利用单字碎片过滤改进汉语分词性能

李双龙[1,2]　刘乐中[3]　刘群[1]

[1]中科院计算所，北京 100080；[2]北京科技大学，北京 100083；[3].慕尼黑大学，慕尼黑

摘　要： 对于一般的分词系统，由于数据稀疏而识别失败的未登录词往往被切分成单字串。这里将切分结果中连续的单字串称之为“单字碎片”。本文提出的一种碎片过滤方法的基本思想就是重新检测出单字碎片中识别失败的未登录词，并将此方法作为“后处理”引入到一个原有的基于统计方法的分词系统（ICTCLAS）中。在第一届 SIGHAN 北大测试语料上测试，新系统未登录词召回率提高了 4%，F 值比原系统提高了 1%。可以看出，利用这种过滤方法在一定程度上削弱了数据稀疏问题，从而提高了汉语分词的性能。

关键词： 汉语分词；单字碎片；过滤；未登录词

Using a Fragment Filter to Improve Chinese Word Segmentation

Li Shuanglong[1,2]　Liu Lezhong[3]　Liu Qun[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080; [2]Beijing University of Science and Technology, Beijing 100083; [3]Munich University, Munich

Abstract: Generally, the unknown words are always segmented to be single characters separated by some mark when the segmenter analyses them unsuccessfully. These consecutive single characters here is called “character fragment”. The basic thinking of the fragment filter method in this paper is to detect the valid unknown words which are segmented falsely in the basic segmentation in the fragments. This method is combined to the origin segmenter(ICTCLAS) as a posting process. Finally, segmentation F-measure can be improved by 1% and OOV recall 4% in PK test data of the 1st SIGHAN bakeoff. We can see that the sparse

基金资助：863 课题“中文平台评价体系研究与基础数据库建设”(2004AA114010)，863 课题“中文信息处理与人机交互技术的测评系统和体系”(2003AA111010)

作者简介：李双龙(1981-)，男，山西人，中科院计算所客座研究生，Email:shuanglongli@eyou.com

data problem of statistical model can be weakened by this mechanism.

Keywords: Chinese Word Segmentation; Fragment; Filter; Unknown Word

1 引言

汉语分词是进行其他汉语语言处理之前非常关键的一项任务。由于汉语自身的特殊性，单字之间没有间隔，当我们对一个句子进行切词以后，独立的词语之间就被我们用特定的分隔符隔开，在本文中我们使用"/"。

对于任何一个已经被正确切分的汉语句子来说，必然会包含许多独立的单字词。比如：

广州/人海/花海/报/春/来/

家乡/的/酒/，/这/家乡/的/夜/，/怎/不/让/人/陶醉/呢/

在这里，我们先假设在切分结果中非单字的词语都是正确的切分。上面有下划线的连续单字序列在这里就被我们称之为"单字碎片"。

由于数据稀疏问题，一些识别失败的未登录词被切分成单字串；比如，利用我们的实验系统 ICTCLAS 就切分出了以下错误的结果：

他/却/被/短/接/发出/的/电弧/重重地/打倒/在/地上/。

露天/雪/浴/则/别/具/挑战性/。

姆拉迪奇/和/什/利/万/查/宁/是/南联盟/军队/的/将领/。

而这里提出的过滤方法的就是从这些单字串中重新检测出未登录词。比如从"/雪/浴/则/别/具/"中识别出新词"雪浴"。

实验的原型系统——ICTCLAS 是基于多层隐马模型的系统，具体细节可以参看(Huaping Zhang, et al 2003)。

2 前人的工作

在过去的相关工作中，Chooi-Ling GOH 曾建立一个基于最大熵的后处理模型，通过裁剪错误的未登录词来提高分词的性能；AnDi Wu 和 Zixin Jiang 在一个基于规则的汉语分词系统中，利用汉字的独立词概率(IWP)及相关信息进行新词的识别；Hongqiao Li 等用 SVM 来进行新词识别，利用了汉字在词中的不同位置的概率，新、旧词类推关系等特征（Hongqiao Li et al.,2004）；其中 Chooi-Ling GOH 与 AnDi Wu 与本文的工作最为相似。Chooi-Ling GOH 采用的是最大熵模型，但是其时间的代价却是很高的，实用性较差。AnDi Wu 在新词识别中只利用了 IWP，而本文利用了更丰富的信息及方法。

由于是后处理模型，我们的方法就要尽量简单有效，所以在处理过程中我们没有使用像最大熵（ME），SVM 等复杂度较高的模型，而是利用一些统计以及非统计的启发式方法，但是却有不错的效果。

3 单字碎片过滤模型

图 1 给出了整个系统的流程。

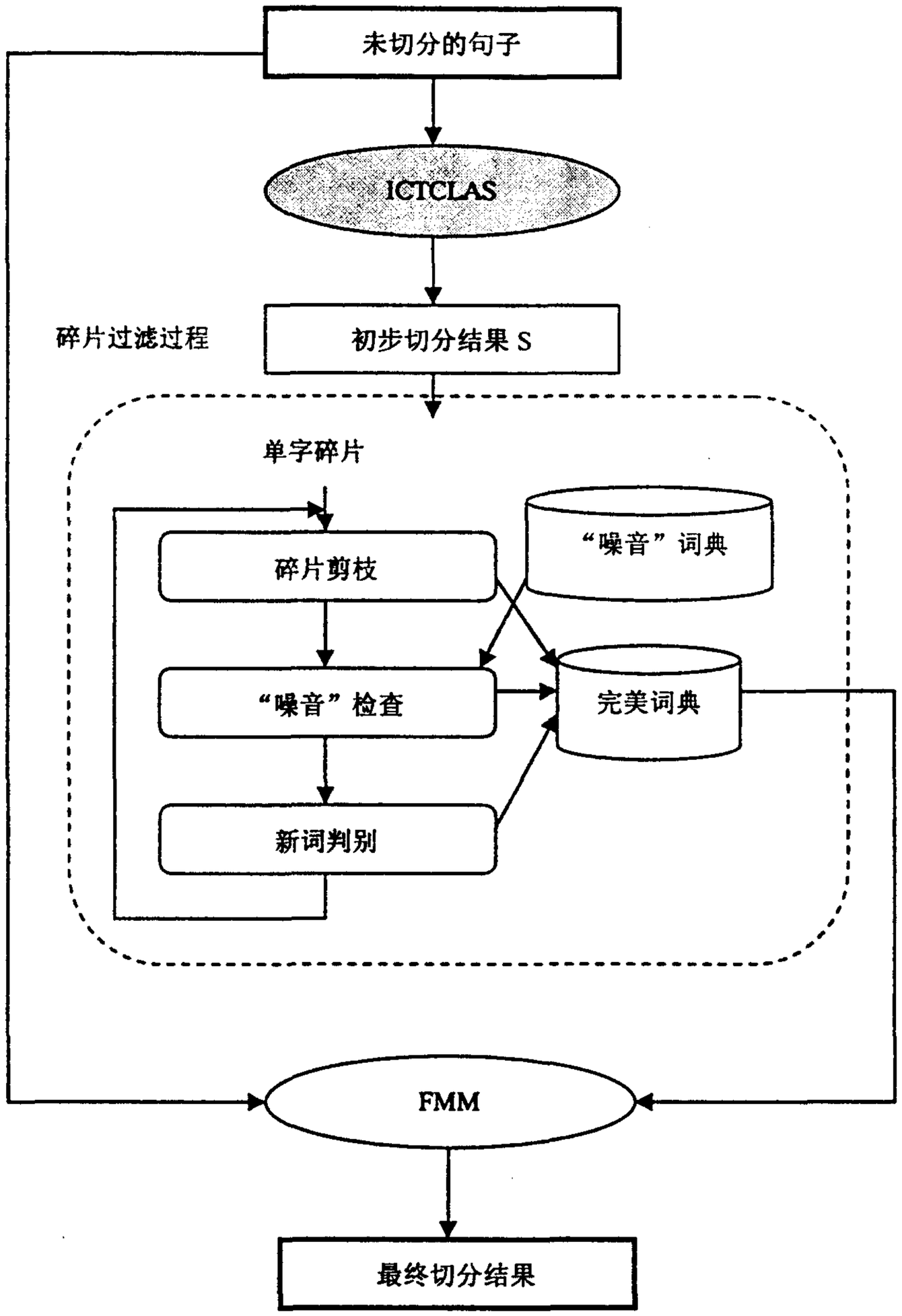

图 1 系统结构图

3.1 完美词典

在进行过滤的时候，我们将建立一部“完美词典”，之所以称“完美”是因为它里边存储的都是我们认为没有歧义，已经切分正确的词语。

一个初步切分结果要过滤之前，我们先为它建立一部空的“完美词典”，在以后过滤的每一步，词典都会被扩充，直到过滤结束。最后，我们将利用 “完美词典”来切分之前的句子，就得到最终的切分结果。

我们发现，同样的一个未登陆词，在不同的上下文就被可能为不同的结果（正确或错误），利用“完美词典”可以在一定程度上避免这个问题。完美词典在整个切分过程中是变化的，所以它可以称作是是“在线”的。

3.2 单字碎片剪枝

3.2.1 独立词概率（Independent Word Probability 简称 IWP）

大部分汉字即可以独立成一个单字词，也可以和其他汉字组成多字词。在这里，独立词概率就是一个汉字独立成词的概率，利用极大似然估计，一个汉字 *c* 的独立词概率 *IWP(c)*就可由下面的公式计算：

$$IWP(c) = \frac{N(Word(c))}{N(c)}$$

其中 *N(Word(c))*是汉字 c 在语料库中作为一个独立的词出现的次数，*N(c)*是指 *c* 在同一语料库中出现的总次数。

*IWP(c)*越接近 1，*c* 独立成词的可能性就越大。像“。”，“的”的独立词概率都很接近 1，这也符合客观事实。

3.2.2 碎片剪枝

在我们得到的碎片集合中，有些是很长的碎片，这样就不利于我们检测出其中包含的未登陆词，所以首先我们利用了独立词概率来对碎片进行“剪枝”。

比如“她/回/了/国/就/会/来/看/你/！”按照上面的假设 “她/回/了/国/就/会/来/看/你/！”整个句子就是一碎片，记为 *f*，如果我们把这样长达十个字的碎片作为候选，对于处理是很不利的。所以，我们用以下的方法来为碎片“剪枝”：

我们为汉字的 *IWP* 设定一阈值 *a*，对于碎片中某个汉字 *c*，如果 *IWP(c) > a* ,就认为 *c* 是正确的划分，就不是碎片，并将 *c* 放入“完美词典”。其中 *a* 是个经验值，通过在训练集上多次实验，取 *a=0.55*。

对于上面的碎片 *f*:

IWP(“ 她 ”)=0.91 ， IWP(“ 回 ”)=0.14 ， IWP(“ 了 ”)=0.92,IWP(“ 国 ”)=0.05 ， IWP(“就”)=0.68，

IWP(“会”)=0.11, IWP(“来”)=0.26；IWP(“看”)=0.4，IWP(“你”)=0.79

其中 *IWP* 大于阈值 *a* 的单字有：“她”，“了”，“就”，“你”；这样 *f* 就被剪枝成“回”，

“国”，“来/看”三个碎片，只有“来/看”为最终的碎片，其他的都放入“完美词典”。这样一个长度为 10 个字的候选碎片就被剪枝为 2 个字的候选碎片。

3.3 “噪音”检查

对于一个碎片，一般既包含正确的切分部分又包含错误的切分部分。正确的部分我们称之为“噪音”，要被过滤掉。而在错误的部分中可能包含被切分错误已知的低频词。所以“噪音”检查要做的就是将“噪音”过滤掉并合并错误切分的已知词。

例如：“他/却/被/短/接/发出/的/电弧/重重地/打倒/在/地上/”

单字碎片“他/却/被/短/接/”，其中“他/却/被/”是正确的切分，就是“噪音”；而“短/接/”是被切分错误的未登陆词。

如何过滤“噪音”，比如“我们/正/乘/汽车/去/北京”，其中“正/乘”是一个单字碎片，它既不是已知词，也不是新词，我们采用了基于实例的方法来进行过滤。

通过从训练语料库中抽取单字碎片实例，建立了一个实例词典，并假设：实例词典中的任何碎片的任何子串都是“噪音”。

以人民日报的切分句子为例：

改革/开放/以来/，/龙舟/之/乡/的/岳阳/人民/发扬/“/先/忧/后/乐/，/团结/求索/”/的/精神/，/“/以/舟/为/媒/”/，/致力/开拓/创新/，/经济/建设/突飞猛进/，/岳阳/正/乘/龙舟/腾飞/！/

像“之/乡/的”，“/先/忧/后/乐/”，“/以/舟/为/媒/”，和“正/乘/”都被看作“噪音”。它们的任何子串，比如：“先/忧”，“先/忧/后”，“以/舟”等也是“噪音”。

对于前面的例子，由于“正/乘”与实例匹配，所以就被过滤掉，“正”和“乘”分别被加入到“完美词典”。

这样，我们利用实例词典对初步切分得到的单字碎片进行检查，如果碎片的某一部分与某实例相匹配，就认为它是噪音，并将其过滤掉。

3.4 新词判别

一个词语之所以被切成了单字碎片，往往都是因为这个词语本身就是个未登录词而造成了，当一个碎片经过了上面的处理，剩下的部分就作为新词的候选。在这个新词判别中我们主要利用了与位置信息有关的概率 *P(c,position)*（Hongqiao Li et al.,2004）以及 *WFP(Word Formation Power)*。

3.4.1 P(c,position)

我们观察发现，对于某个单字它经常出现在一个词语某个特定位置，比如，“劳”经常出现在一个词的开始，而“性”经常出现在一个词语的末尾。我们利用了和位置有关的 *P(c,position)*。根据汉字的位置，我们将其分为 S,B,I,E,4 类（Chooi-Ling GOH et al.,2004），如表 1 所示。*P(c,position)*定义如下：

$$P(c, position) = \frac{N(c, position)}{N(c)}$$

其中，*N(c,position)*是汉字 *c* 在某种类型的位置出现的次数。其中 *P(c,S)*就是独立词概率。我们通过在训练语料库上训练，库中所有汉字的 *P(c,position)*表，部分举例见表 2。

表 1 汉字分类

Position	*Description*
S	单字词(*Single*)
B	多字词的开始(*Begin*)
I	多字词的中间(*Intermediate*)
E	多字词的末尾(*End*)

3.4.2 新词判别

这里我们主要利用 *Word Formation Power*(简称 *WFP*)(Chooi-Ling GOH et al.,2004)来判断一个碎片是不是新词。定义一个词语的 *WFP* 如下：

对于一个长度为 *n* 的词语 *w*, $w = c_1\ c_2 \ldots c_n$ ，

$$WFP(w) = P(c_1, B)\prod_{i=2}^{n-1} P(c_i, I)P(c_n, E)$$

对于一个碎片 $f = c_1\ c_2 \ldots c_n$,*WFP(f)*如果满足下面的任何一条规则它都不是新词，

（1） $WFP(f) < \min\{ WFP(x), length(x) = length(f)\}$;

（2） $WFP(f) < \prod_{i=1}^{n} P(c_i, S)$

（3） 对于 *length(f)=4*, $WFP(f) < P(c_1, S)P(c_2, S)P(c_3, B)P(c_4, E)$

或者 $WFP(f) < P(c_1, B)P(c_2, E)P(c_3, S)P(c_4, S)$

（4） 其中某个 c_i， $P(c_i, S) = 1$；

如果碎片是新词就加入到新词词典中；如果不是新词，我们就将碎片中的每一个汉字都加到“完美词典”中。

通过 *WFP* 我们就可以把如：“亚冬会”，“武科瓦尔”，“乌铁局”等碎片识别成新词。

表 2 汉字 *P(c,position)*举例

汉字 c	*P(c,S)*	*P(c,B)*	*P(c,I)*	*P(c,E)*
中	0.298493	0.539543	0.062117	0.0998465
国	0.0500065	0.348293	0.0349129	0.566788

4 实验结果及错误分析

4.1 切分实验结果

与 SIGHAN Bakeoff 一样，我们也从召回率(Recall)，准确率(Precision)，F-measure，以及未登录词召回率(OOV Recall)，登录词召回率(IV Reacll)这几个方面来评价分词性能的好坏(Richard and Thomas , 2003)。F 值为主要评价指标。我们利用了 SIGHAN Bakeoff data：PK（北大）测试集，以及一个特定领域[6]的测试集来测试我们的新，旧系统，结果分别见表 3，表 4。

$$\text{Recall}=\frac{\text{正确切分的词数}}{\text{答案中的词数}} \qquad \text{Precision}=\frac{\text{正确切分的词数}}{\text{切分结果中的词数}} \qquad \text{F}=\frac{2*\text{Reacll}*\text{Precision}}{\text{Reacll}+\text{Precision}}$$

$$\text{Recall}_{oov}=\frac{\text{正确切分的未登录词数}}{\text{答案中的未登录词数}} \qquad \text{Recall}_{IV}=\frac{\text{正确切分的登录词数}}{\text{答案中的登录词数}}$$

表 3 PK 测试集结果

PK (56K)	Recall	Precision	F-measure	OOV Recall	IV Recall	Time Cost
ICTCLAS	0.955	0.932	0.943	0.785	0.972	1.78s
ICTCLAS+Filter	0.956	0.95	**0.953**	0.825	0.969	1.87s

表 4 特定领域测试集结果

特定领域(289K)	Recall	Precision	F-measure	OOV Recall	IV Recall	Time Cost
ICTCLAS	0.831	0.692	0.759	0.421	0.964	9.08s
ICTCLAS+Filter	0.847	0.762	**0.802**	0.531	0.968	10.66s

可以看出，通过引入过滤模型，新系统在准确率和召回率上都比原系统由一定的提高。PK 测试集测试结果 F 值提高了 1%，其中 OOV Recall 提高了 4%。对特定领域测试，新系统的 F 值提高了 4.3%，其中 OOV Recall 提高了 11.1%，可以看出新系统具有更好的自适应性。在于时间消耗方面，从表中可以看出，时间消耗的增加只有 10%左右。

4.2 新词识别结果

过滤模型之所以提高了性能，一个主要原因就是识别到了更多的新词，所以我们也计算了模型的新词检测的召回率，准确率和 F 值。

[6] 这里的特定领域主要是指按北大标准切分的化学领域的语料。

表 5，表 6 给出了新系统在 PK 测试语料以及特定领域语料上的新词检测结果。

表 5 PK 新词识别结果

PK(56K)	OOV Recall	OOV Precision	OOV F-measure
ICTCLAS	0.785	0.935	0.853
ICTCLAS+Filter	0.825	0.915	**0.867**

表 6 特定领域新词识别结果

特定领域(289K)	OOV Recall	OOV Precision	OOV F-measure
ICTCLAS	0.421	0.90	0.574
ICTCLAS+Filter	0.531	0.844	**0.652**

4.3 错误分析

在我们的过滤过程主要有三步，在对碎片进行剪枝的时候，我们利用了独立词概率过滤掉了独立词较大的词语，并且概率阈值是个经验值，所以会带来一些错误，比如碎片“多/云/转/晴”，当我们进行剪枝时，由于 IWP(“多”)=0.578 大于阈值 0.55，所以“多”就被剪掉，而实际“多/云/转/晴”正确的切分就是“多云转晴”。

然后，我们使用基于实例的过滤方法，这虽然可以将一些正确的切分滤掉，但也会带来一些错误，比如碎片“有/没/有/”，按照我们的过滤规则，“没/有”就会当成已知词合并，切分就会变成“有/没有”，而正确的切分就是“有/没/有”。

最后就是新词的检测，这里主要利用了 WFP，这种方法虽然可以过滤掉一些构词能力比较弱的碎片，但碎片的 WFP 较高也并不意味着碎片一定是新词，比如碎片“耀/灯/辉”，就被错误的识别为新词。

由上可以看出，单字碎片过滤方法还有很多局限性，需要我们去进一步研究。

5 结论

本文介绍了一种单字碎片过滤模型来对分词结果进行“后处理”，原系统是基于多层隐马的系统，主要通过利用词语出现的频率以及词语之间的共现频率来进行切分，所以在未登录词语比较多的时候，我们原系统就会产生很多的错误，过滤模型更多是从“字”的角度出发，通过利用单字的概率信息，在一定程度上降低了由于低频词影响带来的错误。从试验看出，我们的方法还是比较有效的，通过过滤，一方面，提高了已知词的切分召回率，另一方面，也提高了未登录词的召回率，从而在整体上提高了分词的性能。

参考文献

[1] Huaping Zhang, Hongkui Yu, Deyi Xiong et al. 2003.HMM-based Chinese Lexical Analyzer ICTCLAS. Inproceedings of the Second SIGHAN Workshop onChinese Language Processing, July 11-12,

2003,Sapporo, Japan.

[2] Goh Chooi-Ling,Masayuki Asahara,and Yuji Matsumoto.2004.Chinese Word Segmentation by Classification.In Proceedings of COLING,pages 466-472.

[3] Goh Andi Wu,Zixin Jiang.2000.Statistically-Enhanced New Word Identification in a Rule-Based

[4] Chinese System.Proceedings of the Second Chinese Language Processing Workshop，HongKong.

[5] Hongqiao Li, Chang-Ning Huang, Jianfeng Gao and Xiaozhong Fan, 2004. The use of SVM for Chinese new word identification. In IJCNLP-04. Sanya City, Hainan Island, China, March 22-24, 2004.

[6] Hongqiao Li,Aitao Chen. 2003. Chinese Word Segmentation UsingMinimal Linguistic Knowledge. In proceedings of the Second SIGHAN Workshop on Chinese LanguageProcessing, July 11-12, 2003, Sapporo, Japan.

[7] Chooi-Ling Goh, Masayuki Asahara and Yuji Matsumoto.Pruning False Unknown Words to Improve Chinese Word Segmentation.PACLIC-2004, pp.139-149.

[8] Richard and Thomas Emerson. 2003. The First International Chinese Word Segmentation Bakeoff. In Proceedings of Second SIGHAN Workshop, pages 133–143.

[9] Asahara, Masayuki, Chooi-Ling Goh, Xiaojie Wang, and Yuji Matsumoto. 2003. Combining Segmenter and Chunker for Chinese Word Segmentation. In Proceedings of Second SIGHAN Workshop, pages 144–147.

[10] Hua-Ping Zhang,Qun Liu,Xue-Qi Cheng,Hao Zhang and Hong-Kui Yu.2003.Chinese Lexical Analysis Using Hierarchical Hidden Markov Model. In:Proceedings of the Second SIGHAN Workshop 2003