基于边界熵和卡方统计量的多领域适应性中文分词方法*

韩冬煦，常宝宝
北京大学 计算语言所，北京 100871
北京大学 计算语言学教育部重点实验室，北京 100871
E-mail: handx@pku.edu.cn; chbb@pku.edu.cn

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。本文采用有指导的学习方法，基于 CRF 模型，提出使用边界熵和卡方统计量相结合的特征，进一步改善字标注分词方法的性能。同时，我们也就 AV(Accessor Variety)统计量等当前普遍使用的特征进行了对比。从结果来看，边界熵和卡方统计量的引入，在跨领域适应性上，比其他特征有更好的表现。

关键词：中文分词；边界熵；卡方统计量；字标注分词方法

Improving the Domain Adaptability of Chinese Word Segmentation Models with Boundary Entropy and Chi-square Statistics

Han Dongxu, Chang Baobao
Institute of Computational Linguistics, Peking University, Beijing 100871
Key Laboratory of Computational Linguistics (Peking University), Ministry of Education, Beijing 100871
E-mail: handx@pku.edu.cn; chbb@pku.edu.cn

Abstract: Character-based tagging method is one of effective methods in Chinese Word Segmentation (CWS) nowadays. Using the CRF model in a supervised learning method, this paper proposes a new approach to improve the segmentation method by the combination of the boundary entropy and chi-square statistic. We also pose contrasts with AV (Accessor Variety) and other features currently widely used. Experiments show that the approach performs better than other features in improving the domain adaptability.

Keywords: Chinese word segmentation (CWS); boundary entropy; chi-square statistics; character-based tagging approach of CWS

1 引言

中文信息处理中，中文分词作为一项基础工作，具有重要的意义。过去的十几年间，经过一系列研究探索，中文分词已取得长足的进步，准确性大为提升。特别是在使用机器学习和基于统计方法后，分词效果有了显著进步[1]。尤其字标注分词方法[2]，使中文分词准确性有了明显提升。该方法主要将中文语句着眼于每个语素，语素在构词时占据一个确定的构词位置，以此添加分词的标记。比如一个语素在构词时规定有如下四种可能位置：词的开端(begin)、多字词中间部分(middle)、词的末尾(end)及该语素构成单字词(single)。这样就形成了一个四元标记集：{B、M、E、S}，用它们标记每个语素分词。近年来，更多分词工作沿着该方法展开，分词方法进一步完善，如提出用 CRF 模型训练比原来使用最大熵模型更具有可靠性[3]，六词位标记集比四词位标记集能带来更高准确度[4]，引入更多新特征模板，如数字、字母、标点等类型特征模板的引入[5]。

有指导分词方法日渐成熟，其他无指导和半指导分词理论也不断进步。如基于互信息(Mutual Information)的无指导分词方法[6]，完全使用未标注语料进行训练得到模型加以分词；再如半指导使用大量未标记语料和少量已标记语料相互结合的方式训练[7]，反复优化得到最优模型分词。这些方法由于语料中可以得到的信息比有指导方法少，所以结果比有指导要低，比如前者无指导分词，在评价分词结果调和平均值(F 值)为 0.85[6]，远远低于有指导分词方法中字标注方法的 0.94。

* 本文工作得到自然科学基金项目(60975054)和社会科学基金项目(06BYY048)的支持。

有指导方法的高准确度，多数是在相同内容领域下得到的。根据 CIPS (Chinese Information Processing Society)的数据，有指导分词方法在同领域内分词处理，F 值高达 0.95。在不同领域内进行分词处理，结果则远远低于此。字标注分词作为有指导方法，同样具有该问题。中文涉及方方面面，我们不可能为各个领域都制作标注好的训练语料；我们更不能因为领域适应性问题，放弃有指导这种可行的分词手段。所以，跨领域分词适应性问题，成为一个值得关注的课题。

跨领域分词中，影响分词结果的一个主要原因，是 OOV(Out-Of-Vocabulary)词汇的影响。实际上，OOV 词汇的负面影响在同领域也存在，不过由于适应性甚至过适应训练产生了弥补，同领域分词结果没有受到 OOV 词汇明显影响。而放到跨领域语料下分词，这个影响就显而易见了。

在针对于跨领域分词和解决 OOV 词汇识别上，研究工作一直在不断进行中，人们通过在训练语料中不断添加新的特征，加强分词的领域适应性。如引入汉语拼音作为特征的方法[8]，使用最长匹配算法[9]，以及 n 元互信息的方法[10]，这些方法主要立足点在引入一些无指导的方法或者增加词典的方法来平衡训练语料带来的过适应问题，虽然在整体结果上有所提升，但由于着眼点并非在 OOV 词汇上，OOV 词汇带来的首要影响没有解决，效果并不显著。

本文提出利用边界熵和卡方统计量的策略改善分词系统的性能和领域适应性，从实验结果看，边界熵和卡方统计量特征的引入改善了分词效果。

接下来几部分中，第 2 章节简介字标注分词基本特征及目前根据字标注分词建立的常见特征；第 3 章节主要介绍我们提出的新特征——边界熵，以及边界熵和卡方统计量在特征使用上的分析；第 4 章节介绍实验步骤，列举结果和相关的讨论；第 5 章节是结论以及今后进一步的改进。

2 字标注分词及相关特征

2.1 标记集选择

字标注分词方法主要特征为四词位标记集：{B、M、E、S}；后经实验[4]，证实六词位标记集{B、B1、B2、M、E、S}准确性更好。后者在四词位标记集基础上添加 B1 和 B2 两标记，分别代表多字词第二和第三个字的位置，M 变成代表多字词第四个及其后面所有非结尾字的位置。

2.2 基本特征

字标注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，称之为分词基本特征：

(a) $C_n(n=-2,-1,0,1,2)$ (b) $C_nC_{n+1}(n=-2,-1,0,1)$ (c) $C_{n-1}C_{n+1}(n=-1,0,1)$

这里 C 代表字，n 代表当前字相对位置。通过上述模板，我们建立起一个 5 元的特征模板。

2.3 类型特征

为了有效处理数字、字母等问题，人们也提出使用类型特征[5]。在本文中，我们也使用了类型特征，将文本中的字符区分为汉字、数字、字母、标点符号以及其他共 5 种类型，其他类是指不同于前四种任何一个的类型，如一些数学符号、非英文字母等。模板如下：

(d) $T_n(n=-2,-1,0,1,2)$ (e) $T_nT_{n+1}(n=-2,-1,0,1)$ (f) $T_{n-1}T_{n+1}(n=-1,0,1)$

这里 T 代表字对应的类型特征，n 代表当前字的相对位置。

2.4 AV 统计量特征

为了改善分词领域适应性，近年来人们提出用 AV(Accessor Variety)统计量[11]。AV 统计量用于评估字串是否具有独立性，可否单独作为词语，在 OOV 词提取方面有较好效果。方法定义为：

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

其中 s 为预期可能成词的字串，$L_{av}(s)$定义为 s 前面出现不同字的个数 s 作为句首次数的总和；$R_{av}(s)$定义为 s 后面出现不同字的个数与 s 作为句尾次数的总和。根据公式，定义模板为：

(g) $V_n(n=-2,-1,0,1,2)$ (h) $V_nV_{n+1}(n=-2,-1,0,1)$ (i) $V_{n-1}V_{n+1}(n=-1,0,1)$

这里 V 代表字对应的 AV 统计量特征，n 代表当前字的相对位置。

AV 统计量主要针对的是 OOV 词汇的提取，在解决跨领域分词的首要问题上有针对性。所以在跨领域分词上，AV 统计量的使用，使结果有较高提升。不过，AV 统计量特征本身结构简单，带有信息量少，在准确率和 IV(In-Vocabulary)词汇的召回率上产生了负面作用，使结果降低。

2.5 卡方统计量特征

卡方统计量用于计算两个字的关联度，可以用来改善分词系统的领域适应性[12]。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 代表连续的两个字构成的二元字组；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一个字为 c_1 第二个字不为 c_2 的次数；c 代表语料中所有出现的二元字组第一个字不为 c_1 第二个字为 c_2 的次数；d 代表语料中所有出现的二元字组第一个字不为 c_1 且第二个字不为 c_2 的次数；n 代表语料中所有二元组的个数，即 $n=a+b+c+d$。

这样得到的卡方统计量数据离散程度比较大，彼此没有关联性，还需要进行规范化：

$$\chi^2_{norm}(c_1,c_2)=\left\lfloor\frac{\chi^2(c_1,c_2)-\chi^2_{\min}}{\chi^2_{\max}-\chi^2_{\min}}\times 10\right\rfloor$$

卡方统计量值越大，说明两个字结合越紧密，越可能成为词或多字词的一部分；相反，卡方统计量值越小，说明两个字不成词概率越高。卡方统计量提升 OOV 词召回率上有显著效果[12]。

由于卡方统计量特征是基于两个字计算得到的特征值，每一个字对应的卡方统计量特征值实际上是与其前后的字相关的，所以卡方统计量在加入特征模板时没有单字特征，特征模板如下：

(j) $X_nX_{n+1}(n=-2,-1,0,1)$ (k) $X_{n-1}X_{n+1}(n=-1,0,1)$

X 为卡方统计量特征，n 为字相对位置。模板中加入 $X_{n-1}X_{n+1}$ 特征，意在加强字之间关联性。

与 AV 统计量特征类似，卡方统计量特征的引入，也带来了准确率和 IV 词汇的召回率的降低，产生了负面作用，使结果降低。基于此，我们继续探究了卡方统计量新的使用方法，并提出了新的特征方案——边界熵特征，进一步改进跨领域分词的结果。

3 边界熵特征和卡方统计量的改进

3.1 边界熵特征

边界熵常被用来提取文本中的高频词[13]或者短语[14]。本文提出将其用于分词特征，改善有指导分词系统的性能和领域适应性。边界熵基于字左侧和右侧出现不同汉字的次数计算该字的条件概率而得。因为汉字左侧和右侧有不同的结果，故分为左边界熵和右边界熵。它的公式如下：

$$Le(W)=-\sum_{\forall a\in\{x|count(xW)\geq 1\}}P(a|W)\log(P(a|W))$$

$$Re(W)=-\sum_{\forall b\in\{x|count(Wx)\geq 1\}}P(b|W)\log(P(b|W))$$

公式中 Le 和 Re 分别表示左边界熵(Left Entropy)和右边界熵(Right Entropy)；W 表示字符串，$W=w_1w_2\ldots w_n$；$count(s)$表示候选字符串 s 在语料中出现次数。$\{x|count(xW)\geq 1\}$表示候选字符串 W 左侧出现所有字符组成的集合，其中 xW 表示候选字符串 W 和其左边出现的 x 结合构成的字符串。

$P(a|W)$表示在出现 W 的前提下，在 W 左侧出现字符 a 的条件概率。同样，$\{x|count(Wx)\geq 1\}$表示候选字符串 W 右侧出现所有字符组成的集合，其中 Wx 表示候选字符串 W 和其右边出现的 x 结合构成的字符串，$P(b|W)$表示在出现 W 的前提下，在 W 左侧出现字符 b 的条件概率。

根据边界熵理论，边界熵分别反映词语左右侧的不确定性。词语 W 的 Le 和 Re 的数值越大，W 就越有可能是一个完整的词语。由于我们采取字标注分词方法，在使用边界熵时，$W=w_1w_2...w_n$ 的 n 为 1。与之对应，W 由字符串变为字符，求词语边界熵变为求每个字符的边界熵。这样，当一个字符的 Le 值较大，说明这个字符极有可能是一个词的开端；同样，当一个词的 Re 值较大，说明这个字符极有可能是一个词的末尾。与卡方统计量类似，边界熵也需要规范化，规范化如下：

$$x_{norm}=\lfloor x\times 5\rfloor$$

实验中发现，两个熵值整体即“字－左边界熵－右边界熵”构成特征时，会带来负面影响，且远大于正面作用。因为，当字处于一个词开端时，它的左边界熵很高，而右边界熵不确定，这取决于这个字本身会接多少种字成词，可能有各种不同结果，此时右边界熵的用途远不如左边界熵；同理，当字是词末尾时，右边界熵作用更重要。现在将一个重要特征与不重要特征作为整体加以处理，导致本应体现出的特征被抹杀，并带来许多本不应该有的特征，扰乱结果。在实验中，我们进一步提出最小边界熵的概念，选取左边界熵和右边界熵的最小值作为该字的边界熵。即：

$$E(w)=\min\{LE(s),RE(s)\}$$

其中 w 为字，$LE(w)$为 w 的左边界熵；$RE(w)$为 w 的右边界熵。这样，特征定义模板为：

(l) $E_n(n=-2,-1,0,1,2)$ (m) $E_nE_{n+1}(n=-2,-1,0,1)$ (n) $E_{n-1}E_{n+1}(n=-1,0,1)$

这里 E 代表字的最小边界熵，n 代表当前字的相对位置。

3.2 卡方统计量的改进

卡方统计量特征是通过比对训练语料和测试语料的结果，寻求匹配的结果加以分词处理。它本身是基于字之间关系计算的，对各个领域适应度较高，有利于提升 OOV 词汇的召回率。文献[12]发现卡方统计量在提升 OOV 召回率时，也导致准确率和 IV 词汇召回率的降低，造成整体结果的下降。因为文献[12]的方法是将训练语料和测试语料分别计算卡方统计量的值，各自以其自身求得的结果作为卡方统计量特征进行分词。由于测试语料本身的规模小，IV 词汇计算得到的卡方统计量的值与测试语料计算的实际值有一定程度上的偏差，导致 IV 词汇的结果降低。

基于此，我们提出新的改进方法：计算卡方统计量时，将测试语料合并到训练语料一起计算得到新值，作为训练和测试语料特征。结果表明，不但 OOV 词召回率提升，其他各项也稍有提升。说明将测试语料合并入训练语料计算卡方统计量，可以避免给 IV 词汇带来的负面影响。

4 实验及结果讨论

实验采用人民日报 1998 年 1 月份全部内容[1]作为训练标记语料，使用 CRF[2] 训练。测试语料有 3 种，分别为新闻(相同领域)、小说和科技文。在以下结果列举中，我们使用：B 代表基本特征(a)(b)(c)；T 代表类型特征(d)(e)(f)；AV 代表 AV 统计量特征(g)(h)(i)；X 代表卡方统计量特征(j)(k)；X^+代表将测试语料合并到训练语料计算的卡方统计量特征(j)(k)；BE 代表边界熵特征(l)(m)(n)。

4.1 优化卡方统计量的比对

表 1 是测试语料是否合并到训练语料计算卡方统计量的跨领域分词对比结果。从结果看出，测试语料未合并到训练语料时，OOV 召回率明显上升，却带来 F 值下降，因为对 IV 词汇产生负

[1] http://icl.pku.edu.cn/icl_groups/corpus/dwldform1.asp

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

面影响。当采取测试语料加入训练语料计算的方法后，OOV 词召回率虽没大幅度提升，但各项都有正面作用，说明测试语料并入训练语料计算卡方统计量，对 IV、OOV 词汇具有正面作用。

表 1

文章领域	特征	Precision	Recall	F	IV Recall	OOV Recall
小说	*B*、*T*	0.912851	0.908529	0.910685	0.92498	0.652651
	B、*T*、*X*	0.91237	0.903214	0.907769	0.915724	**0.708638**
	B、*T*、X^+	0.915826	0.910958	**0.913386**	0.92739	0.655393
科技文	*B*、*T*	0.919212	0.888648	0.903672	0.924364	0.686122
	B、*T*、*X*	0.912178	0.897375	0.904716	0.937915	**0.701273**
	B、*T*、X^+	0.926301	0.899413	**0.912659**	0.938736	0.693115

4.2 卡方统计量对分词结果的影响

表 2 是对比卡方统计量跨领域分词的作用。为了进一步体现卡方统计量作用，我们在实验中使用了 AV 统计量做对比参照。从结果看出，卡方统计量的加入，对跨领域分词结果有提升。

表 2

文章领域	特征	Precision	Recall	F	IV Recall	OOV Recall
小说	*B*、*T*	0.912851	0.908529	0.910685	0.92498	0.652651
	B、*T*、X^+	0.915826	0.910958	**0.913386**	0.92739	0.655393
	B、*T*、*AV*	0.915083	0.910586	0.912829	0.92786	0.64191
	B、*T*、*AV*、X^+	0.916209	0.91053	**0.913361**	0.927199	0.65128
	B、*T*、*BE*	0.915968	0.910379	0.913164	0.927066	0.650823
	B、*T*、*BE*、X^+	0.921206	0.916259	**0.918726**	0.933766	0.643967
科技文	*B*、*T*	0.919212	0.888648	0.903672	0.924364	0.686122
	B、*T*、X^+	0.926301	0.899413	**0.912659**	0.938736	0.693115
	B、*T*、*AV*	0.920475	0.898609	0.90941	0.940257	0.680115
	B、*T*、*AV*、X^+	0.919916	0.900044	**0.909872**	0.94024	0.68917
	B、*T*、*BE*	0.917303	0.89492	0.905973	0.938445	0.666577
	B、*T*、*BE*、X^+	0.919386	0.909159	**0.914244**	0.944922	0.721535

4.3 与 AV 统计量特征的对比实验

AV 统计量是目前流行的分词特征，我们在实验中使用了 AV 统计量做对比参照(表 3)。从结果看出，边界熵特征其本身带有信息量比 AV 统计量大，可用性高，效果比 AV 统计量要好一些。

表 3

文章领域	特征	Precision	Recall	F	IV Recall	OOV Recall
小说	*B*、*T*、*AV*、X^+	0.916209	0.91053	0.913361	0.927199	0.65128
	B、*T*、*BE*、X^+	0.921206	0.916259	**0.918726**	0.933766	0.643967
科技文	*B*、*T*、*AV*、X^+	0.919916	0.900044	0.909872	0.94024	0.68917
	B、*T*、*BE*、X^+	0.919386	0.909159	**0.914244**	0.944922	0.721535
新闻	*B*、*T*、*AV*、X^+	0.958272	0.95774	0.958006	0.963994	0.711422
	B、*T*、*BE*、X^+	0.958521	0.957604	**0.958062**	0.963615	0.721256

4.4 最终结果

经过边界熵特征和卡方统计量的共同训练下，最终结果如表 4 所示。

表 4

文章领域	特征	Precision	Recall	F	IV Recall	OOV Recall
小说	B、T、BE、X^+	0.921206	0.916259	0.918726	0.933766	0.643967
科技文	B、T、BE、X^+	0.919386	0.909159	0.914244	0.944922	0.721535
新闻	B、T、BE、X^+	0.958521	0.957604	0.958062	0.963615	0.721256

5 结论及进一步的工作

从上述实验的结果来看，边界熵和卡方统计量在分词模型的建立上是可行的，它在跨领域的分词中起到了一定的效果。上述结果都是在训练模型下直接生成的结果，未进行任何后处理工作。相信后续处理如果再引入，结果还能够有一定的提升。

此外，卡方统计量在 OOV 词汇分词上有显著提升，边界熵在 IV 和 OOV 词汇上有平衡作用。两者可以尝试 co-training 方式，彼此互补，反复训练，或许可以进一步提升跨领域分词结果。

参 考 文 献

[1] 黄昌宁，赵海，2007，中文分词十年回顾，中文信息学报，第 21 卷第 3 期，8-20 页.

[2] Nianwen Xue, 2003, Chinese Word Segmentation as Character Tagging, Computational Linguistics and Chinese Language Processing, 8(1), pages 29-48.

[3] Tseng, Huihsin et al., 2005, A conditional random field word segmenter for SIGHAN Bakeoff 2005, Proceedings of the fourth SIGHAN workshop on Chinese language processing. Jeju Island, Korea, pages 168-171.

[4] Zhao, Hai et al., 2006, 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, pages 87-94.

[5] Low, Jin Kiat et al., 2005, A Maximum Entropy Approach to Chinese Word Segmentation, Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea, pages 161-164.

[6] 孙茂松等，基于无指导学习策略的无词表条件下的汉语自动分词。计算机学报第 27 卷第 6 期，736-742 页.

[7] Hai Zhao and Chunyu Kit, 2007, Incorporating global information into supervised learning for Chinese word segmentation, Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, pages 66-74.

[8] Huixing Jiang et al., 2010, An Double Hidden HMM and an CRF for Segmentation Tasks with Pinyin's Finals, Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing(CLP2010), pages 277-281.

[9] Xiaoming Xu, et al., 2010, High OOV-Recall Chinese Word Segmenter, Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP2010), pages 252-255.

[10] Ling-Xiang Tang, et al., 2010, A Boundary-Oriented Chinese Segmentation Method Using N-Gram Mutual Information, Proceedings of CIPS-SIGHAN Joint Conference on CLP2010, pages 234-239.

[11] Haodi Feng, et al., 2004. Accessor Variety Criteria for Chinese Word Extraction, Association for Computational Linguistics, 30(1), pages 75-93.

[12] Baobao Chang, Dongxu Han, 2010, Enhancing domain portability of Chinese segmentation model using chi-square statistics and bootstrapping, Proceedings of the 2010 Conference on EMNLP, pages 789-798.

[13] 任禾等，2006，一种基于信息熵的中文高频词抽取算法，中文信息学报，第 20 卷第 5 期，第 40-43 页.

[14] 姜柄圭，2006，面向中文专著的汉韩机器辅助翻译研究，北京大学博士论文.