侧出现的不同词语的个数和 s 作为句尾的次数的总和.公式中，统计不同词语数目时，词语来源于预先引入的词表.根据公式，模板定义为

(7) $V_n(n=-2,-1,0,1,2)$

(8) $V_nV_{n+1}(n=-2,-1,0,1)$

(9) $V_{n-1}V_{n+1}(n=-1,0,1)$

这里 V 代表一个字对应的 AV 统计量特征，n 代表当前字的相对位置.

因为在处理中文分词中引入了词表，AV 统计量特征在针对未登录词的提取，解决跨领域中文分词等问题上有针对性.不过，AV 统计量特征本身结构简单，带有的信息量少，分词准确度提升有限.对此，我们提出新的特征——边界熵特征，以及引入和改进使用卡方统计量特征，提升中文分词未登录词的处理能力，进一步改善字标注分词方法在跨领域中文分词方面的性能.

3 边界熵特征

边界熵过去常被用来提取中文文本中的短语[20].我们认为，将汉字的边界熵作为特征引入中文分词模型，可以捕获未标注目标领域语料的构词特征，改善有指导中文分词系统的性能和领域适应性.根据汉字左侧和右侧有不同的结果，边界熵分为左边界熵(Left Entropy)和右边界熵(Right Entropy).它的公式如下：

$$LH(c)=-\sum_{\forall a\in\{x|count(xc)\geq 1\}}P(a|c)\log(P(a|c)),$$

$$RH(c)=-\sum_{\forall b\in\{x|count(cx)\geq 1\}}P(b|c)\log(P(b|c)).$$

在上面的公式中，c 代表文本中的字符；$LH(c)$ 和 $RH(c)$ 分别表示字符 c 的左边界熵和右边界熵；$count(s)$ 表示字符串 s 在语料中出现的次数.$\{x|count(xc)\geq 1\}$ 表示字符 c 左侧出现的所有字符组成的集合，其中 xc 表示由字符 c 和其左边出现的 x 结合构成的字符串.$P(a|c)$ 表示在出现 c 的前提下，在 c 左侧出现字符 a 的条件概率.RH 公式与此同理.

通过上述公式计算得到的边界熵数据结果的离散程度比较大，彼此没有关联性，需要进行如下规范化处理：

$$LH_{\mathrm{norm}}(c)=\left\lfloor\frac{LH(c)-LH_{\min}}{LH_{\max}-LH_{\min}}\times k\right\rfloor,$$

$$RH_{\mathrm{norm}}(c)=\left\lfloor\frac{RH(c)-RH_{\min}}{RH_{\max}-RH_{\min}}\times k\right\rfloor,$$

其中 $LH_{\mathrm{norm}}(c)$ 表示字符 c 的左边界熵规范化结果，$LH_{\min}$ 是语料中所有字符的左边界熵中的最小值，$LH_{\max}$ 是语料中所有字符的左边界熵中的最大值.k 代表离散规范系数.RH 规范化公式与此同理.

根据边界熵理论，边界熵分别反映一个字左侧和右侧的不确定性.字符 c 的 LH 值较大，说明这个字符极有可能是一个词的开端；同样，当一个字符 c 的 RH 值较大，说明这个字符极有可能是一个词的末尾.

特征模板定义为

(10) $LH_n(n=-2,-1,0,1,2)$

(11) $LH_nLH_{n+1}(n=-2,-1,0,1)$

(12) $LH_{n-1}LH_{n+1}(n=-1,0,1)$

(13) $RH_n(n=-2,-1,0,1,2)$

(14) $RH_nRH_{n+1}(n=-2,-1,0,1)$

(15) $RH_{n-1}RH_{n+1}(n=-1,0,1)$

这里 LH 和 RH 分别表示一个字对应的左边界熵和右边界熵特征，n 代表当前字的相对位置.

边界熵特征统计了每个字与上下文的关系，同时又带有了每个字一定的位置信息，涵盖的信息量较多，能够整体提升中文分词准确度.

4 卡方统计量特征

4.1 卡方统计量特征介绍

卡方统计量特征用于计算两个字的关联度，能够有效识别未登录词，改善中文分词系统的性能和领域适应性[6].它的公式如下：

$$\chi^2(c_1,c_2)=\frac{n\times(a\times d-b\times c)^2}{(a+b)\times(a+c)\times(b+d)\times(c+d)},$$

其中，c_1 和 c_2 分别代表一个二元字组中连续的两个

① http://www.mandarintools.com/segmenter.html

字；a 代表语料中所有出现的二元字组为 c_1c_2 的次数；b 代表语料中所有出现的二元字组第 1 个字为 c_1 但第 2 个字不为 c_2 的次数；c 代表语料中所有出现的二元字组第 1 个字不为 c_1 但第 2 个字为 c_2 的次数；d 代表语料中所有出现的二元字组第 1 个字不为 c_1 且第 2 个字不为 c_2 的次数；n 代表语料中所有二元组的个数，即 $n=a+b+c+d$.

与边界熵特征类似，通过公式计算得到的卡方统计量的数据结果离散程度比较大，彼此没有关联性，需要进行如下规范化处理：

$$\chi^2_{\text{norm}}(c_1,c_2)=\left\lfloor\frac{\chi^2(c_1,c_2)-\chi^2_{\min}}{\chi^2_{\max}-\chi^2_{\min}}\times k\right\rfloor,$$

其中 $\chi^2_{\text{norm}}(c_1,c_2)$ 表示二元字组 c_1c_2 规范化后的结果，$\chi^2_{\min}$ 是所有卡方统计量中的最小值，$\chi^2_{\max}$ 是所有卡方统计量中的最大值. k 代表离散规范系数.

如果两个字的卡方统计量的值相对较大，说明这两个字的共现率高，结合程度紧密，很可能成为词或多字词的一部分；相反，卡方统计量的值相对较小，说明这两个字共现率低，难以结合，不成词的可能性高. 卡方统计量特征在提升未登录词的召回率方面有较为显著的效果[6].

由于卡方统计量特征值是基于两个字共同计算得到的，每一个字对应的卡方统计量特征值实际上是该字与其前或后的字计算得到的结果，所以卡方统计量在加入特征模板时要加入两个字共同的特征，特征模板如下：

(16) $X_nX_{n+1}(n=-2,-1,0,1)$

(17) $X_{n-1}X_{n+1}(n=-1,0,1)$

这里 X 代表一个字对应的卡方统计量特征，n 代表当前字的相对位置. 在特征模板中，我们加入了 $X_{n-1}X_{n+1}$ 特征，意在进一步加强字与字之间的关联性.

4.2　卡方统计量特征改进

卡方统计量特征是通过统计两个字的共现率，来判定两个字是否成词的，所以在统计时，语料中词汇的共现频率起着至关重要的作用. 在以往的实验中，我们通过分别计算训练语料和未标注语料的卡方统计量将结果用作中文分词特征，通过比对训练语料和未标注语料的卡方统计结果，寻求相同或匹配的值进行词语的识别，加以分词处理. 实验表明，这样做虽然使未登录词召回率有了明显的提升，但导致已登录词召回率降低，造成整体结果的下降[6].

经过分析，这是因为训练语料和未标注语料规模不同，致使两份统计数据在卡方分布上不完全一致，同一个词(已登录词)在训练语料和未标注语料分别计算得到的卡方统计量结果不同，在比对时并不具备可比性，导致已登录词召回率下降.

为了使两份语料的卡方统计量的计算结果一致，我们采取了新的方法：引入与测试语料相同领域的大规模未标注语料并将其合并到训练语料中一起计算得到统一的卡方统计量值，作为训练语料和未标注语料的特征值. 这样做避免了计算出两份彼此不一致的卡方统计量结果的情况，改进了卡方统计量特征的使用方法. 从结果上来看，改进后的方法保留了原先卡方统计量特征的效果，使得未登录词的召回率提升，同时又避免了给已登录词带来的负面影响.

5　半指导中文跨领域分词实现

为充分发挥大规模未标注语料的作用，我们也将自举式训练引入中文分词模型，分别使用了自学习(self-train)和协同学习(co-train)两种训练策略，结合使用字标注分词方法下的 CRF 模型，进一步改进分词模型的领域适应性，以提升跨领域中文分词的效果.

自学习训练策略[21-23]是根据已标注好的训练语料，训练生成出一个初始模型，用以标注未标注语料，再从标注结果中选择较好的结果加入训练语料，扩充训练语料并重新训练以得到新的模型，如此反复进行，直到模型处理性能收敛结束.

协同学习训练策略[24]与自学习不同之处在于协同学习需要建立两组不同的模型. 在本文中，我们使用相同的训练语料，通过使用两个不同的特征集，分别训练生成两个初始模型. 由于自学习训练策略的设定，每次只能将优秀的结果加入训练语料，不好的结果总难以得到改善，所以迭代到一定次数后，自学习会使本能够标注好的词语标注得更好，不能够很好标注的词语则一直得不到改善. 而协同学习与此不同，协同学习的策略总是在初始标注器的结果中选择优秀的标注结果补充到另一个标注器的训练语料中，因此在一定程度上弥补自学习策略的缺陷，一个标注器处理缺陷，可由另一个标注器帮助改善，两者相互取长补短，使结果进一步提升.

在本文中，我们使用已标注语料作为初始训练语料，并将跨领域的未标注语料引入模型，反复迭代，以得到具有领域适应性的分词模型. 在自学习训练中，我们选取基本特征、类型特征、边界熵特征和

卡方统计量特征作为特征集,反复迭代进行训练.在协同学习训练中,我们通过实验将几种特征集进行组合,比对效果,发现采用基本特征、类型特征与边界熵特征一组,以及采用基本特征、类型特征与卡方统计量特征一组,会得到较好的两组结果.所以我们将这两组分别用作协同学习的两组特征集,训练建立两套分类器,以提升跨领域中文分词的结果.

在中文分词中,词是基于它所在句子的上下文关系加以识别的.因此,在自举训练中,当一个字得到置信度高的标注结果后,单独将它放入训练语料中是没有意义的,而应该以整句话的方式加入训练语料中.在使用 CRF 标记未标注语料时,既可以得到每一个字的标注的置信度,也可以同时得到每个整句标注结果的正确置信度.当一个整句标注结果的正确置信度高于一定的阈值时,我们就可以将这句话和句中每个字对应的标注结果一并加入到训练语料中,作为训练集加以训练.在协同学习中,两个特征集会产生两组标注结果,但是由于两个特征集的分词能力相当,因而往往在同一句话的分词结果整句置信度上,两者相差无几.所以,当整句的标记置信度高于设定的阈值后,我们把比对的信息进一步细化至每一个字的层级上.实验中,使用 CRF 标记未标注语料时,可以将每个字对应的各个标记的边缘概率均求出来,用于计算选择合适的标记作为分词结果.算法如下:

(1)先通过阈值筛选,优先保留本特征集标注结果中置信度高的字;(2)当本特征集得不到一个字的置信度高的结果时,从另一个特征集的标注结果中寻找相同位置的汉字标注结果,通过阈值筛选,从另一个特征集中采纳结果协同学习;(3)当两个特征集均无法得到置信度高的标注结果时,通过统计的方法,加以使用概率模型计算,进一步分析出合适的标记:由于标记本身制定的规则,在标记与标记之间,有一定的转移关系.如标注为"S"或"E"的字,下一个字的标记只能为"B"或"S";标注为"B"的字,下一个字的标记只能是"B1"或者"E".设 w_1w_2 表示前后两个字,c_1c_2 为 w_1w_2 对应的标记序列,概率模型的公式如下:

$$c_2^{\#}=\arg\max_{c_2} p(c_2|w_2)p(c_2|c_1).$$

上式中 $c_2^{\#}$ 指汉字 w_2 的最佳标记.

在该概率模型中,每一项的概率值都需要事先统计.先将训练语料进行读取,统计语料中每个字与标记之间的对应次数,标记与标记之间的转移关系,并适当加以平滑,计算各自的对应概率,以此可以计算出最合适的标注结果.

通过上面的两种自举方法,我们将训练语料不断扩大语句数量,尽可能改善训练语料的领域适应性.从训练的结果来看,通过自举式学习,跨领域中文分词的 F 值、未登录词的召回率,得到了很好的提升.

6 实验及结果分析

本次实验采用字标注分词方法,使用 CRF[①] 进行训练.训练标记语料采用北京大学为 SIGHAN bakeoff2005 提供的训练语料[②].该训练语料全部为《人民日报》1998 年 1 月份文本内容(170 万字),属于新闻性质的语言,语句规范,而且文本中所有的拉丁字母、阿拉伯数字和标点符号均采用中文全角格式书写.

测试语料分为 3 种,文体领域分别为新闻、文学和电脑.新闻语料来自于《人民日报》1998 年 2 月份部分文本内容(16.6 万字),其语句规则和字符格式与训练语料一致.该语料用于对比,测试跨领域分词模型在相同领域下的分词效果.文学(10.4 万字)和电脑(18.7 万字)领域的语料用于测试跨领域分词效果,语料中含有许多不规范的语句,字符中含有大量半角形式的数字、英文字母和标点符号,甚至含有数学符号、法文单词以及拼写错误.这两份语料的标准答案为纯手工切分,切分标准依据《北京大学汉语语料库加工规范》[③].

测试语料字数及未登录词所占比重如表 1 所示.

表 1 测试语料字数和未登录词比重对比

测试语料	总字数	未登录词比重
新闻	166000 字	0.08024
文学	104774 字	0.10181
电脑	131001 字	0.31730

实验中,我们使用近几年来被普遍使用的 AV 统计量特征作为对比,使用标准的中文分词评测标准,包括准确率、召回率、F 值以及已登录词的召回率、未登录词的召回率.

在以下结果列举中,我们使用:

B 代表基本特征(1)(2)(3);

T 代表类型特征(4)(5)(6);

① http://crfpp.sourceforge.net/

② 可以从 http://www.sighan.org/bakeoff2005/下载.

③ 详情见 http://www.sighan.org/bakeoff2005/.

AV 代表 AV 统计量特征(7)(8)(9)；

H 代表边界熵特征(10)(11)(12)(13)(14)(15)；

X 代表改进后的卡方统计量特征(16)(17).

6.1　使用前人提出的特征得到的结果

表 2 是使用基本特征、类型特征和 AV 统计量特征得到的最好的结果. 从结果来看，字标注分词方法是中文分词领域中一种较为有效的分词方法. 在相同领域下，仅使用基本特征，F 值就高达 0.96. 但是在不同领域下，F 值则出现明显的下降. 特别是在电脑领域，在只使用基本特征时，F 值的准确率仅为 0.78，未登录词的召回率仅为 0.33，中文分词的领域适应性问题凸显.

表 2　使用基本特征、类型特征和 AV 统计量特征下的分词结果

领域	特征	Precision	Recall	F	IV Recall	OOV Recall
新闻	B	0.966952	0.964722	0.965835	0.970625	0.722769
	B、T	0.968689	0.967388	0.968038	0.973191	0.729576
	B、T、AV	0.968951	0.967686	**0.968318**	0.973551	0.727307
文学	B	0.910128	0.901999	0.906045	0.919177	0.634826
	B、T	0.912851	0.908529	0.910685	0.924980	0.652651
	B、T、AV	0.916209	0.910530	**0.913361**	0.927199	0.651280
电脑	B	0.744152	0.829369	0.784445	0.922997	0.339519
	B、T	0.919212	0.888648	0.903672	0.924364	0.686122
	B、T、AV	0.919916	0.900044	**0.909872**	0.940240	0.689170

当加入类型特征、AV 统计量特征后，中文分词的领域适应性有所改善，这也是近年来 AV 统计量被广泛使用的原因. 不过，从结果来看，在跨领域分词中，提升效果并不是很明显. 而且，与相同领域下的中文分词相比，F 值的差距十分明显.

6.2　边界熵特征和卡方统计量特征规范参数选择

在实验中，我们使训练语料与未标注语料字数之比约为 1∶10，引入与测试语料相同领域的未标注语料，合并到训练语料中一起计算得到统一的边界熵特征和卡方统计量特征.

边界熵和卡方统计量特征都需要进行离散化处理，基于实验，我们将边界熵特征离散度 k 值设定为 5，将卡方统计量特征的离散度 k 值设定为 20. 表 3 为分别使用边界熵特征和卡方统计量特征所取得的分词结果.

表 3　边界熵特征和卡方统计量特征的离散度选择结果比对

领域	特征	离散度 k	Precision	Recall	F	IV Recall	OOV Recall
文学	B、T、H	5	0.920667	0.914147	0.917396	0.929285	0.678702
	B、T、X	20	0.927459	0.920277	0.923854	0.933002	0.722349
电脑	B、T、H	5	0.926535	0.928091	0.927312	0.943384	0.847857
	B、T、X	20	0.920368	0.929813	0.925066	0.941966	0.866057

从结果来看，相比于 AV 统计量特征，边界熵特征和卡方统计量特征在跨领域分词结果的已登录词召回率和未登录词召回率均得到了明显提升，特别是未登录词的召回率，文学领域达到了 0.72，电脑领域达到了 0.86，已经超过了相同领域下的未登录词召回率. 在 F 值方面，相比与 AV 统计量特征又提升了 1 个百分点. 整体结果得到了较好改善.

6.3　边界熵使用方法选择

在实验中，因为边界熵特征有左右两个数值，我们考虑过只采用左边界熵和右边界熵的最小值或最大值作为特征，抑或在选择左右边界熵共同使用作为特征时，增加了一个新的特征模板：

$LH_nRH_n(n=-2,-1,0,1,2)$

这些方案中，我们采用文学领域的测试语料做对比，结果如表 4. 从结果可以看出，在同时选取左右边界熵作为特征，且不改变特征模板的情况下，结果最好.

表 4　边界熵特征模板不同使用方法的结果比对

方案	特征	Precision	Recall	F	IV Recall	OOV Recall
最小边界熵	B、H	0.916486	0.910793	0.913631	0.927404	0.652422
最大边界熵	B、H	0.916816	0.911994	0.914398	0.928492	0.655393
左右边界熵	B、H	**0.920667**	**0.914147**	**0.917396**	**0.929285**	**0.678702**
左右边界熵(改变模板)	B、H	0.919760	0.913360	0.916549	0.929491	0.662477

6.4 优化卡方统计量的比对

表 5 是使用改进前和改进后的卡方统计量，选用文学领域的测试语料得到的一份跨领域分词的对比结果. 从结果可以看出，当测试语料未合并到训练语料时，未登录词召回率明显上升，却带来 F 值下降，对已登录词产生了负面影响. 当采取引入未标注语料，改进卡方统计量的计算的方法后，未登录词召回率仍旧有大幅度提升，且各项都有正面作用，说明引入同领域未标注语料，改进卡方统计量的计算方法，对已登录词、未登录词具有正面作用.

表 5 卡方统计量特征使用方法优化前后的比对

特征	Precision	Recall	F	IV Recall	OOV Recall
B、T	0.912851	0.908529	0.910685	0.924980	0.652651
B、T、X(原)	0.912370	0.903214	0.907769	0.915724	0.708638
B、T、X(改)	**0.927459**	**0.920277**	**0.923854**	**0.933002**	**0.722349**

6.5 自学习结果

表 6 和表 7 是使用自学习算法分别在文学和电脑领域得到的迭代分词结果. 边界熵特征和卡方统计量特征共同使用时，既有互补的一面，也有干扰彼此的一面，所以从结果中可以看到，初始状态时的未登录词召回率比单独使用卡方统计量特征要低. 随着迭代次数增加，各项值稳步增加. 随着迭代达到一定次数之后，F 值和未登录词的召回率开始出现波动，说明自学习所起的作用开始变小.

表 6 自学习算法在文学领域的迭代分词结果

迭代次数	Precision	Recall	F	IV Recall	OOV Recall
1	0.927973	0.920953	0.924450	0.934075	0.716865
2	0.928426	0.921505	0.924953	0.934516	0.719150
3	0.928231	0.921478	0.924842	0.934427	0.720064
4	0.927980	0.921574	0.924766	0.934604	0.718921
5	0.928191	0.921464	0.924815	0.934545	0.718007
6	0.931444	0.924487	0.927952	0.937160	0.727377
7	0.932042	0.924321	0.928165	0.936484	0.735146
8	0.933872	0.926434	**0.930138**	0.938306	**0.741773**
9	0.933212	0.926461	0.929824	0.939291	0.726920
10	0.933029	0.926061	0.929532	0.938806	0.727834

表 7 自学习算法在电脑领域的迭代分词结果

迭代次数	Precision	Recall	F	IV Recall	OOV Recall
1	0.920573	0.929268	0.924900	0.943247	0.855926
2	0.926445	0.931205	0.928819	0.944854	0.859602
3	0.924027	0.932038	0.928015	0.945349	0.862202
4	0.926747	0.932985	0.929855	0.945247	0.868657
5	0.927892	0.933100	0.930489	0.944871	0.871347
6	0.928352	0.933789	**0.931063**	0.945247	**0.873678**
7	0.926186	0.932368	0.929267	0.944375	0.869374
8	0.926886	0.933272	0.930068	0.945195	0.870719
9	0.926545	0.933487	0.930003	0.945298	0.871526
10	0.926073	0.933516	0.929779	0.945418	0.871078

图 1 是文学领域和电脑领域的 F 值变化曲线. 从图中看到随着迭代次数增加，F 值不断提升，当迭代达到一定次数时，F 值出现波动和下降，说明自学习所起的作用开始变小.

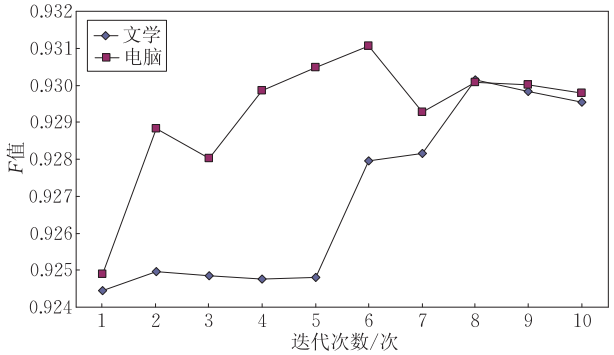

图 1 自学习算法在文学及电脑领域的 F 值迭代变化曲线

6.6 协同学习结果

表 8 和表 9 是使用协同学习算法在文学和电脑领域得到的分词结果. 从协同学习的结果来看，协同学习比自学习需要的迭代次数更少，得到的结果更好. 随着迭代的次数增加，准确率、召回率和未登录词的召回率逐步提升，最后趋于稳定.

表 8　协同学习算法在文学领域的迭代分词结果

特征集 1：基本特征、类型特征与边界熵特征					
迭代次数	Precision	Recall	F	IV Recall	OOV Recall
1	0.920667	0.914147	0.917396	0.929285	0.678702
2	0.927431	0.920415	0.923909	0.933311	0.719835
3	0.928560	0.921202	0.924866	0.934016	0.721892
4	0.930033	0.922831	0.926418	0.935103	**0.731947**
5	0.930200	0.922803	**0.926486**	0.935382	0.727148
6	0.929189	0.921864	0.925512	0.934898	0.719150
特征集 2：基本特征、类型特征与卡方统计量特征					
迭代次数	Precision	Recall	F	IV Recall	OOV Recall
1	0.927459	0.920277	0.923854	0.933002	0.722349
2	0.929651	0.923272	0.926451	0.935691	0.730119
3	0.933434	0.926682	0.930046	0.938732	0.739260
4	0.933890	0.927869	0.930870	0.939746	0.743144
5	0.934655	0.928242	**0.931437**	0.939967	**0.745887**
6	0.934308	0.927317	0.930799	0.939100	0.744059

表 9　协同学习算法在电脑领域的迭代分词结果

特征集 1：基本特征、类型特征与边界熵特征					
迭代次数	Precision	Recall	F	IV Recall	OOV Recall
1	0.926535	0.928091	0.927312	0.943384	0.847857
2	0.924560	0.932124	0.928326	0.944905	0.865071
3	0.926705	0.933861	0.930269	0.946067	0.869822
4	0.929954	0.933731	0.931839	0.945896	0.869912
5	0.930303	0.934162	0.932229	0.946118	0.871436
6	0.931653	0.935411	**0.933528**	0.946990	**0.874664**
特征集 2：基本特征、类型特征与卡方统计量特征					
迭代次数	Precision	Recall	F	IV Recall	OOV Recall
1	0.920368	0.929813	0.925066	0.941966	0.866057
2	0.924402	0.932124	0.928247	0.944683	0.866236
3	0.927362	0.934736	0.931034	0.946563	0.872691
4	0.929951	0.934837	0.932388	0.946699	0.872602
5	0.931052	0.935382	0.933212	0.946563	0.876726
6	0.932152	0.935698	**0.933922**	0.946733	**0.877802**

图 2 和图 3 分别是在文学和电脑领域内准确率特征集和召回率特征集的 F 值变化曲线. 从图中看到两个特征集的 F 值先不提升，随着迭代的次数增加，文学领域的两个 F 值开始下降，电脑领域的两个 F 值趋于相同.

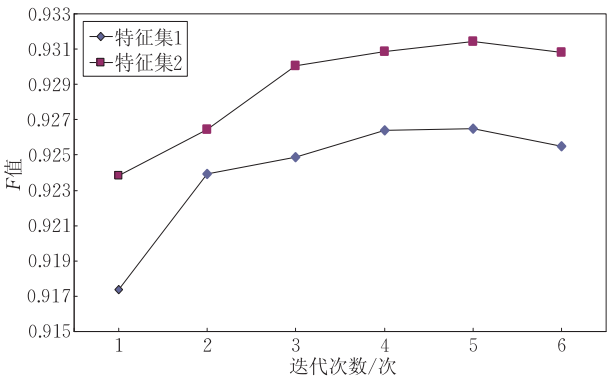

图 2　协同学习算法在文学领域 F 值迭代变化曲线

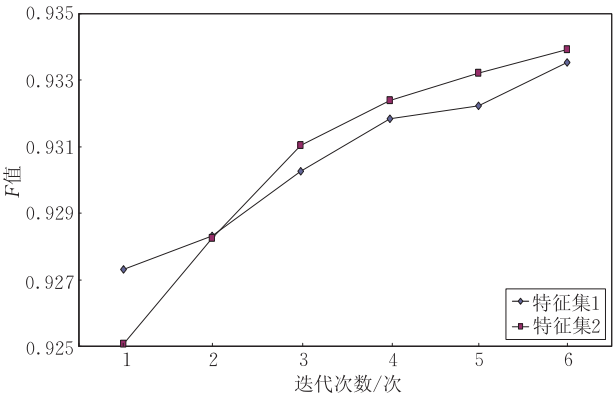

图 3　协同学习算法在电脑领域 F 值迭代变化曲线

6.7　最终结果

经过边界熵特征和卡方统计量特征的共同训练，结合自举学习训练，最终结果如表 10 所示. 从结果中，我们可以看到跨领域分词结果有了大幅度的提升. 在文学和电脑领域的 F 值都达到了 0.93，相比于之前使用 AV 统计量特征的结果，提升了两个百分点. 特别是在未登录词召回率方面，提升结果更加明显，文学领域达到了 74.59%，电脑领域达到了 87.78%，甚至超过了同领域下的未登录词召回率. 这是由于

电脑领域的未登录词的比重有 31.7%，且语料中含有大量专业术语以及英文单词，如“触摸屏”、“电阻”、“分区表”和“MySQL”、“Java”等，本文引入的各种统计量特征可以有效地识别这些未登录词.

表 10 最终实验结果

领域	特征	Precision	Recall	F	IV Recall	OOV Recall
新闻	B、T、H、X	0.969043	0.968091	0.968566	0.973920	0.747391
文学	B、T、H、X	0.934655	0.928242	**0.931437**	0.939967	**0.745887**
电脑	B、T、H、X	0.932152	0.935698	**0.933922**	0.946733	**0.877802**

在同领域下，我们也做了对比分词实验，从结果来看，分词结果也有一定的提升，不过并不是很明显. 说明边界熵特征和卡方统计量特征在领域适应性方面的提出是非常有针对性的. 在实验中，我们没有使用预处理和后处理，以上数据均完全由自动分词结果获得. 如果再使用预处理和后处理，结果可以有更进一步的提升.

7 分析和结论

从上述实验的结果来看，边界熵特征和卡方统计量特征在跨领域中文分词模型的建立上是可行的，它为改善中文分词模型的领域适应性方面起到了一定的效果. 再结合自举学习技术的使用，有效改善了跨领域中文分词的效果. 从实验的结果来看，跨领域中文分词的未登录词的召回率得到了明显的提升，同时已登录词的召回率也有小幅度的增长.

根据分词结果分析，由于有了卡方统计量的引入，使得很多结合紧密的未登录词因为卡方值较高，得到了有效的识别，例如文学领域中的词汇“百感丛生”、“恶鬼”等，电脑领域中的“编程”、“服务器”等领域术语以及大量数字或单词等.

通过实验分析，我们发现协同训练对初始标注结果也有一定的纠正能力，例如在测试语料中，有一句话为“年岁渐长远走他乡”，因为训练语料（新闻语料）中出现了较多次的“长远”，且这两个字的卡方统计量分值也比较高，因而初始标注器给出的结果为

BESBEBME 年岁 /渐 /长远 /走他乡

通过引入领域未标注语料，经过协同学习和概率模型的修正，分词模型最终得到了正确的结果：

BEBEBMME 年岁 /渐长 /远走他乡

当然，分词结果中也仍然存在负面的例子，但总体而言，本文通过结合使用边界熵特征、改进后的卡方统计量特征以及协同学习策略，中文分词模型的领域适应性有了较大的改善，有效提升了跨领域未登录词汇的识别性能，也小幅提升了跨领域已登录词的召回率. 不过，与同领域中文分词结果相比，总体分词性能仍有一定差距. 考虑到已登录词在测试语料中占有更大的比重，已登录词准确率、召回率的提升，可以给准确度带来进一步的提升. 这也是跨领域中文分词需要进一步考虑的问题.

参 考 文 献

[1] Huang Chang-Ning, Zhao Hai. Chinese word segmentation: A decade review. Journal of Chinese Information Processing, 2007, 21(3): 8-20(in Chinese)
(黄昌宁，赵海. 中文分词十年回顾. 中文信息学报，2007，21(3)：8-20)

[2] Xue Nianwen. Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processing, 2003, 8(1): 29-48

[3] Feng Haodi, et al. Accessor variety criteria for Chinese word extraction. Association for Computational Linguistics, 2004, 30(1): 75-93

[4] Feng Haodi, et al. Unsupervised segmentation of Chinese corpus using accessor variety//Proceedings of the 1st International Joint Conference on Natural Language Processing. Hainan Island, China, 2004: 255-261

[5] Huang Degen, Tong Deqin, Luo Yanyan. HMM revises low marginal probability by CRF for Chinese word segmentation //Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing 2010. Beijing, China, 2010: 216-220

[6] Chang Baobao, Han Dongxu. Enhancing domain portability of Chinese segmentation model using chi-square statistics and bootstrapping//Proceedings of the 2010 Conference on Empirical Methods on Natural Language Processing. Massachusetts, USA, 2010: 789-798

[7] Shen Jianping, et al. Chinese word segmentation based on mixing multiple preprocessor and CRF//Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing 2010. Beijing, China, 2010: 270-273

[8] Xu Xiaoming, et al. High OOV-recall Chinese word segmenter //Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing 2010. Beijing, China, 2010: 252-255

[9] Jiang Huixing, Dong Zhe. An double hidden HMM and an CRF for segmentation tasks with Pinyin's finals//Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing 2010. Beijing, China, 2010: 277-281

[10] Wang Kun, et al. A character-based joint model for CIPS-SIGHAN word segmentation Bakeoff 2010//Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing 2010. Beijing, China, 2010: 245-248

[11] Qin Xiao, et al. CRF-based experiments for cross-domain Chinese word segmentation at CIPS-SIGHAN-2010// Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing 2010. Beijing, China, 2010: 261-265

[12] Jiang Tianjian, et al. Term contributed boundary tagging by conditional random fields for SIGHAN 2010 Chinese word segmentation bakeoff//Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing 2010. Beijing, China, 2010: 266-269

[13] Tang Ling-Xiang, et al. A boundary-oriented Chinese segmentation method using n-gram mutual information// Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing 2010. Beijing, China, 2010: 234-239

[14] Zhang Chongyang, et al. A Chinese word segmentation system based on structured support vector machine utilization of unlabeled text corpus//Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing 2010. Beijing, China, 2010: 221-227

[15] Gao Wenjun, et al. Chinese word segmentation with online passive-aggressive algorithm//Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing 2010. Beijing, China, 2010: 240-244

[16] Wu Yu-Chieh, et al. Chinese word segmentation with conditional support vector inspired Markov models//Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing 2010. Beijing, China, 2010: 228-233

[17] Tseng H, et al. A conditional random field word segmenter for SIGHAN Bakeoff 2005//Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing. Jeju Island, Korea, 2005: 168-171

[18] Zhao Hai, et al. Effective tag set selection in Chinese word segmentation via conditional random field modeling// Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation. Wuhan, China, 2006: 87-94

[19] Low J K, et al. A maximum entropy approach to Chinese word segmentation//Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing. Jeju Island, Korea, 2005: 161-164

[20] Kang Byeong-Kwu. A Study on the Chinese-Korean Computer-Aided Translation for Chinese Monographs [Ph. D. dissertation]. Peking University, Beijing, 2006(in Chinese)
(姜柄圭. 面向中文专著的汉韩机器辅助翻译研究[博士学位论文]. 北京大学, 北京, 2006)

[21] Efron B. Bootstrap methods: Another look at the jackknife. Annals of Statistics, 1979, 7(1): 1-26

[22] Efron B. Computers and the theory of statistics: Thinking the unthinkable. SIAM Review, 1979, 21(4): 460-480

[23] Efron B. Censored data and the bootstrap. Journal of the American Statistical Association, 1981, 76(374): 312-319

[24] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training//Proceedings of the 11th Annual Conference on Computational Learning Theory. 1998: 92-100

HAN Dong-Xu, born in 1988, M. S. candidate. His major research interest is natural language processing.

CHANG Bao-Bao, born in 1971, Ph. D., associate professor. His major research interest is natural language processing.

Background

As the basis of Chinese information processing, Chinese Word Segmentation (CWS) plays a very significant role nowadays. In Chinese Word Segmentation, character-based tagging method is currently one of effective methods, which has achieved content results in the same domain; yet constrained by domain and size of the training corpus, the results in the out of domain could not be approved.

In this paper, we put forward the use of chi-square statistic and boundary entropy to enhance the segmentation method in handling the Out-Of-Vocabulary words. Moreover, this paper brings up the method of using two bootstrapping methods: self-training and co-training in Chinese Word Segmentation, which is in the way of training the better results of segmentation model to achieve optimal segmentation results. With the use of these proposed methods, the performance of Chinese Word Segmentation domain adaptability is effectively improved.

The research is supported by the National Natural Science Foundation of China (Grant Nos. 60975054, 61273318) and the National Social Science Foundation of China (Grant No. 06BYY048).

During the past few years, the authors of this paper have performed some researches on Chinese Word Segmentation. They have published some papers in this area.