统计与词典相结合的领域自适应中文分词*

张梅山，邓知龙，车万翔，刘 挺
哈尔滨工业大学 信息检索研究中心，哈尔滨 150001
E-mail: {mszhang, zldeng, car, tliu}@ir.hit.edu.cn

摘 要: 基于统计的中文分词方法往往不具有良好的领域自适应性。本文通过将外部词典信息融入统计分词模型（本文使用 CRF 统计模型）来实现领域自适应性。实验表明，这种方法具有良好的领域自适应性。当测试领域和训练领域相同时，分词的 F-measure 值提升了 2%；当测试领域和训练领域不同时，分词的 F-measure 值提升了 6%。最终优化后的分词速度也得到了很大的改善。

关键词: 中文分词；CRF；领域自适应

Combining Statistical Model and Dictionary for Domain Adaption of Chinese Word Segmentation

Zhang Meishan, Deng Zhilong, Che Wanxiang, Liu Ting
Center for Information Retrieval of Computer Science & Technology School, Harbin Institute of Technology, Harbin 150001
E-mail: {mszhang, zldeng, car, tliu}@ir.hit.edu.cn

Abstract: Generally, statistical methods for Chinese Word Segmentation don't have good domation adaption. We propose an approach which can integrate extern dictionary information into statistical models to realize domain adaption for Chinese Word Segmentation. In the paper, we use the CRF statistical model. Experimental results show that our approach have good domain adaption. When domain of test corpus is identical to the training corpus, the F-measure value has an increase of 2%; when domain of test corpus is different with the training corpus, the F-measure value has an increase of 6%. The final speed of segmentation has also been improved greatly after optimized.

Keywords: Chinese word segmentation; CRF; domain adaption

1 引言

中文分词是中文自然语言处理中最基本的一个步骤，非常多的研究者对它做了很深入的研究，也因此产生了很多不同的分词方法，这些方法大体上可以分为两类：基于词典匹配的方法和基于统计的方法。

基于词典的方法[1]利用词典作为主要的资源，这类方法不需要考虑领域自适应性的问题，它只需要有相关领域的高质量词典即可，但是这类方法不能很好的解决中文分词所面临的歧义性问题以及未登录词问题。

基于统计的方法[2][3][4][5]是近年来主流的分词方法，它采用已经切分好的分词语料作为主要的资源，最终形成一个统计模型来进行分词解码。基于统计的方法在分词性能方面有了很大的提高，但是在跨领域方面都存在着很大的不足，它们需要针对不同的领域训练不同的统计分词模型。这样导致在领域变换后，必须为它们提供相应领域的分词训练语料，但是分词训练语料的获得是需要大量人工参与的，代价昂贵。而基于词典的方法却在领域自适应方面存在着一定优势，当目标分词领域改变时，只需要利用相应领域的外部词典即可。外部词典的获取相比训练语料而言要容易很多。如果把这两种方法结合起来，使得统计的方法能够合理应用外部词典，则可实现中文分词的领域自适应性。

* 本文承国家自然科学基金（60803093；60975055），哈尔滨工业大学科研创新基金（HIT.NSRIF.2009069）和中央高效基本科研业务费专项资金（HIT.KLOF.2010064）的资助。

赵海等人(2007)[6]以及张碧娟等人(2008)[7]都曾提出将外部词典信息融入统计分词模型中大大改善了分词的性能。但是他们实际上都始终把词典当做一种内部资源，训练和解码都使用同样的外部词典信息，并没有解决中文分词的领域自适应性问题。本文借鉴在 CRF 模型中融入外部词典的方法来解决中文分词的领域自适应性问题。在训练 CRF 分词模型时，使用通用的外部词典；而分词阶段通过额外再加入领域词典来实现领域自适应性。当分词领域改变时，只需要加载相应领域的外部词典，而且不需要改变原有已经训练得到的统计中文分词模型，就可以大大改善该领域的分词准确率。

最后本文利用 SIGHAN CWS BAKEOFF 2005 中提供的 PKU corpora 进行训练，训练过程中采用通用的外部词典，训练得到的统计分词模型分别在 PKU test corpus 和人工标注的金融领域语料上进行了测试。测试时，PKU 语料所用的词典保持与训练语料所用的词典一致，而金融领域所用的词典则额外再加入了部分金融领域的专业词汇。最后的结果显示，PKU 语料上取得了 2%的 F-measure 提升；金融领域上取得了 6%的 F-measure 提升，最终达到 93.4%。

本文组织内容为：第二部分介绍 CRF 中文分词原理；第三部分介绍领域自适应性的实现；第四部分为实验部分；第五部分为结论及进一步工作。

2 CRF 中文分词原理

薛念文[0]等人 2003 年提出将中文分词问题看成序列标注问题。句子中每个字根据它在词中的位置进行分类，共分为 B，M，E，S 四类。其中 B 代表该字符是每个词的开始，M 表示该字符在某个词的中间位置，E 表示该字符是某个词的结束位置而 S 表示该字符能独立的构成一个词。

CRF 是目前主流的序列标注算法，它在序列标注问题上取得了很大的成功。对于给定的句子 $\mathbf{x}=c_1\cdots c_n$ 及其某个分词标注结果为 $\mathbf{y}=y_1\cdots y_n$，其中 c_i 为输入字符，$y_i\in(B,M,E,S)(1\leqslant i\leqslant n)$，我们可以用如下的方法表示 $\mathbf{y}$ 的概率：

$$P_\lambda(\mathbf{y}\mid\mathbf{x})=\frac{1}{z(\mathrm{x})}\exp(\boldsymbol{\lambda}\cdot\sum\nolimits_{i=1}^{n}\Phi(y_{i-1},y_i,\mathbf{x})) \tag{1}$$

其中 $z(\mathrm{x})$ 为一个归一化因子，$\Phi(y_{i-1},y_i,\mathbf{x})$ 为特征向量函数，$\boldsymbol{\lambda}$ 为特征权重向量。

对于 CRF 模型，特征的选择尤为重要。本文首先使用了三类基本特征：字符 n-gram 特征，字符重复信息特征和字符类别特征。这三类特征和论文 Tseng(2005)[3]中提到的特征很类似，这里对他们的字形态特征做了一定的扩展，将输入字符分为九类：Single，Prefix，Suffix，Long，Punc，Digit，Chinese-Digit，Letter 以及 Misc。下表 1 是对它们的定义以及相应的示例：

任何一个输入字符只可能属于这些类别中的一类。其中 Punc、Digit、Chinese-Digit、Letter 可以直接通过其属性来直接判断一个字符是否属于该类别；而判断一个字符是否属于 Single、Prefix、

表 1　字符类别定义以及示例

字符类别	属性	示例
Single	通常单独是一个词	的、呢
Prefix	通常作为词语的开始	他、惆
Suffix	通常作为词语结束	式、所
Long	通常构成长词	帕、穆
Punc	标点符号	，、。
Digit	数字	1、2
Chinese-Digit	汉字数字	一、二
Letter	字母	A、b
Other	其他	行、练

Prefix 或 Long，通过统计该字符在外部词典中满足这些类别属性的比例来判断，阈值设为 95%；如果一个字符属于多个类别，那么按照 Punc、Digit、Chinese-Digit、Letter、Single、Prefix、Prefix、Long、Other 这个优先次序来判定，越靠前的优先级别越高。

最后这里列举一下在我们的 CRF 中文分词模型中所使用的基本特征，如表 2 所示：

表 2　CRF 中文分词模型中所使用的基本特征

特征类别	具体特征形式化描述
字符 n-gram 特征	$c_i (i=-2,-1,0,1,2)$
	$c_i c_{i+1} (i=-2,-1,0,1)$
	$c_i c_{i+2} (i=-1,0)$
字符重复信息特征	$Reduplication\,(c_0,c_i)\,(i=-2,-1)$
字符类别特征	$Type\,(c_i)\,(i=-1,0,1)$
	$Type\,(c_{-1})Type(c_{-1})Type(c_{-1})$

其中下标代表考虑的相对位置，$Reduplication\,(c_0,c_i)$ 表示 c_0 和 c_i 是否为两个完全一样的字符，$Type(c_i)$ 表示字符 c_i 的类别。

3　领域自适应性的实现

外部词典对中文分词有着很大的用处，而且外部词典的获得所需要的代价远远小于为相关领域标注分词语料所需要的代价。如果统计中文分词方法能充分合理的利用外部词典，一方面可以提高中文分词的准确率，另一方面还可以使中文分词具有良好的领域自适应性。当为特定领域进行中文分词时，只需要加载该领域的专属词典，便可以很好的解决该领域的中文分词问题。整个系统框架如图 1 所示。

当领域改变之后，原有的 CRF 分词模型是不需要再改变的，只需改变领域词典即可，因此不需要针对不同领域重新去训练不同的分词模型。

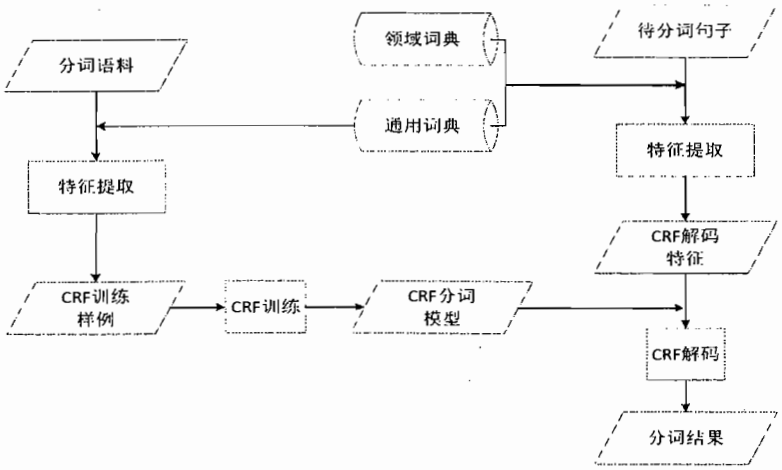

图 1　领域自适应性分词系统框架图

下面介绍外部词典特征的融入方法，给定句子 $\mathbf{x}=c_1\cdots c_n$，以及外部词典 $\mathbf{D}$，考虑其中的第 j 个字符 $c_j (1\leqslant j\leqslant n)$，定义如下三个函数：

$$f_B(\mathbf{x},j,\mathbf{D})=\max l,\ s.t.\begin{cases}\mathbf{w}=c_j\cdots c_{j+l-1}\in\mathbf{D}\\ j+l-1\leqslant n\end{cases}$$

$$f_M(\mathbf{x},j,\mathbf{D})=\max l,\ s.t.\begin{cases}\mathbf{w}=c_s\cdots c_{s+l-1}\in\mathbf{D}\\ j<s+l-1\leqslant n\\ 1\leqslant s<j\end{cases}\qquad(2)$$

$$f_E(\mathbf{x},j,\mathbf{D})=\max l,\ s.t.\begin{cases}\mathbf{w}=c_{j-l+1}\cdots c_j\in\mathbf{D}\\ 1\leqslant j-l-1\end{cases}$$

其中 **w** 表示词语；$f_B(\mathbf{x},j,\mathbf{D})$ 表示对于句子 **x** 在 j 位置根据词典 **D** 采用正向最大匹配所获得的词的长度；$f_M(\mathbf{x},j,\mathbf{D})$ 表示对于句子 **x** 在 j 前面的某个位置根据词典 **D** 采用正向最大匹配所获得的经过 j 位置而且不以 结尾的最长词的长度；$f_E(\mathbf{x},j,\mathbf{D})$ 表示对于句子 **x** 在 j 位置根据词典 **D** 采用逆向最大匹配所获得的词的长度。

本文对 CRF 分词模型所引入的与外部词典 D 相关的扩展特征如表 3 所示：

表3　CRF 中文分词模型中所使用的外部词典特征

外部词典特征	具体特征形式化描述
uni-gram 特征	$[f_B]_i(i=-1,0,1)$
	$[f_M]_i(i=-1,0,1)$
	$[f_E]_i(i=-1,0,1)$
tri-gram 特征	$[f_B]_{-1}[f_B]_0[f_B]_1$
	$[f_M]_{-1}[f_M]_0[f_M]_1$
	$[f_E]_{-1}[f_E]_0[f_E]_1$

假设目前考虑位置为 ,则上面相应的 $[f_B]_i=f_B(\mathbf{x},j+i,\mathbf{D})$, $[f_M]_i=f_M(\mathbf{x},j+i,\mathbf{D})$, $[f_E]_i=f_E(\mathbf{x},j+i,\mathbf{D})$.

4　实验

本文利用 SIGHAN CWS BAKEOFF 2005 中提供的 PKU 训练语料进行训练，训练过程中使用北京大学中国语言学研究中心公开的词典[1]，该词典一共包含大约 10 万多个词。最后分别在相应的 PKU 测试语料和人工标注的金融领域语料上进行了评测，表 2 给出了两个测试语料的统计信息。本文使用准确率(P)、召回率(R)和 F-measure 值(F)来评价分词系统。本文采用 CRF++工具包[2]来进行训练和标注。

表4　测试语料相关统计信息

测试语料	句子数目	词语数目
PKU 测试语料	1 944	104 372
金融测试语料	51 763	1 326 711

4.1　实验结果及分析

CRF-basic 代表仅使用基本特征训练出来的模型；CRF-post 表示使用拼接的后处理方法去纠正被 CRF 错误切分的词，这个方法假定外部词典中没有在训练语料中出现的词都应该是不可切分的；CRF-extern 表示融入了词典信息特征之后所得到模型。

[1] http://ccl.pku.edu.cn/doubtfire/Course/Chinese%20Information%20Processing/Source_Code/Chapter_8/Lexicon_full_2000.zip

[2] http://chasen.org/~taku/software/CRF++/

在 PKU 的测试语料上，使用训练出来的模型，测试时所使用的词典和训练时所使用的词典一致。表 3 给出了 PKU 语料上测试的结果。从表中可以看出，CRF-extern 与 CRF-basic 相比，F-measure 提升了 1.8%；与 CRF-post 相比，提升了 0.3%。

在金融领域的测试语料上，保持训练出来的 CRF 分词模型不变，使用的词典是在训练语料的外部词典基础上增加了 1 000 个左右的金融领域专用词典。表 4 给出了金融领域测试语料上的结果。从表中可以看出，CRF-extern 与 CRF-basic 相比，F-measure 提升了 7.6%；与 CRF-post 相比，提升了 3.2%。

表 5　SIGHAN BAKEOFF 2005 PKU 测试语料上分词性能比较

分词方法	P	R	F
CRF-basic	94.3%	95.4%	94.8%
CRF-post	96.4%	96.2%	96.3%
CRF-extern	**96.6%**	**96.6%**	**96.6%**

表 6　金融领域测试语料上分词性能比较

分词方法	P	R	F
CRF-basic	84.0%	89.7%	86.8%
CRF-post	88.8%	91.8%	90.2%
CRF-extern	**93.3%**	**93.5%**	**93.4%**

从上面的两个实验可以看出，

a) 无论是测试语料的领域与训练语料领域是否相同，CRF-extern 对比 CRF-basic 显著提高了分词的性能。

b) 当训练语料和测试语料领域相同时，CRF-extern 和 CRF-post 相比，分词性能有稍微的提高；但是当领域不同时，CRF-extern 对比 CRF-post 而言，有了非常显著的提高。

c) 测试领域和训练语料不同时，最终的分词 F-measure 值达到了 93.4%，已经非常接近于 CRF-basic 在不考虑跨领域时的 F-measure 值 94.8%。

综上所述，在统计模型中融入词典信息特征后，一方面分词性能有了一定的提高；另外一方面领域迁移后，分词性能依然能够保持在一定的水平。因此统计模型与词典结合后，使得中文分词具有良好的领域自适应性。

5　结论及下一步工作

本文通过在 CRF 统计分词模型中融入外部词典特征来实现中文分词的领域自适应性。当面向不同的领域时，只需通过加载相应领域的词典。因为领域词典的获取与为该领域标注分词训练语料相比代价要小很多。最终实验结果表明，该方法不仅仅在原有领域上取得了比较好的效果，而且在金融领域上也取得了不错的效果。

下一步我们需要自动挖掘各种领域相关的词，从而使得我们的分词系统能适应各个领域的需求。

参 考 文 献

[1] Keh-Jiann Chen and Shing-Huan Liu. 1992. Word identification for mandarin chinese sentences. In Proceedings of the 14th conference on Computational linguistics, pages 101-107, Morristown, NJ, USA. Association for Computational Linguistics.

[2] Nianwen Xue. 2003. Chinese word segmentation as character tagging. International Journal of Computational Linguistics and Chinese Language Processing, 8(1).

[3] Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In Proceedings of the fourth SIGHAN workshop, pages 168-171.

[4] Yue Zhang and Stephen Clark. 2007. Chinese segmentation with a word-based perceptron algorithm. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 840-847, Prague, Czech Republic, June. Association for Computational Linguistics.

[5] Xu Sun, Yaozhong Zhang, Takuya Matsuzaki, Yoshimasa Tsuruoka, and Jun'ichi Tsujii. 2009. A discriminative latent variable chinese segmenter with hybrid word/character information. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 56-64, Boulder, Colorado, June. Association for Computational Linguistics.

[6] Hai Zhao; Chang-Ning Huang; Mu Li. An Improved Chinese Word Segmentation System with Conditional Random Field. In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 162-165.

[7] Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. Optimizing Chinese Word Segmentation for Machine Translation Performance. In ACL 2008 Third Workshop on Statistical Machine Translation.