基于词边界分类的中文分词方法

李寿山　黄居仁
香港理工大学 中文及双语学系　中国香港
E-mail: {shoushan.li, churenhuang}@gmail.com

摘　要：本文研究和探讨一种新的分词方法：基于词边界分类的方法。该方法直接对字符与字符之间的边界进行分类，判断其是否为两个词之间的边界，从而达到分词的目的。相对于目前主流的基于字标注的分词方法，该方法的实现和训练更加直接、简单和快速，但却能获得比较接近的分词效果。更重要的是，我们很容易地从词边界分类方法获得在线分词学习方法。该方法能够使我们的分词系统非常迅速的学习新的标注样本。

关键词：　中文分词，WBD 方法，在线学习

Chinese Word Segmentation Based on Word Boundary Decision

Li Shoushan and Huang Chu-Ren
The Hong Kong Polytechnic University, Department of Chinese & Bilingual Studies. Hong Kong.
E-mail: {shoushan.li, churenhuang}@gmail.com

Abstract: This paper focuses on the word boundary decision (WBD) approach to Chinese word segmentation. This new approach classifies a boundary between two characters into either a word boundary or not. Compared to the stat-of-the-arts approach of character-based tagging approach, this approach is easier to implement and faster to execute, with similar performances. More crucially, robust online learning module can be added to enable a WBD system to adapt to new data quickly and serve as a reliable online Chinese segmentation system without domain or training data constraints.
Keywords: Chinese word segmentation, WBD approach, online learning.

1　引言

自动分词是中文计算机处理中的一个基本任务[1]。该技术是实现众多中文应用系统的前期基础，例如：中英（英中）机器翻译、信息检索、文本自动分类等等。由于中文词边界本身的歧义性，加上大量新词的不断涌现，中文自动分词一直是中文信息处理中的一个长期且艰巨的任务。针对该任务，中文信息处理研究领域先后出现了大量的自动分词方法。

长期以来，分词的方法都是基于词（或词典）的[2]。例如基于规则的最大匹配方法[3]或者基于统计的词的 N 元语法方法[4]。在大规模词典的帮助下，基于词的分词方法取得了比较好的效果。但是该方法在识别未登录词（OOV）时，结果并不理想，然而未登录词的识别在分词系统的应用中是不可避免的问题。近几年来，基于字的分词方法渐渐得到相当的关注，在近几年的 Sighan 评测中占主导地位。该方法将分词问题转化为字的分类问题，即判断文本的字是否为某词的开始或者结束等等[5]。相对于基于词的分词方法，基于字的分词方法的最大优点就是能够取得令人满意的 OOV 识别率。尽管如此，由于基于字的分词方法需要构建高维的向量空间，而且分类的样本非常庞大，这使得该方法训练和测试的时间复杂度和空间复杂度都非常高。因此，该方法还不能满足一些真正实用的实时分词系统。我们认为一个优秀的实时分词系统不仅仅需要

快速对新样本进行分词，还需要能够迅速学习新的标注样本（该特性我们称之为在线学习能力[6]）。而目前基于字的分词方法训练需要时间太长，很难满足以上的要求。

文献[7]首先提出一种基于词边界分类的分词方法（Word Boundary Decision，以下简称该方法为 WBD）。该方法直接对分词的任务建模，判断字同字之间的边界是否为词的边界。换言之，该方法视分词为一个两类分类问题。相对于一般的基于字的分类方法，类别的数目有所减少，从而减低了分类的复杂度。本文将介绍该分词方法并进一步完善该方法，使得它能够获得接近目前主流方法的分词效果。在此基础上，我们将提出一种基于 WBD 的在线学习分词方法，能够快速学习新标注样本。鉴于 WBD 方法的时间复杂度和空间复杂度远远小于基于字的分词方法，我们能够利用它构建一个更实用的分词系统。

本文其他部分安排如下：第 2 节详细介绍基于词边界分类的分词方法并提出我们的一些改进；第 3 节提出基于 WBD 的在线学习分词方法；第 4 节给出实验结果及分析；第 5 节给出相关结论。

2 基于词边界分类的分词方法（WBD）

2.1 WBD 方法

自动分词的基本任务是将一段由字串组成的文本切割成由词组成的词序列。举例来说，一段分词前的字序列为“共同创造美好的新世纪”，而分词后的词序列为：“共同 创造 美好 的 新世纪”。这样做是便于计算机识别和处理比较完整的一些语义单位，如“创造”、“美好”等等，从而有利于进一步的文本处理，如自动翻译、信息检索等。

WBD 方法的目标是判断每两个字之间的边界是否为词边界。我们形式化表示一段文本为：

$$c_1 I_1 c_2 I_2, ..., c_i I_i, ..., c_{n-1} I_{n-1} c_n$$

其中 c_i 表示一个中文字符，I_i 表示任意两个字之间的字边界。在原始的中文文本中，这些字边界没有明显的显示出是否为词边界。我们设定如果该字边界为词边界，则记为 $I_i = 1$，否则 $I_i = 0$。WBD 的方法就是直接判断某个字边界是否为词边界。因此，在 WBD 中，分词任务被转化为一个两类分类问题。然而在基于字的分词方法中，一般会把字的类别数目定义为“开始”、“中间”、“结束”和“单字”。在类别数目上面来说，WBD 方法要比基于字的分词方法更简单。

WBD 方法大致分为两个步骤：N-gram 概率信息统计和边界的向量表示。下面分别介绍这两个步骤。

在第一个步骤中，该方法通过收集训练样本中的词边界信息，即 N-gram 字串关于词边界的统计信息。文献[7]中给出了 5 种不同的 unigram 和 bigram 特征的统计信息，它们分别为 P_{CB}、P_{BC}、P_{CCB}、P_{CBC} 和 P_{BCC}。这些统计信息具体是指 N-gram 字串相对于词边界的出现的概率。其中 P_{CB} 定义如下：

$$P_{CB}(I_i = 1 | c_i) = \frac{C(c_i, I_i = 1)}{C(c_i)}$$

其中 $C(c_i, I_i = 1)$ 表示在训练语料中字符 c_i 出现在词边界前面的次数。而 $C(c_i)$ 表示字符 c_i 在训练语料中出现的总次数。

另外，P_{CCB} 的定义如下

$$P_{CCB}(I_i=1|c_{i-1},c_i)=\frac{C(c_{i-1},c_i,I_i=1)}{C(c_{i-1},c_i)}$$

其中 $C(c_{i-1},c_i,I_i=1)$ 表示在训练语料中二元字符串 c_{i-1},c_i 出现在词边界前面的次数。而 $C(c_{i-1},c_i)$ 是二元字符串 c_{i-1},c_i 在训练语料中出现的总次数。

其他三个概率信息，$P_{BC}(I_i=1|c_{i+1})$，$P_{CBC}(I_i=1|c_i,c_{i+1})$和$P_{BCC}(I_i=1|c_{i-2},c_{i-1})$ 具有类似的定义。为了简单起见，下面我们分别用 $P_{CCB}(I_i)$，$P_{CBC}(I_i)$，$P_{BCC}(I_i)$，$P_{CB}(I_i)$，$P_{BC}(I_i)$ 表示这 5 个概率信息统计量。一旦获得了所有 N-gram 字符串的概率信息，我们将这些数据保存在一个字典中，供以后构建向量使用。我们命名这个包含统计信息的字典为“N-gram 数据源”。

在第二个步骤中，WBD 方法将所有的字边界表示为下面的向量：

$$<P_{CCB}(I_i), P_{CB}(I_i), P_{CBC}(I_i), P_{BC}(I_i), P_{BCC}(I_i)>$$

WBD 的训练过程和测试过程都会将所有字边界表示为这种向量。剩下的就是一个典型的模式分类问题。我们可以采用各种不同的统计分类方法去训练和测试。由于该向量的维度很低，仅仅需要几千个样本就足够训练出一个比较好的分类器[8]。这一点与基于字的分词方法不同，基于字的分词方法需要训练大量的样本向量。因此，WBD 方法大大降低了训练时间。为了更清晰的了解 WBD 方法，我们给出文本“共I_1同I_2创I_3造I_4美I_5好”中每个字边界的向量实例，如表 1 所示。其中，在N-gram 数据源中未出现的字串概率一律赋值为 0.5。

P_{CCB}	P_{CB}	P_{CBC}	P_{BCC}	P_{BC}	I_i	字边界
0.50	0.29	0.00	0.50	0.69	0	共同
0.94	0.41	0.43	0.99	0.90	1	同创
0.44	0.17	0.00	0.50	0.37	0	创造
0.50	0.57	0.50	0.99	0.85	1	造美
0.50	0.23	0.01	0.55	0.56	0	美好

表 1　一段字串的字边界向量实例

2.2 WBD 方法的进一步完善

在已有的 WBD 方法框架的基础上，我们提出一系列进一步完善的措施，包括概率估计的平滑、新 N-gram 的引入和数字英文字符的预处理。

在估计概率的时候，往往会出现某些字符在语料中出现次数很少的情况。利用上面的统计公式统计概率的时候会带来很大的误差。为了减小这些误差给分类带来的影响，我们尽量使频率出现很少的概率趋向于 0.5。以 P_{CB} 为例，我们改写概率估计函数如下：

$$P_{CB}(I_i=1|c_i)=\frac{C(c_i,I_i=1)+1}{C(c_i)+2}$$

另外，上面介绍的 WBD 方法仅仅考虑到了 unigram 和 bigram 字符串的统计信息。其实在实际分词任务中，为了捕捉更远的分词信息，从而更好的识别多字符的词，我们引入 trigram 字

符串的统计信息。trigram 字符串包括 *CCCB*、*CCBC*、*CBCC* 和 *BCCC*，对应的概率值分别为 $P_{CCCB}(I_i)$、$P_{CCBC}(I_i)$、$P_{CBCC}(I_i)$ 和 $P_{CCCB}(I_i)$。我们将在实验部分测试这些新特征对分词效果的影响。

为了进一步提高 WBD 方法的分词效果，我们在进行统计 N-gram 字符串概率之前利用正则表达式进行所有数字和英文字符的识别和替换。

3 基于 WBD 方法的在线学习分词方法

优秀的实时分词系统不仅仅需要快速对新样本进行分词，还需要能够迅速学习新的标注样本，只有这样，才能让系统能够快速引入不断涌现的新词。换言之，系统应该具备在线学习的能力。具体来讲，在线学习（online learning）是指系统在输入新的标注样本后，不需要重新学习原始的训练样本，只是针对新来的标注样本学习就可以了。目前，在线学习本身作为机器学习的一个重要的研究问题，备受模式识别，机器学习，自然语言处理等学术届的重视。

需要强调的是 WBD 方法的核心部分是 N-gram 数据源，如果加入新的标注样本，我们只需要更新 N-gram 数据源。以其中一个概率 $P_{CCB}(I_i)$ 为例，更新的概率公式如下：

$$P'_{CCB}(I_i=1\,|\,c_{i-1},c_i)=\frac{C(c_{i-1},c_i,I_i=1)+C_{new}(c_{i-1},c_i,I_i=1)+1}{C(c_{i-1},c_i)+C_{new}(c_{i-1},c_i)+2}$$

其中 $C(c_{i-1},c_i,I_i=1)$ 和 $C(c_{i-1},c_i)$ 是 N-gram 数据源里面字符串 $c_{i-1}c_i$ 的词频信息。这些数据都是已经有的，不需要重新学习。$C_{new}(c_{i-1},c_i,I_i=1)$ 和 $C_{new}(c_{i-1},c_i)$ 则是新标注样本里面字串 $c_{i-1}c_i$ 的统计信息。具体来讲，$C_{new}(c_{i-1},c_i,I=1)$ 表示在新训练语料中二元字符串 c_{i-1},c_i 出现在词边界前面的次数，而 $C_{new}(c_{i-1},c_i)$ 是二元字符串 c_{i-1},c_i 在新训练语料中出现的总次数。其他几种 N-gram 的更新公式同上式类似。在实际应用中，如果待测样本和新标注样本比较接近的话，可以通过提高新样本里面统计词频的权重来加重新标注样本对后期分词的影响。

除了更新 N-gram 数据源之外，训练阶段不需要任何其他的操作。因此，基于 WBD 的在线学习方法，需要再学习的代价非常小，足以满足实时系统中的分词需要。

4 实验

在本实验中，我们将首先详细给出 WBD 的分词效果，其次，我们将测试上面提出的基于 WBD 的在线学习方法。

我们使用第二届国际分词竞赛（Bakeoff-2005）中的四组语料对 WBD 分词方法进行测试。每组语料中的训练语料用于生成各自的 N-gram 数据源，并随机生成训练样本中的 1000 个字边界的向量训练一个 SVM（Support Vector Machine）分类器。在测试过程中，测试语料中的所有字边界由该组的 N-gram 数据源生成向量。然后用训练好的 SVM 分类器进行测试。

我们使用基于词的 F 值作为评估标准，它是准确率 P 和召回率 R 的调和平均值：$F=2RP/(R+P)$。为了和相关工作比较，我们也列出了未登录词的召回率（OOV recall）。表 2-表 5 分别给出了 WBD 方法及其一些改进方法在四组测试语料上面的分词结果。其中，基准的实现过程完全按照文献[7]里面的描述，只用了 5 个 N-gram 的字符串特征。在改进的方法中，我们只报告了 *CCBC* 和 *CBCC* 的结果，这是因为 *CCCB* 和 *BCCC* 字串的加入对结果基本没有影响。从这些表中可以看出，平滑在两个语料上都有了好的表现，在另外两个语料中基本保持原来的分词效果。其他两个改进在所有的语料中都有不同程度上的提高。

Pku 语料	Precision	Recall	F1-score	OOV recall
基准	0.905	0.870	0.888	0.370
平滑	0.908	0.880	0.895	0.382
CCBC，CBCC	0.917	0.911	0.914	0.440
数字英文字符	**0.939**	**0.923**	**0.931**	**0.690**

表 2　WBD 及其改进方法在 Pku 语料测试语料上的结果

Cityu 语料	Precision	Recall	F1-score	OOV recall
基准	0.904	0.915	0.910	0.512
平滑	0.895	0.922	0.908	0.500
CCBC，CBCC	0.913	0.922	0.917	0.540
数字英文字符	**0.915**	**0.932**	**0.924**	**0.575**

表 3　WBD 及其改进方法在 Cityu 语料测试语料上的结果

Msr 语料	Precision	Recall	F1-score	OOV recall
基准	0.933	0.933	0.933	0.526
平滑	0.925	0.940	0.932	0.467
CCBC，CBCC	0.937	0.960	0.949	0.418
数字英文字符	**0.940**	**0.960**	**0.950**	**0.479**

表 4　WBD 及其改进方法在 Msr 语料测试语料上的结果

As 语料	Precision	Recall	F1-score	OOV recall
基准	0.901	0.930	0.919	0.483
平滑	0.912	0.932	0.922	0.504
CCBC，CBCC	0.914	0.942	0.928	0.475
数字英文字符	**0.926**	**0.946**	**0.936**	**0.541**

表 5　WBD 及其改进方法在 As 语料测试语料上的结果

为了便于比较 WBD 同基于字的分词方法，我们利用 Porket CRF [9]工具实现了基于字的分词方法，实验使用了字符的四种标识类别[10]。表 6 中给出该方法在 Pku 和 Msr 两个语料上面的封闭测试结果（没有用到数字英文字符识别）。相对于 CRF 实现的基于字的分词方法，虽然 WBD 的分词效果要稍微差一点，但是 WBD 所需要的训练时间（包括 N-gram 字符串概率的搜集和 SVM 分类器的训练）要远远小于 CRF 的训练时间（我们用 python 实现的 WBD 方法）。

在实际分词系统的应用中，训练样本和测试样本并不能像分词竞赛那样限制为来自相同来源。为了模拟实际分词无法预知测试语料来源的状况，我们交换了 Pku 和 Msr 的测试

语料。也就是说，我们让 CRF 和 WBD 在 Pku 的训练语料上面训练，但是在 Msr 的测试语料上面测试，这样分词的结果会大大减低到 0.856 和 0.850（Pku→Msr）。在这个时候，CRF 和 WBD 方法的差异已经小到没有统计上的相关性了。

	Pku	Msr	Pku（训练）→ Msr（测试）	Msr（训练）→ Pku（测试）	训练时间
CRF	0.914	0.962	0.856	0.850	大于 1 小时
WBD	0.931	0.950	0.850	0.851	小于 2 分钟

表 6　WBD 与 CRF 实现的基于字标注的分词方法的比较结果

最后，我们利用 WBD 方法构建一个真正实用的分词系统，该系统使用 Sinica 研究院平衡语料产生 N-gram 数据源。该语料是来自台湾研究员的繁体语料，由于 Sinica 语料在各个领域的分布比较均匀，我们认为该语料库能够比较好的反映各种字符串同词边界的关系[11]。为了测试我们系统的效果以及上面提到的基于 WBD 的在线学习方法，我们使用部分 Cityu 的训练样本用作我们系统新的标注样本。图 1 给出了我们系统在 Cityu 测试语料上面的 F 值结果。横坐标表示加入新样本的规模。从图中可以看出，在不使用任何 Cityu 的训练样本的情况下，我们的系统已经取得了近 88%的 F 值。如果利用我们的在线学习方法，渐渐融入少量 Cityu 的训练样本，系统的表现会越来越好。

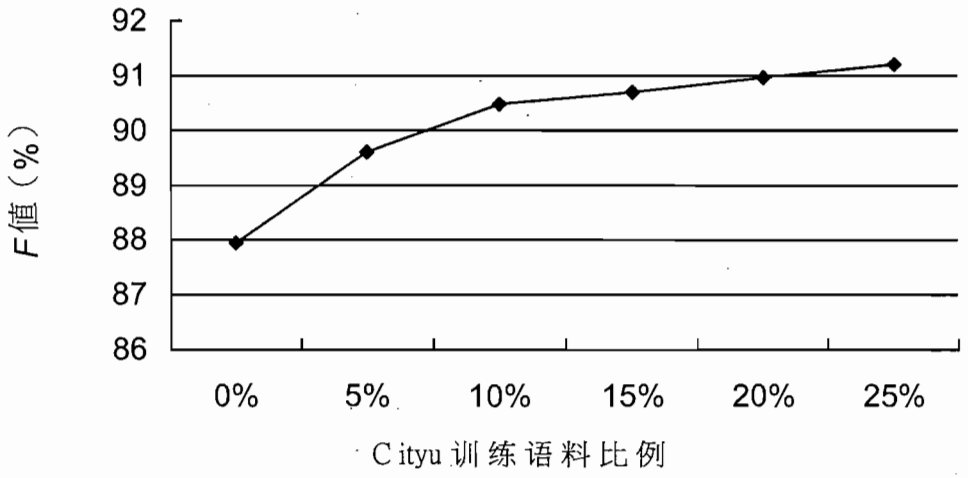

图 1　系统在 Cityu 测试语料上面的 F 值

5 结论

本文研究了一种基于词边界分类的分词方法并提出改进方法。在此基础上，我们实现了一种在线学习的分词方法。实验结果表明，这种新的分词方法能够获得接近目前主流方法的分词效果，但只需要很少的训练时间。同时，我们利用提出的在线学习分词方法和 Sinica 研究院平衡语料库构建了我们的分词系统，该系统能够在来自不同地方的语料中获得比较满意的分词效

果，并且能够很迅速的学习新的样本，使我们的系统具备很好的更新能力。

参 考 文 献

[1] 黄昌宁. 中文信息处理的分词问题. 语言文字应用, 1997, (1):72-78.

[2] 黄昌宁, 赵海. 中文分词十年回顾. 中文信息学报, 2007, 21(3):8-20.

[3] 骆正清, 陈增武, 胡上序. 一种改进的 MM 分词方法的算法设计. 中文信息学报, 1996, 30-36.

[4] 吴春颖, 王士同. 基于二元语法的 N-最大概率中文粗分模型. 计算机应用, 2007, 27(12): 332-339.

[5] Xue N. and Shen L. Chinese word segmentation as LMR tagging. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*. 2003.

[6] Crammer K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, vol.7, 2006. 551–585.

[7] Huang C., P. Šimon, S. Hsieh, and L. Prevot. Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Wordbreak Identification. In *Proceedings of the Association of Computational Linguistics Annual Meeting (ACL)*. 2007.

[8] Huang C., Yo T., P. Šimon, and S. Hsieh. A Realistic and Robust Model for Chinese Word Segmentation. In *Proceedings of ROCLING*. 2008.

[9] http://sourceforge.net/project/showfiles.php?group_id=201943.

[10] Ng H. and Low J. Chinese part-of-speech tagging: one-at-a-time or all-at-once? Word-based or character-based? In *Proceedings of EMNLP*. 2004.

[11] CKIP. Academia Sinica Balanced Corpus of Modern Chinese. http://www.sinica.edu.tw/SinicaCorpus/. 2001.