ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.9, September 2009, pp.2366−2375
doi: 10.3724/SP.J.1001.2009.03606

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种基于字词联合解码的中文分词方法*

宋 彦[1+], 蔡东风[1], 张桂平[1], 赵 海[2]

[1](沈阳航空工业学院 知识工程中心,辽宁 沈阳 110034)

[2](香港城市大学 中文、翻译及语言学系,香港)

Approach to Chinese Word Segmentation Based on Character-Word Joint Decoding

SONG Yan[1+], CAI Dong-Feng[1], ZHANG Gui-Ping[1], ZHAO Hai[2]

[1](Knowledge Engineering Center, Shenyang Institute of Aeronautical Engineering, Shenyang 110034, China)

[2](Department of Chinese, Translation and Linguistics, City University of Hong Kong, Hong Kong, China)

+ Corresponding author: E-mail: mattsure@gmail.com

Song Y, Cai DF, Zhang GP, Zhao H. Approach to Chinese word segmentation based on character-word joint decoding. *Journal of Software*, 2009,20(9):2366−2375. http://www.jos.org.cn/1000-9825/3606.htm

Abstract: The performance of Chinese word segmentation has been greatly improved by character-based approaches in recent years. With the help of powerful machine learning strategies, the words extraction via combination of characters becomes the focus in Chinese word segmentation researches. In spite of the outstanding capability of discovering out-of-vocabulary words, the character-based approaches are not as good as word-based approaches in in-vocabulary words segmentation with some internal and external information of the words lost. In this paper we propose a joint decoding strategy that combines the character-based conditional random field model and word-based Bi-gram language model, for segmenting Chinese character sequences. The experimental results demonstrate the good performance of our approach, and prove that two sub models are well integrated as the joint model of character and word could more effectively enhance the performance of Chinese word segmentation systems than any of the single model, thus is fit for many applications in Chinese information processing.

Key words: Chinese word segmentation; joint decoding; language model; conditional random field model

摘 要: 近年来基于字的方法极大地提高了中文分词的性能,借助于优秀的学习算法,由字构词逐渐成为中文分词的主要技术路线.然而,基于字的方法虽然在发现未登录词方面有其优势,却往往在针对表内词的切分效果方面不及基于词的方法,而且还损失了一些词与词之间的信息以及词本身的信息.在此基础上,提出了一种结合基于字的条件随机场模型与基于词的 Bi-gram 语言模型的切分策略,实现了字词联合解码的中文分词方法,较好地发挥了两个模型的长处,能够有效地改善单一模型的性能,并在 SIGHAN Bakeoff3 的评测集上得到了验证,充分说明了合理的字词结合方法将有效地提高分词系统的性能,可以更好地应用于中文信息处理的各个方面.

关键词: 中文分词;联合解码;语言模型;条件随机场模型

* Supported by the National Natural Science Foundation of China under Grant No.60842005 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z148 (国家高技术研究发展计划(863)); the Key Project of the Ministry of Education of China under Grant. No.207148 (国家教育部科学技术研究重点项目)

Received 2008-09-12; Accepted 2009-03-05

中图法分类号: TP301 文献标识码: A

中文分词作为中文信息处理的基本环节,近年来重新得到了广泛的关注.随着近几届 SIGHAN 国际中文分词评测(Chinese Word Segmentation Bakeoff)[1−3]的举行,中文分词领域的研究取得了许多令人振奋的成果.自第 1 届 SIGHAN 分词评测中 Xue 采用基于字标注的学习方法[4]以来,使用由字分词的思想在后来的分词评测及相关研究中得到了较大的发展并结合不同的机器学习方法得到了集中使用,同时也取得了超越以往传统方法的分词效果,由字构词一度成为中文分词领域的焦点[5,6].同一时期,完全基于词的分词方法几乎难觅踪影,不难理解,因为中分分词面临的一个主要难题就是未登录词(out-of-vocabulary word,简称 OOV)识别,一般,基于词的方法都是使用了某种指定的词表,不在词表之列的未登录词显然无法进行处理.相对而言,基于字的分词方法的合理性在于,它将词表词和未登录词同等看待,同时使用字这种更小的单元进行标注,而且字集合的规模比词相对要小很多,使用基于字的机器学习方法往往可以更准确地进行决策,因此,这种使用字组合的分词过程能够得到较好的分词效果.但是,相对于字而言,词的信息往往十分重要,一些词所包含的组合信息在基于字的方法中可能被忽略,Zhang 的基于 Subword 的方法[7]和 Zhao 提出的基于子串标注的分词方法[8]尝试解决这个问题,也得到了比单一字标注方法更好的效果,尽管如此,词和词之间的转换关系(不同于前一个词的尾字和后一个词的首字的转换关系)仍然无法在这些方法中得以描述.另外,针对词表词(in-vocabulary word,简称 IV)的处理,基于字的方法也往往得不到比基于词的方法更优的结果,而且通常会因为字和字的组合造成一些莫名其妙的切分错误.因为字和词的方法各有特点,有必要将字和词的方法结合起来,实现效果更好的中文分词.

本文尝试了一种字词结合的分词思想,通过采用联合解码在组合字和词的模型方面进行了一定程度的探索,充分发挥了 Bi-gram 语言模型切分词表词的作用以及条件随机场(conditional random fields,简称 CRF)模型发现未登录词的作用,并在 SIGHAN Bakeoff3[3]的四组测试项目上验证了本文方法的有效性.

1 由字构词的一些问题

由字构词方法虽然可以平等地对待词表词和未登录词,并且得到了较高的 R_{OOV},但是正因为如此,其对于词表词的切分效果反而不如基于词的方法,Zhang 的工作在 SIGHAN Bakeoff2[2]的测试集上很好地证实了这一点[7].

表 1 中使用词的方法得到的词表词召回率(R_{iv})普遍高于使用字的方法,只不过因为在这些测试集上,未登录词造成的精度损失比其他因素大很多[6],才使得基于字的方法在最终的结果上占据了较大的优势(这里是指全局 F 值,按照 $F=2PR/(P+R)$计算,其中 P 是全局准确率,R 是全局召回率).而且,针对一个语料库中的 IV 部分,切分歧义将成为其中主要的问题,基于字的方法相比基于词的方法可能并不能更好地解决这些歧义,因此需要在 IV 和 OOV 部分区别对待,才能在一定程度上提高全局召回率,即基于词的方法针对词表词的召回率以及基于字的方法在未登录词上的召回率.

Table 1 Recall rate of IV and OOV by different segmentation approaches

表 1 不同切分方法得到的 IV 和 OOV 召回率

Corpus	OOV rate	R_{IV}		R_{OOV}	
		Word based approach	Character based approach	Word based approach	Character based approach
AS	0.043	**0.982**	0.967	0.038	0.647
CityU	0.074	**0.989**	0.967	0.164	0.736
MSRA	0.026	**0.993**	0.972	0.048	0.716
PKU	0.058	**0.981**	0.955	0.408	0.754

另外,对具体的分词错误进行分析可以发现,基于字的方法得到的分词结果总是倾向于将两个或者多个词合成为一个词,例如在 SIGHAN Bakeoff3 的 MSRA 测试语料上,不论采用 4 标注集还是 6 标注集的 CRF 模型,总会发生如下的切分错误:

台/港/澳/同胞 → 台/港澳同胞 (前面为正确结果,后面为错误结果,下同)

我国/驻/纽约/总/领事馆 → 我国/驻/纽约总领事馆

环境保护/中心 → 环境保护中心

不可否认,如果不参考标准结果,按照主观的理解,有些切分结果亦可以认为是正确的.但是因为测试集的标准结果是按照训练语料的切分形式进行处理的,从训练数据得到的模型应该可以很好地反映这种切分情况.现在普遍采用的条件随机场模型基本上可以认为是按照某种模式进行字串的标注,倘若遇到训练语料中没有出现的字串(即未登录词),只能遵循训练时得到的标注模式,因此受训练语料的颗粒度影响较大,如发生上述错误的 MSRA 语料颗粒度较大,切分结果往往会产生一些多词粘连的情况,影响了最终的分词性能.针对这类情况,词信息往往可以提供较好的切分指导,不至于使多个词切分到一起.

2 基于字词联合解码的中文分词

本文融合字词信息的基本思路是将它们的切分词分别作为候选结果加入到一个统一的概率空间下,通过求解联合概率空间最大概率下的切分路径,得到最终的切分结果.倘若将各个字信息概率模型表示为 P_{ci},词信息概率模型表示为 P_{wj},那么联合概率空间可以表示为如下形式:

$$P=\prod_i P_{ci}\prod_j P_{wj} \tag{1}$$

各个字模型和词模型通过乘积形式组合,从数学上来说,将两种完全不同的概率模型融合到一起是借用了对数线性模型(最大熵模型[10,11])的原理:将不同的概率模型作为特征函数形式累积在同一个数学框架下,共同作用于最终的决策.通常其中的特征函数包含权重系数,可以调整各个子模型对整个模型的贡献.这里本文只是针对字词模型组合的最简单形式,没有为每个模型指定权重系数,因此式(1)可以认为是一种具有最简单形式的对数线性模型,如若输出结果为 y_{out},那么针对式(1)的求解可以表示为

$$y_{out}=\arg\max_y P=\arg\max_y \prod_i P_{ci}\prod_j P_{wj} \tag{2}$$

对于一个输入结果,通常需要计算它对于所有可能输出 y 的条件概率,并从中选择最大概率对应的结果.这个过程通常称为解码,因此,我们将上述联合概率空间的求解称为字词联合解码.

在实际使用时,一般采用式(1)的对数形式:

$$\log P=\sum_i \log P_{ci}+\sum_j \log P_{wj} \tag{3}$$

这时,字和词的概率模型按照线性加权组合的方式结合起来,式(2)可以转化为

$$y_{out}=\arg\max_y\left\{\sum_i \log P_{ci}+\sum_j \log P_{wj}\right\} \tag{4}$$

其中,字和词的概率子模型可以采用完全不同的建模方法,对于某个给定的输入,整个系统按照所有概率模型的共同作用选择最终的输出信息.

针对上述联合解码作为融合字模型与词模型的集成方法,本文分别使用了一个基于字的模型和一个基于词的模型作为子模型.其中,基于字的模型是使用 6 个标记的条件随机场(CRF)标注器,基于词的模型使用的是二元语法(bi-gram)语言模型.因此求解过程可以描述为

$$y_{out}=\arg\max_y\{\log P_{CRF}+\log P_{LM}\} \tag{5}$$

下面分别对这里所使用的各个模型及融合方法进行阐述.

2.1 基于字的6标记CRF标注器

条件随机场模型进入自然语言处理领域即得到了成功的应用[12−14].近年来,尤其在中文分词领域,自从 SIGHAN Bakeoff2 以后,几乎所有取得较好名次的基于字的分词方法都采用了条件随机场作为学习工具,其在 OOV 的召回率上通常可以取得较好的效果.

为了便于比较,本文使用了基于6个标记的条件随机场模型[9]作为由字构词的标注模型,标注集见表2,采用了6种对应不同字位的标记.

Table 2 Tag set used in the CRF tagger

表 2 CRF 标注器使用的标记集

ID	Tag	Description
1	B1	The first character position in a multi-character word
2	B2	The second character position in a *n*-character word ($n\geq3$)
3	B3	The third character position in a *n*-character word ($n\geq4$)
4	I	The position in a *n*-character word from the fourth character to the $n-1$th character ($n\geq5$)
5	E	The last character position in a multi-character word
6	S	Single-character word

除此之外,为了有效地配合上述标记集,我们还使用了8个对应的特征模板,见表3.

Table 3 Feature templates used in the CRF tagger

表 3 CRF 标注器使用的特征模板

ID	Template	Description
1	$C_{-2},C_{-1},C_0,C_1,C_2$	Unigram character feature
2	$C_{-2}C_{-1},C_{-1}C_0,C_0C_1,C_1C_2$	Bigram characters feature
3	$C_{-1}C_0C_1$	Trigram characters feature
4	$C_{-1}C_1$	Jump characters feature
5	$T(C_{-1}),T(C_0),T(C_1)$	Unigram character type feature: punctuation – *P*, English letter – *E*, numbers – *D*, Chinese or other characters – *C*
6	$T(C_{-1})T(C_0),T(C_0)T(C_1)$	Bigram characters feature
7	$T(C_{-1})T(C_1)$	Jump characters feature
8	$C_0T(C_0)$	Combined feature using current character and its type

其中,模板1~模板4为基本的字符特征,仅涉及分词字面的信息.这里需要指出的是,模板1、2中定义的字符特征相当于使用了5个字的上下文“窗口”,加上我们使用的6字位的标记集,因为其上下文窗口宽度的增加,该模型可以得到比文献[15]所描述的采用6字位标记集与3字上下文得到的模型更好的性能.另外,本文还引入了字符的类别特征,以帮助条件随机场模型学习到不同类型的字符对分词结果的影响,它们反映在模板5~模板7中,我们使用的4种类别信息如表中所述.严格地说,采用类别标记相当于引入了一个外部词典,因此不能认为训练数据是完全从语料库中获得的,即并非严格的封闭测试,考虑到现有的基于字的方法基本上都采用了这种类型的特征,为了更好地与以往的实验结果进行比较,本文仍然在封闭测试中引入这类特征.最后,还使用了一个字符及其类别的混合特征(模板8).本文采用Taku Kudo编写的CRF++0.50(http://crfpp.sourceforge.net/#download)实现训练和标注过程.

仅采用上述标记集和模板的条件随机场的模型就已经可以实现一个较好性能的中文分词系统,这是目前许多相关工作都采取的技术路线,这里将其作为一个子模型加入到最终的联合解码框架下.

2.2 基于词的Bi-gram语言模型

在分词过程中引入词的信息可以有很多种方法,使用词典就是一个很好的例子.但是,考虑到我们的方法还需要能够处理切分歧义,词和词之间的关系也应该在这个模型中体现出来.因此,本文采用N-gram语言模型来反映训练语料中的词信息,N-gram语言模型不但可以很好地反映词和词之间的转移关系,同时也包含了词典(所有Uni-gram的集合),较符合本文分词任务的要求.为了限制计算量的大小以及数据稀疏带来的影响,在本文中我们仅使用Bi-gram语言模型.实际上,Bi-gram模型对于分词任务而言已经足够,采用该模型的方法在SIGHAN Bakeoff2的测试结果中取得了较好的成绩[16].

本文在实验中采用SRI Language Modeling Toolkit (SRILM)(http://www.speech.sri.com/projects/srilm/download.html)进行语言模型的训练,得到的语言模型使用标准的ARPA格式,其中包括N-gram串及其对应的概率和回退系数,便于合理地计算某两个词之间的转移概率.在训练语言模型时使用的是Good-Turing方法进行概率平滑,因此在计算Bi-gram概率时,我们使用Katz方法[17]进行回退,对应到Bi-gram的情况,可以表述为

$$p(w_n \mid w_{n-1}) = \begin{cases} p(w_n \mid w_{n-1}), & \text{if } C(w_{n-1}w_n) > 0 \\ \alpha(w_{n-1})\,p(w_n), & \text{otherwise} \end{cases} \tag{6}$$

其中$\alpha(w_{n-1})$是(w_{n-1})的回退系数.采用这样的计算方法(回退方法)是为了更好地表达某些在训练语料中尚未出现的 Bi-gram 串(其中不包含 OOV)在测试语料中的概率.

同样地,单独使用语言模型也可以构造一个分词系统,但是往往需要后处理以弥补其 OOV 发现能力的不足[16],因此,需要结合字标注模型以发挥其各自的优点.

2.3 算法描述

虽然已经指定使用的具体模型,但是如何有效地将它们融合到联合解码的框架下仍然需要仔细设计,由于字词两个模型的形式差别较大,如何合理地选取候选切分词对最终的结果影响很大.特别是,要想实现扬长避短,发挥 CRF 标注器发现 OOV 的能力和 Bi-gram 语言模型解决歧义及较高的 IV 召回能力,同时又不致使 Bi-gram 较差的 OOV 发现能力影响 CRF 标注器的结果.因此,不同模型在整个过程中的作用应该是均等的,因此,联合解码过程中 CRF 标注器和 Bi-gram 语言模型对整个概率空间具有同样的贡献是有意义的,这样有助于突出 CRF 发现的 OOV 以及 Bi-gram 语言模型切分的 IV.另外,对于 CRF 标注器的标注词,不能按照语言模型的定义给予其一个近似的"零概率",因此需要解决它与前后词的概率转移.我们将这个概率指定为语言模型中最低的 IV Bi-gram 概率,这实际上认可了其中包含的 OOV 作为一个潜在的切分结果,而且赋予转移概率以后可以将所有标注词(对于语言模型而言,其中包括 IV 与 OOV)与语言模型中的 IV 等同对待,我们认为,这对于语言模型和 CRF 标注器而言都是公平的.

基于上述原则,整个分词过程可以认为是一个寻找最大概率切分的过程,本文使用柱搜索(beam search)实现这个过程.假设存在一个字串 S_1^n,算法具体流程可以描述为:

(1) 根据语言模型中的 Uni-gram 寻找 S_1^n 所有可能的切分词,再依据 CRF 标注得到 S_1^n 的切分结果,取出所有的切分词,并从这两组切分结果中得到原字串上的所有候选切分词;

(2) 建立 n+1 个栈[0,1,2,…,n],用于存放不同长度的切分片段(其中第 1 个栈为起始空栈,不放入信息),同时,指定栈空间的大小用于剪枝,本文使用的栈空间为 100;

(3) 将步骤 1 得到的候选词在 S_1^n 上自前向后以当前位置为指针在步骤 2 建立的栈上扩展,每次扩展时根据联合解码模型为可能的切分词按照转移概率进行打分,并记录当前最后一个词与扩展词的打分结果并与之前得到的总分进行累积(IV 词之间使用其语言模型概率,OOV 与 IV 或者 OOV 与 OOV 之间按照前面提到的原则累积转移概率),显然,当一个词在语言模型和 CRF 中同时被切分出来,其具有较大的可能是正确的切分结果,因为采用联合解码模型的关系,其打分值也将得到倍增;或者,如果一个切分结果仅通过语言模型得到,那么它将损失 CRF 模型那部分的打分值;

(4) 每次扩展结果按照已覆盖的原 S_1^n 长度(字符个数)压栈(例如,如果已经得到 2 个 2 字词的切分结果,那么在其结果上进行扩展时,再得到一个 2 字词时应压入第 6 个栈),压栈时,如栈中的记录已满,则需要按照目前为止的积分进行剪枝,将当前记录与栈中最小积分的记录进行比较,保留积分较大的记录入栈,同时记录其扩展来源,便于回溯;

(5) 在 S_1^n 上自前向后移动指针并重复步骤 3、步骤 4,直至切分结果已经覆盖 S_1^n 所有字符;

(6) 从最后一个栈中取出一个(1-Best)或者多个(N-Best)积分最大的记录,往前回溯,生成整个 S_1^n 上的切分结果.

不难发现,采用上述算法,CRF 的标注序列实际上已经通过不同的扩展记录存在于整个栈空间的回溯路径中,最后均以累积的转移概率来判断采用何种切分结果.数学上,相当于把 CRF 的切分词线性插值到原语言模型中,其有效性用后面的实验进行分析.

3 实验结果及分析

本文使用 SIGHAN Bakeoff3(2006)的测试集——繁体中文的 CityU(City University of Hong Kong)语料和 CKIP(Chinese Knowledge Information Processing Laboratory)语料,以及简体中文的 MSRA(Microsoft Research Asia)语料和 UPUC(University of Pennsylvania and University of Colorado)语料——对我们的方法进行评价.这里的评价方法沿用 Bakeoff 的封闭测试规则,即在某组语料的测试中,只允许从该组训练语料中获取的知识指导其测试语料的分词,不允许采用任何其他资源.对应地,Bakeoff 的开放测试则可以使用任何可用的方法和数据进行分词.考虑到需要评估分词算法本身的有效性,因此 Bakeoff 的封闭测试形式更符合我们的要求,而且也便于与相关工作进行比较.

具体地,这几组测试集语料的相关信息[3],见表 4.

Table 4 Bakeoff corpus statistics

表 4 评测语料的信息

Corpus	Encoding	Training data size**	Test data size	OOV rate
CityU	BIG5HKSCS	1.6M	220K	0.040
CKIP	BIG5	5.5M	91K	0.042
MSRA	GB18030	1.3M	100K	0.034
UPUC	GB	509K	155K	0.088

在利用这几组训练语料生成语言模型时,考虑到概率值的可靠性以及数据稀疏的影响,我们取 Cutoff=1,相应地,通过 SRILM 得到的语言模型规模见表 5.

Table 5 Language model statistics from training corpus

表 5 从训练语料中得到的语言模型规模

Corpus	Uni-Grams	Bi-Grams
CityU	75 597	728 251
CKIP	145 957	1 716 803
MSRA	62 634	547 094
UPUC	37 384	253 752

我们的方法和其他方法的结果(F 值与 R_{OOV})列于表 6,其中,包括 SIGHAN Bakeoff3 的官方底线成绩***和顶线成绩****,当届 Bakeoff 各个参赛队提交的封闭测试最佳成绩,以及 Zhao 在同样语料上得到的目前最佳研究结果[18],这样可以方便地对我们的算法进行最全面的分析.

Table 6 Experimental results(F-Score/R_{IV}/R_{OOV})

表 6 实验结果比较(F-Score/R_{IV}/R_{OOV})

	CityU	CKIP	MSRA	UPUC
Baseline	0.906/0.969/0.009	0.892/0.954/0.030	0.924/0.981/0.022	0.828/0.951/0.011
Topline	0.984/0.981/0.993	0.983/0.979/0.997	0.992/0.991/0.999	0.968/0.958/0.989
Bakeoff Best	0.972/0.981/0.787	0.958/0.972/0.702	0.963/0.976/0.612	0.933/0.963/0.707
Zhao*****	0.975/-------/0.801	0.959/-------/0.694	0.966/-------/0.662	0.943/-------/0.761
Joint Decoding	0.973/0.981/0.805	0.957/0.972/0.694	0.964/0.971/0.663	0.939/0.963/0.721

相比之下,联合解码的方法表现出较优越的性能,在与 SIGHAN Bakeoff3 的结果比较中,CityU,MSRA 和 UPUC 这 3 个语料上的 F 值达到了最好.由于本文联合解码的模型均采用监督学习(supervised learning)的模型合成,而 Zhao 的结果采用了非监督学习(unsupervised learning)方法辅助其 CRF 建模,在此仅将其结果列为参考.在 R_{OOV} 方面,即使考虑 Zhao 的结果,在除 UPUC 之外的 3 个语料上我们的 R_{OOV} 均取得了最高成绩.而且,在 4 个语料的测试中,我们的方法表现出了足够的稳定性,在不同的语料上都能达到几乎最好的效果.同时,语言模

** 此处训练语料和后面的测试语料大小均按照词数量统计.

*** 采用训练语料中抽取的词典进行的最大匹配分词结果.

**** 采用训练语料和测试语料的标准切分结果中共同抽取的词典进行的最大匹配分词结果.

***** Zhao 的结果来源于文献[18],因其中并未列出 R_{IV} 的结果,所以本文也没有给出.

型的确较好地处理了 CRF 的 OOV,在保证 IV 效果的前提下,改善了相当部分 OOV 的结果,使每个测试集上的 R_{OOV} 都有了一定程度的提升.具体地,我们的方法得到的全局准确率和召回率及其与 Bakeoff3 的封闭测试前三名成绩对比见表 7～表 10.

Table 7 Global precision, recall and *F*-score on CityU corpus, compared to top 3 scores of Bakeoff3

表 7 CityU 语料的全局准确率和召回率及 *F* 值与 Bakeoff3 前三名的比较

	P	*R*	*F*
Our approach	**0.974**	**0.973**	**0.973**
Bakeoff3 1st	0.972	**0.973**	0.972
Bakeoff3 2nd	0.972	**0.973**	0.972
Bakeoff3 3rd	0.971	0.972	0.971

Table 8 Global precision, recall and *F*-score on CKIP corpus, compared to top 3 scores of Bakeoff3

表 8 CKIP 语料的全局准确率和召回率及 *F* 值与 Bakeoff3 前三名的比较

	P	*R*	*F*
Our approach	**0.956**	0.958	0.957
Bakeoff3 1st	0.955	**0.961**	**0.958**
Bakeoff3 2nd	0.953	**0.961**	0.957
Bakeoff3 3rd	0.952	**0.961**	0.957

Table 9 Global precision, recall and *F*-score on MSRA corpus, compared to top 3 scores of Bakeoff3

表 9 MSRA 语料的全局准确率和召回率及 *F* 值与 Bakeoff3 前三名的比较

	P	*R*	*F*
Our approach	**0.963**	**0.965**	**0.964**
Bakeoff3 1st	0.961	0.964	0.963
Bakeoff3 2nd	0.953	0.961	0.957
Bakeoff3 3rd	0.955	0.959	0.957

Table 10 Global precision, recall and *F*-score on UPUC corpus, compared to top 3 scores of Bakeoff3

表 10 UPUC 语料的全局准确率和召回率及 *F* 值与 Bakeoff3 前三名的比较

	P	*R*	*F*
Our approach	**0.935**	**0.944**	**0.939**
Bakeoff3 1st	0.926	0.940	0.933
Bakeoff3 2nd	0.923	0.936	0.930
Bakeoff3 3rd	0.914	0.940	0.927

本文的方法在 4 个测试集上得到的全局准确率和召回率都比较稳定,只有在 UPUC 语料上的准确率和召回率有较大偏差.分析其原因,可能是由于 UPUC 语料测试集与训练集的不太一致造成的.文献[3]也指出,因为 UPUC 的训练语料完全来源于 CTB(Chinese TreeBank,中文树库),其测试集还包括了一部分其他新闻和广播内容,因此它具有比其他几个测试集高出 2 倍以上的 OOV 比率******,而且,Bakeoff3 当届测试结果普遍偏低也说明了这个问题.由于较多 OOV 的存在,联合解码过程中语言模型相应地也作了较多的回退计算,因此可能损失了一部分较好的切分路径,相当于使用了大部分的 CRF 切分结果.实际上,虽然本文的方法在 UPUC 语料上的切分准确率相比 Bakeoff3 的最好成绩已有提高,但是相比于召回率,仍然还可以有较大的提升空间,在其他几个语料上的结果也说明了切分准确率还值得改进.

从有效联合不同的概率模型角度分析,因为 OOV 的插入对语言模型预测准确率的影响较大,很多情况下,切分概率可能都回退到 Uni-gram 的概率.然而,考虑到组合模型分别来自于两种不同的学习方法,它们有不同的特点,因此这种组合也具有一定程度上投票的合理性.从具体的切分结果出发,可以发现,对于一个中文字串,当语言模型与 CRF 具有相同的切分结果时,系统会倾向于使用这样的结果,按照前面的描述不难理解,因为这样的切分串在两个模型中都成立,显然这样的结果比其他结果具有更高的打分值,因此,基于语言模型的切分结果被保留下来.当出现 OOV 时,按照语言模型的处理会将其切为单字成词的序列,而这个时候如果 CRF 标注器能把

****** UPUC 语料的 OOV 比率为 0.088(见表 4),其余测试集的 OOV 比率均远远小于 UPUC 的 OOV 比率.

这个 OOV 识别出来,则其作为一个整体的概率比 3 个单字的组合概率显然要更大,最终保留了 CRF 发现的 OOV.为了更好地说明这种组合方法的优势,表 11 中列出了分别仅采用语言模型(LM)和 CRF 模型的切分结果与联合解码模型的性能比较,足以说明组合方法的显著性.

Table 11 Comparison between different single models and joint decoding on segmentation performance

表 11 使用单一模型与联合解码模型的分词性能比较(全局 *P*/*R*/*F*-Score)

	CityU	CKIP	MSRA	UPUC
LM	0.897/0.952/0.924	0.881/0.943/0.911	0.907/0.960/0.932	0.804/0.898/0.849
CRF	0.968/0.968/0.968	0.952/0.956/0.954	0.957/0.958/0.957	0.931/0.936/0.933
Joint decoding	**0.974/0.973/0.973**	**0.956/0.958/0.957**	**0.963/0.965/0.964**	**0.935/0.944/0.939**

表 11 很好地体现了联合解码相比单个模型的优势,一方面可以认为,对于 CRF 的切分结果,语言模型较好地帮助解决了一部分 IV 的歧义问题和一部分 OOV 的处理;另一方面,CRF 所发现的 OOV 较好地补充了语言模型 OOV 发现能力不足的弱点,因此,我们可以得出在同源语料上得到的字和词模型的结合可以改善单一方法的结论.在此之外,我们还测试了基于文献[18]中 Zhao 提到的使用非监督学习方法作为特征的 CRF 模型与语言模型结合的效果,结果同样得到了有效提升,可见,该方法使用的技术路线是合理的.

另外,当本文接近完成的时候,我们注意到最新发表的文献[19]也使用了集成多种模型的思想进行中文分词的研究,并采用了级联的线性加权模型作为集成方法,就这个方面而言,本文的方法类似于其技术路线.但文献[19]的研究焦点是一种感知机模型以及在其基础上多模型分词结合词性标注的总体性能,其实现分词的仍然是单一模型,而且尽管其核心感知机模型使用的是基于字的方法,并且在级联模型中也包含了语言模型,但其并非针对字词结合研究分词方法的提升.同时,虽然其感知机算法在 Bakeoff2 语料上达到了较为可观的分词性能,但在 CTB 语料的实验中,我们注意到在使用其他模型的情况下,级联模型中的语言模型对分词性能几乎没有影响,对此文献[19]认为,感知机使用的字特征一定程度上“扮演了低阶语言模型的角色”,这实际上是在该语料上所选择的字特征功能与语言模型发生了重叠,我们认为 CTB 语料较小的颗粒度可能也促使了这种情况的发生.然而,如同本文前面描述的,字和字的转换关系与词和词的转换关系应该具有完全不同的语言学信息,即使感知机或者 CRF 模型对 OOV 的识别可以挽救相当程度的召回率,但单纯针对 IV 的切分效果仍然不及基于词的方法就是很好的证明.实验结果说明,本文使用的 CRF 特征和 Bi-gram 语言模型可以有效地组合,而且无须其他更复杂的模型支持也可以较大地提升基于单一方法的分词性能,而且关键在于,该方法有助于提升系统的稳定性,在不同的语料环境下都可以很好地发挥各个子模型的长处.

4 结 论

虽然基于字的机器学习方法已经成为当前中文分词的主要手段,但是合理的词信息引入可以得到更好的分词结果.本文提出了一种使用联合解码模型的中文分词方法,结合使用语言模型和 CRF 标注器,可以很好地发挥基于词的方法处理词表词的能力以及基于字的方法发现未登录词的能力,并且可以利用语言模型对未登录词作进一步的调整,相比已有结果,同时在整体性能或未登录词召回率上均有提升.实验结果证明了使用字词结合的方法可以得到更稳定的分词性能,在不同语料上都可以达到最佳或与其相当的成绩,显然更符合中文分词的要求.

然而,我们针对具体的分词结果的分析也发现了一些问题,首先,对于语料中 CRF 和语言模型都无法正确切分的问题,集成系统显然也无能为力;其次,针对一部分词语,两个模型也带来了一定的相反效果,即原来单个模型可以正确切分的字串,使用集成系统以后却产生切分错误,我们认为,这是因为集成模型在“改正”很多错误的同时也“改错”了另一部分内容,该部分的损失换来了全局结果的提升.因此在下一步工作中,在改进集成方法的同时也需要考虑解决一些子模型均无法解决的问题,同时,我们将尝试调整各个子模型的权重,仔细研究不同子模型对最终结果的影响.

需要特别提到的是,在仔细比较本文的分词结果与标准结果的过程中,我们发现,对于一些与标准答案不同的“错误”切分结果,我们更倾向于认为这些“错误”的切分结果是正确的,尽管这里存在标准答案出现问题的可

能,但针对很多正常的错误,我们也认为其切分结果相比标准答案更合适.实际上这种现象很多,目前通常使用基于SIGHAN Bakeoff形式的学习和比较法评价分词,因此唯一的标准答案决定了评价本身的效果,然而这个过程需要弹性,即在很多情况下,多种切分结果应该都是符合标准的,更重要的是这种优劣往往取决于分词之后的应用.当然,不可否认的是,Bakeoff 的方法采用了简便、易行的策略,同时也避免了分词标准规范之争,大大推动了分词技术的发展.但是,要将优秀的方法应用于实际,仍然还有很多工作要做,目前的很多方法都是在封闭测试的情况下考察结果贴近标准答案的能力,本文也不例外.因此,如果将其中较好的算法发展到实用环境下,特别是能够针对不同需求情况下的分词,将是一件非常有意义的工作.从这个角度而言,我们在将来研究新的分词技术或者集成算法时,其实用性也应当成为一个重要的评价标准.

致谢 在本文工作的研究过程中,我们与中国资深计算语言学专家董振东教授进行了多次颇具启发性的交流,受益良多,尤其得益于董教授对分词和应用问题的观点,特此表示诚挚的谢意.在本文的撰写过程中,香港城市大学的揭春雨教授也给予了作者很多有益的帮助和建议,在此一并鸣谢.

References:

[1] Sproat R, Emerson T. The 1st Int'l Chinese Word Segmentation Bakeoff. In: Proc. of the 2nd SIGHAN Workshop on Chinese Language Processing. 2003. http://www.aclweb.org/anthology-new/W/W03/W03-1719.pdf

[2] Emerson T. The 2nd Int'l Chinese Word Segmentation Bakeoff. In: Proc. of the 4th SIGHAN Workshop on Chinese Language Processing. 2005. http://www.aclweb.org/anthology-new/I/I05/I05-3017.pdf

[3] Levow G. The 3rd Int'l Chinese Language Proc. Bakeoff: Word segmentation and name entity recognition. In: Proc. of the 5th SIGHAN Workshop on Chinese Language Proc. 2006.

[4] Xue N, Shen L. Chinese word segmentation as LMR tagging. In: Proc. of the 2nd SIGHAN Workshop on Chinese Language Proc. 2003. http://www.aclweb.org/anthology-new/W/W03/W03-1728.pdf

[5] Huang C, Zhao H. Which is essential for chinese word segmentation: Character versus word. In: Proc. of the 20th Pacific Asia Conf. on Language, Information and Computation (PACLIC-20). 2006. 1−12.

[6] Huang C, Zhao H. Chinese word segmentation: A decade review. Journal of Chinese Information Processing, 2007,21(3):8−18 (in Chinese with English abstract).

[7] Zhang R, Kikui G, Sumita E. Subword-Based tagging by conditional random fields for Chinese word segmentation. In: Proc. of the HLT/NAACL-2006. 2006.

[8] Zhao H, Kit C. Effective subsequence-based tagging for chinese word segmentation. Journal of Chinese Information Processing, 2007,21(5):8−13 (in Chinese with English abstract).

[9] Zhao H, Huang C, Li M, Lu B. Effective tag set selection in Chinese word segmentation via conditional random field modeling. In: Proc. of the 20th Pacific Asia Conf. on Language, Information and Computation (PACLIC-20). 2006. 87−94.

[10] Berger A, Pietra SAD, Pietra VJD. A maximum entropy approach to natural language processing. Computational Linguistics, 1996, 22:39−71.

[11] Ratnaparkhi A. A maximum entropy model for part-of-speech tagging. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. 1996. http://www.aclweb.org/anthology-new/W/W96/W96-0213.pdf

[12] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the 18th Int'l Conf. on Machine Learning (ICML 2001). 2001. http://www.cis.upenn.edu/~pereira/papers/crf.pdf

[13] Sha F, Pereira F. Shallow parsing with conditional random fields. In: Proc. of the HLT-NAACL 2003. 2003. http://www.aclweb.org/anthology-new/N/N03/N03-1028.pdf

[14] Peng FC, Feng FF, McCallum A. Chinese segmentation and new word detection using conditional random fields. In: Proc. of the 20th Int'l Conf. on Computational Linguisticsd. 2004. http://www.aclweb.org/anthology-new/C/C04/C04-1081.pdf

[15] Zhao H, Huang C, Li M. An improved chinese word segmentation system with conditional random field. In: Proc. of the 5th SIGHAN Workshop on Chinese Language Processing. 2006. 162−165.

[16] Zhang H, Liu T, Ma J, Liao X. Chinese word segmentation with multiple postprocessors in HIT-IRLab. In: Proc. of the 4th SIGHAN Workshop on Chinese Language Processing. 2005. http://www.aclweb.org/anthology-new/I/I05/I05-3028.pdf

[17] Katz SM. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Trans. on Acoustics, Speech, and Signal Processing, 1987,35(3):400−401.

[18] Zhao H, Kit C. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In: Proc. of the 6th SIGHAN Workshop on Chinese Language Processing (SIGHAN-6). 2008. http://www.aclweb.org/anthology-new/I/I08/I08-4017.pdf

[19] Jiang W, Huang L, Liu Q, Lü Y. A cascaded linear model for joint Chinese word segmentation and part-of-speech tagging. In: Proc. of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08). 2008.

附中文参考文献:

[6] 黄昌宁,赵海.中文分词十年回顾,中文信息学报,2007,21(3):8−18.

[8] 赵海,揭春雨.基于有效子串标注的中文分词.中文信息学报,2007,21(5):8−13.

宋彦(1981－),男,湖南长沙人,硕士,主要研究领域为自然语言处理,机器翻译.

张桂平(1962－),女,博士,教授,主要研究领域为自然语言处理,机器翻译,知识管理.

蔡东风(1958－),男,博士,教授,主要研究领域为自然语言处理,人工智能,信息检索.

赵海(1976－),男,博士,研究员,主要研究领域为自然语言处理,机器学习.