DOI:10.16515/j.cnki.32-1722/n.2019.03.001

# 基于HMM的算法优化在中文分词中的应用

朱咸军[1,2],洪 宇[1],黄雅琳[1],张馨予[1],肖芳雄[1*]

(1.金陵科技学院软件工程学院、网络安全学院,江苏 南京 211169;
2.江苏省软件测试工程实验室,江苏 南京 211169)

**摘 要**:随着社交软件的普及,社交软件中社会关系分析日益凸显。中文分词是社会关系分析的一种重要手段,但是现有中文分词方法的效果不好。提出基于隐马尔科夫模型(Hidden Markov Model,HMM)的中文分词优化算法。它们是将基于词典分词算法产生的结果作为附加信息,添加到HMM模型中,在不改动HMM模型的情况下,有效地增加了HMM模型的分词效果。实验结果表明,改进HMM算法能显著提高中文分词的准确率、召回率和$F$值。

**关键词**:隐马尔科夫模型;优化HMM;中文分词

**中图分类号**:TP391 **文献标识码**:A **文章编号**:1672-755X(2019)03-0001-07

## Application in Chinese Segmentation Based on Optimization-HMM Algorithms

ZHU Xian-jun[1,2], HONG Yu[1], HUANG Ya-lin[1], ZHANG Xin-yu[1], XIAO Fang-xiong[1*]

(1. Jinling Institute of Technology, Nanjing 211169, China;
2. Jiangsu Software Testing Engineering Laboratory, Nanjing 211169, China)

**Abstract**: With the popularity of social software, the analysis of social relationships in social software has become increasingly highlighted. Chinese word segmentation is an important means in the study of social relationship analysis. However, the existing method does not perform well. This study proposes some optimization algorithms based on Hidden Markov Model (HMM). These algorithms add the results based on the dictionary segmentation algorithms as additional information to the HMM model, and effectively increase the segmentation effect of the HMM model without changing the model. Experimental results show that the improved HMM algorithms can significantly increase the accuracy, recall rate and $F$ value of Chinese word segmentation.

**Key words**: HMM; optimization-HMM algorithms; Chinese word segmentation

中文分词作为中文信息处理最基础的一个方向,是机器翻译、信息检索的重要基础。即使在深度学习发展迅猛的今天,仍然需要进行分词,将词语进行训练转化为词向量(Word Embedding),从而进一步转化为神经网络的输入。词语是表达语义的最小单位,英语等拉丁语系因其天然的优势,分词过程中可以利用

**收稿日期**:2019-07-03

**基金项目**:江苏省高等学校自然科学研究面上项目(18KJB520018);金陵科技学院高层次人才科研启动基金(jit-b-201703,jit-rcyj-201802);江苏省现代教育技术研究课题(2018-R-63099);金陵科技学院教育教改课题(jyjg2017-21)

**作者简介**:朱咸军(1977—),男,江苏建湖人,讲师,博士,主要从事软件工程、协同计算、智能信息处理与智能系统方面的研究。

**通信作者**:肖芳雄(1971—),男,广西桂林人,教授,博士,主要从事软件工程和大数据方面的研究。

空格作为分界。而中文等东方语言，只有段落之间和句子之间有分界符，词语与词语之间并没有分界符，这就给中文分词造成了困难。目前中文的分词可分为三大类：基于词典的方法、基于统计的方法和混合方法[1]。刘晨晖等提出了基于Kert改进算法的中文主体关键短语提取算法[2]。刘勇等提出了基于双字哈希结构的最大匹配改进算法，确保最大匹配算法分词精度的前提下提高分词速度[3]。歧义消解和未登录词识别是提升分词准确率和召回率的两个突破口[4]。其中，未登录词识别造成的精度失落远高于歧义识别造成的精度失落[5]。基于词典的方法依赖词典中词目的选择和词条数目，不存在足够大的词典来包含所有中文词语，所以未登录词的识别是影响该算法效果的重要因素[6]。隐马尔科夫模型(Hidden Markov Model，HMM)模型是典型的统计模型算法，结合字标注的思想，能够在不依赖词典下自主识别未登录词。但由于HMM输出独立性的假设，不能很好地考虑上下文特征，识别准确率不高[7]($F$值约为80%)。宫法明等提出了一种基于自适应隐马尔可夫模型的分词算法，它结合领域词典和相互信息，以语义约束和词义约束校准分词，实现对石油领域专业术语和组合词的精确识别[8]。蒋卫丽提出了一种基于领域词典的动态规划分词算法，通过对特定领域的信息进行分词实验，验证了该文提出的分词算法可获得较高的分词准确率与召回率[9]。

因此，本文从解决HMM假设的局限性的角度出发，在不改动HMM算法的前提下，通过扩充HMM模型中的观测矩阵，从而提升HMM的分词效果。

## 1 三类分词模式

基于词典的分词算法，是指按一定的文本扫描顺序对词典进行搜索，对符合匹配规则的词语进行划分。按照文本扫描顺序的不同可分为正向扫描、逆向扫描和双向扫描。按照匹配原则不同可分为最大匹配、最佳匹配、最小匹配和逐词匹配。本文主要采用最大正向匹配和最大逆向匹配。最大正向匹配(MM，Maximum Matching)是指从左往右扫描文本，将文本中连续的字符串与词典进行比较，选取长度最长的字符串作为划分单元。最大逆向匹配(RMM，Reverse Maximum Matching)与MM类似，但文本扫描方式从右往左。双向最大匹配法(Bi-directction Maximum Matching，BMM)是MM得到的分词结果和RMM得到的结果进行比较，从而决定正确的分词方法。

如，待分词句为“我去天安门城楼”。MM和RMM都会切分为“我/去/天安门/城楼/”。统计表明[10]，最大正向匹配的错误率为1/169，逆向最大匹配的错误率为1/245。

当MM与RMM切分结果不同时，表明该文本存在歧义。例如，“结婚的和尚未结婚的”。MM会切分成“结婚/的/和尚/未/结婚/的/”。而RMM会切分成“结婚/的/和/尚未/结婚/的”。同时使用最大正向匹配和逆向最大匹配可以发现链长等于1的交叉歧义，但依旧无法解决未登录词造成的精度失落问题。

## 2 基于HMM的中文分词

HMM是一种双重随机过程，包含一个不可见、隐藏的状态转换随机过程和一个用于产生观测符号的随机过程。两个随机过程产生的随机序列分别称为状态序列和观测序列[11]。

利用HMM实现中文分词采用了字标注的思想，该思想将分词视为字的词位标注问题，从而可以利用机器学习算法进行分类。最常见的是通过4tag进行标注(Begin，词开头；Middle，词中间；End，词结尾；Single，单字词)[4]。例如，“毕业于北大”可以标注为“BE/S/BE”。4tag并不是唯一的标注方式，Wenchao等人就标注的精细度做过相关研究，理论上标注越多效果会越好，但是实际实验过程中发现标注越精细，样本不足的问题越突出，导致$F$值不高[12]。

设状态集合$E=\{e_1,e_2,\cdots,e_M\}$，观测集合$O=\{o_1,o_2,\cdots,o_N\}$，观测序列$X=\{x_1,x_2,\cdots,x_n\}$，$x_i\in O$，状态序列$Y=\{y_1,y_2,\cdots,y_n\}$，$y_i\in E$。在中文分词的应用场景中，$E$为汉字集合，$S$为标注状态集合(即$\{B,M,E,S\}$)，$X$为待切分的汉字序列，$Y$为输出字状态序列。

目标即求：

$$MAXP(Y|X)=MAX\frac{P(Y)P(X|Y)}{P(X)} \tag{1}$$

由于 $P(X)$已定，即求 $P(Y)P(X|Y)$的最大值。

根据一阶马尔科夫假设，当前隐藏状态值只与前一个隐藏状态值有关。

$$P(Y)=P(y_1)\prod_{i=2}^{n}P(y_t\mid y_{t-1}) \tag{2}$$

根据观测独立性假设，当前观测值只与当前隐藏状态有关：

$$P(X\mid Y)=\prod_{i=1}^{n}P(x_t\mid y_t) \tag{3}$$

所以，

$$P(XY)=P(x_1y_1)\prod_{i=2}^{n}P(y_t\mid y_{t-1})P(x_t\mid y_t) \tag{4}$$

$P(XY)$的最优解可以通过维特比算法求得[13]。

由此引入 HMM 模型参数，$\lambda=(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\pi})$，其中 $\boldsymbol{A}=[a_{ij}]_{M\times M}$为状态转移矩阵，$\boldsymbol{B}=[b_{ij}]_{M\times N}$为发射矩阵，$\boldsymbol{\pi}=(\pi_1,\pi_2,\cdots,\pi_M)$为初始概率矩阵。

$$a_{ij}=P(y_{t+1}=s_j\mid y_t=s_i),1\leqslant i\leqslant M,1\leqslant j\leqslant M \tag{5}$$

$$b_{ij}=P(x_t=o_j\mid y_t=s_i),1\leqslant i\leqslant M,1\leqslant j\leqslant N \tag{6}$$

$$\pi_i=P(y_1=s_i),1\leqslant i\leqslant M \tag{7}$$

设 $c(x)$表示语料中 $x$ 出现的次数，通过极大似然估计，有

$$P(x_i\mid y_i)=\frac{c(x_i,y_i)}{c(y_i)} \tag{8}$$

$$P(y_i\mid y_{i-1})=\frac{c(y_i,y_{i-1})}{c(y_{i-1})} \tag{9}$$

为了解决语料不足而导致上述概率估计的稀疏性问题，会使用平滑估计法，常用的平滑技术包括 Linear Interpolation[14] 和 Discounting Methods。

## 3　优化 HMM

HMM 的两大假设对问题做了极大的简化，导致分词效果不理想（$F$ 值约为 80%）。本文提出一种 HMM 优化算法，通过将观测集合和状态集合组合，形成新的观测集合，并利用词典匹配算法产生的信息扩充观测序列，从而提升分词效果。

定义⊕为拼接符号，$\bar{o}_{ij}=o_i\oplus e_j$ 表示观察集合中的观察值 $o_i$ 与状态集合中的状态值 $s_j$ 相结合，形成新的观察值。例如，$o_{ij}=$“我”，$s_j=$“$E$”，则 $\bar{o}_{ij}=$“我 $E$”。通过定义拼接符号，观测集从 $ON$ 扩充为 $\bar{O}_{N\times M}$。

通过词典匹配算法来扩充观测序列也有很多方案。基于词典匹配算法的可选方案就包括最大正向匹配（MM）、最大逆向匹配（RMM）、双向最大匹配（BMM）。根据附加信息的添加位置又可分为对位添加（记为＋0）、后错一位（记为＋1）和前错一位（记为－1）。

例如，sentence＝“结婚的和尚未结婚的”，MM 产生的结果为“BMSBMSBMS”，RMM 产生的结果为“BMSSBMBMS”。在对位添加附加信息的情况下，MM＋0 表示使用正向匹配和对位添加附加信息，其将 sentence 扩充为“结 B 婚 M 的 S 和 B 尚 M 未 S 结 B 婚 M 的 S”。

又如 BMM＋1 表示使用双向匹配和后错一位添加附加信息，将词典匹配算法产生的状态序列与其对应的观测序列向后错开，sentence 扩充为“结 # # 婚 BB 的 MM 和 SS 尚 BS 未 MB 结 SM 婚 BB 的 MM”。“婚 B”表示“婚”前一个字的字典匹配结果为词头（B），“#”是占位符。

测试语句“结婚的和尚未结婚的”的三类匹配模式下中文分词结果，以及每类匹配模式对应的三种优化算法修正后的测试语料如表 1 所示。

**表 1 三类匹配模式与添加位置**

| 算法 | 结婚的和尚未结婚的 |
|---|---|
| MM 结果 | BMSBMSBMS |
| RMM 结果 | BMSSBMBMS |
| BMM 结果 | BBMMSSBSMBSMBBMMSS |
| MM−1 修正 | 结 M 婚 S 的 B 和 M 尚 S 未 B 结 M 婚 S 的 # |
| MM+0 修正 | 结 B 婚 M 的 S 和 B 尚 M 未 S 结 B 婚 M 的 S |
| MM+1 修正 | 结 # 婚 B 的 M 和 S 尚 B 未 M 结 S 婚 B 的 M |
| RMM−1 修正 | 结 M 婚 S 的 S 和 B 尚 M 未 B 结 M 婚 S 的 # |
| RMM+0 修正 | 结 B 婚 M 的 S 和 S 尚 B 未 M 结 B 婚 M 的 S |
| RMM+1 修正 | 结 # 婚 B 的 M 和 S 尚 S 未 B 结 M 婚 B 的 M |
| BMM−1 修正 | 结 MM 婚 SS 的 BS 和 MB 尚 SM 未 BB 结 MM 婚 SS 的 # # |
| BMM+0 修正 | 结 BB 婚 MM 的 SS 和 BS 尚 MB 未 SM 结 BB 婚 MM 的 SS |
| BMM+1 修正 | 结 # # 婚 BB 的 MM 和 SS 尚 BS 未 MB 结 SM 婚 BB 的 MM |

记输入序列为 $X\subseteq R^n$，扩充函数为 $f(X)\subseteq R^n$，则

$$P(Y\mid X)\approx P(Y\mid X,f(x)) \tag{10}$$

$$P(XY)=P(x_1,f(x_1),y_1)\prod_{i=2}^{n}P(y_t\mid y_{t-1})P(x_tf(x_t)\mid y_t) \tag{11}$$

在本文中，扩充函数 $f(X)$就是词典匹配函数。通过将词典匹配的结果加入到 HMM 模型中，可以在不更改 HMM 模型的前提下，提高中文分词的效果。

基于一阶马尔科夫链的概率语言模型又可以被称为二元语法模型(Bigram)。对于 N-gram 语法模型，其空间复杂度为 $O(|V|^N)$，其中 $V$ 为字典长度，$N$ 为 N-gram 语法模型中窗口大小，在 Bigram 模型中，$N=2$。在使用单向词典匹配作为优化策略添加到 HMM 模型的情况下，观测集合由 $|V|$ 扩充为 $M\cdot|V|$，$M$ 为状态集大小，故优化 HMM 的空间复杂度为 $O((M\cdot|V|)^N)$，在使用双向词典匹配的情况下，空间复杂度进一步扩大为 $O((2M\cdot|V|)^N)$。

一阶马尔科夫模型使用维特比算法解码的时间复杂度为 $O(n|M|^2)$，使用 Trie 树结构的词典单向分词算法时间复杂度为 $O(n)$，$n$ 为待分词序列长度，$M$ 为状态集合大小。在使用单向词典匹配作为优化策略添加到 HMM 模型的情况下，时间复杂度为 $O(n+n|M|^2)$；使用双向匹配作为优化策略的时间复杂度为 $O(2n+n|M|^2)$。

## 4 实验验证

为了验证本文的研究设计了两个实验进行验证。实验环境为 CPU Intel Core i5 2.4 GHZ，6 G 内存，Ubuntu 14.04 操作系统，Python 3.5 应用软件。

第二届国际汉语分词测评中有四家单位(Academia Sinica、City University、Peking University 和 Microsoft Research)提供了测试语料。四家提供的测评资源(icwb2-data)中包含了它们各自的训练集(Training)、测试集(Testing)以及根据各单位分词标准下的相应测试集标准答案(icwb2-data/scripts/gold)。在 icwb2-data 资源目录下 scripts 目录中包含了对分词自动评分的 pert 脚本 score。

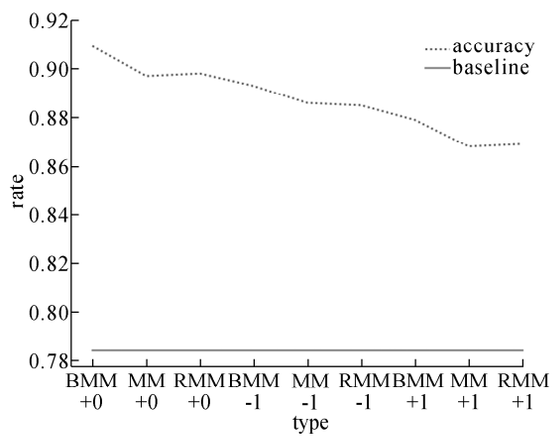
在本节实验环节采用了两种具有代表性的测试语料，即 Peking University 、Microsoft Research，其中 Peking University 测试语料训练集包含 19 056 条分词语句和 1 945 条待分词语句的测试集合(pku-test)，Microsoft Research 测试语料训练集包含 86 924 条分词语句和 3 985 条待分词语句的测试集合(msr_test)。

表 2 统计了 10 种算法的准确率、召回率、$F$ 值以及实验的耗时情况，其中 HMM 测试结果为本节实验的参照。

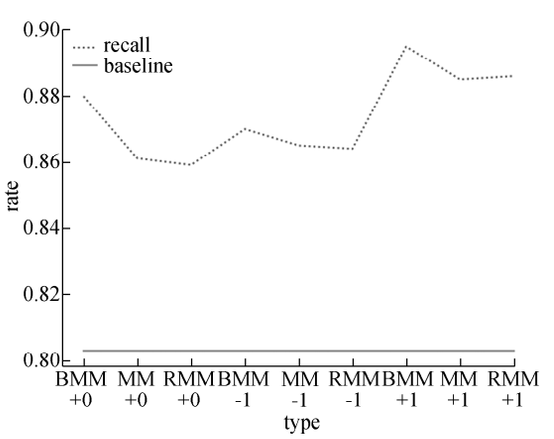
**表 2　10 种算法的测试结果**

| 算法 | Msr_test 测试结果 | | | | Pku_test 测试结果 | | | |
|---|---|---|---|---|---|---|---|---|
| | 准确率 | 召回率 | $F$ 值 | 时间/$s$ | 准确率 | 召回率 | $F$ 值 | 时间/$s$ |
| MM＋0 | 0.897 | 0.861 | 0.879 | 1.22 | 0.909 | 0.861 | 0.884 | 1.32 |
| RMM＋0 | 0.898 | 0.859 | 0.878 | 1.12 | 0.902 | 0.862 | 0.882 | 1.21 |
| BMM＋0 | 0.909 | 0.880 | 0.894 | 1.99 | 0.923 | 0.861 | 0.891 | 1.91 |
| MM＋1 | 0.868 | 0.885 | 0.876 | 1.23 | 0.888 | 0.883 | 0.885 | 1.33 |
| RMM＋1 | 0.869 | 0.886 | 0.877 | 1.22 | 0.880 | 0.891 | 0.885 | 1.12 |
| BMM＋1 | 0.879 | 0.895 | 0.887 | 2.01 | 0.890 | 0.901 | 0.895 | 1.92 |
| MM-1 | 0.886 | 0.865 | 0.875 | 1.31 | 0.893 | 0.872 | 0.882 | 1.31 |
| RMM-1 | 0.885 | 0.864 | 0.874 | 1.34 | 0.887 | 0.879 | 0.883 | 1.29 |
| BMM-1 | 0.893 | 0.870 | 0.881 | 2.10 | 0.897 | 0.876 | 0.886 | 1.97 |
| HMM | 0.784 | 0.803 | 0.793 | 0.63 | 0.763 | 0.782 | 0.772 | 0.63 |

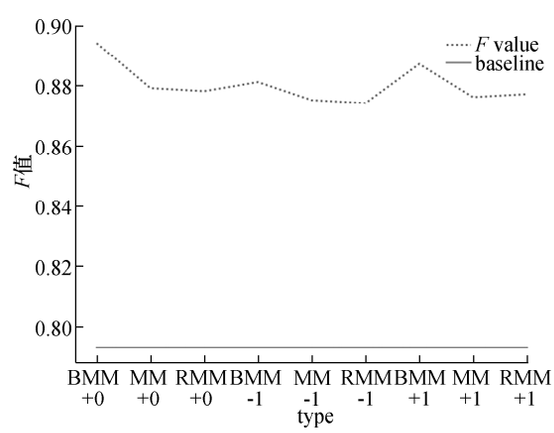
由表 2 可知，所有方法的准确率、召回率和 $F$ 值均优于 HMM 算法的测试结果，但测试耗时均高于 HMM 算法耗时，BMM 的准确率、召回率和 $F$ 值优于 MM 和 RMM。由图 1(a)和图 2(a)可知，对位添加(＋0)算法的准确率优于其它两种改进算法，其中 BMM＋0 算法的准确率最高。由图 1(b)和图 2(b)可知，后错一位(＋1)算法的召回率相较于其它两种改进算法，BMM＋1 算法的召回率最高。由图 1(c)和图 2(c)可知，BMM＋0 算法和 BMM＋1 算法的 $F$ 值相近且优于其它算法。



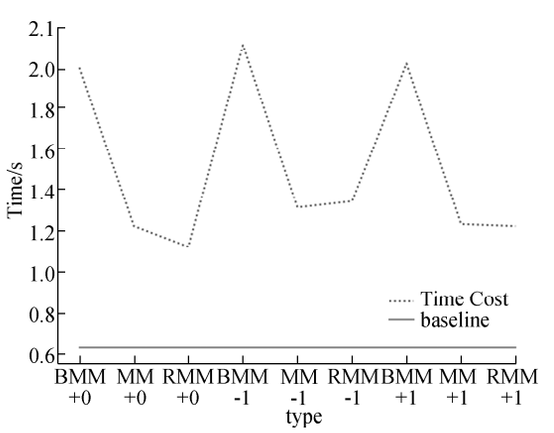

(a) Microsoft Research 的准确率




(b) Microsoft Research 的召回率

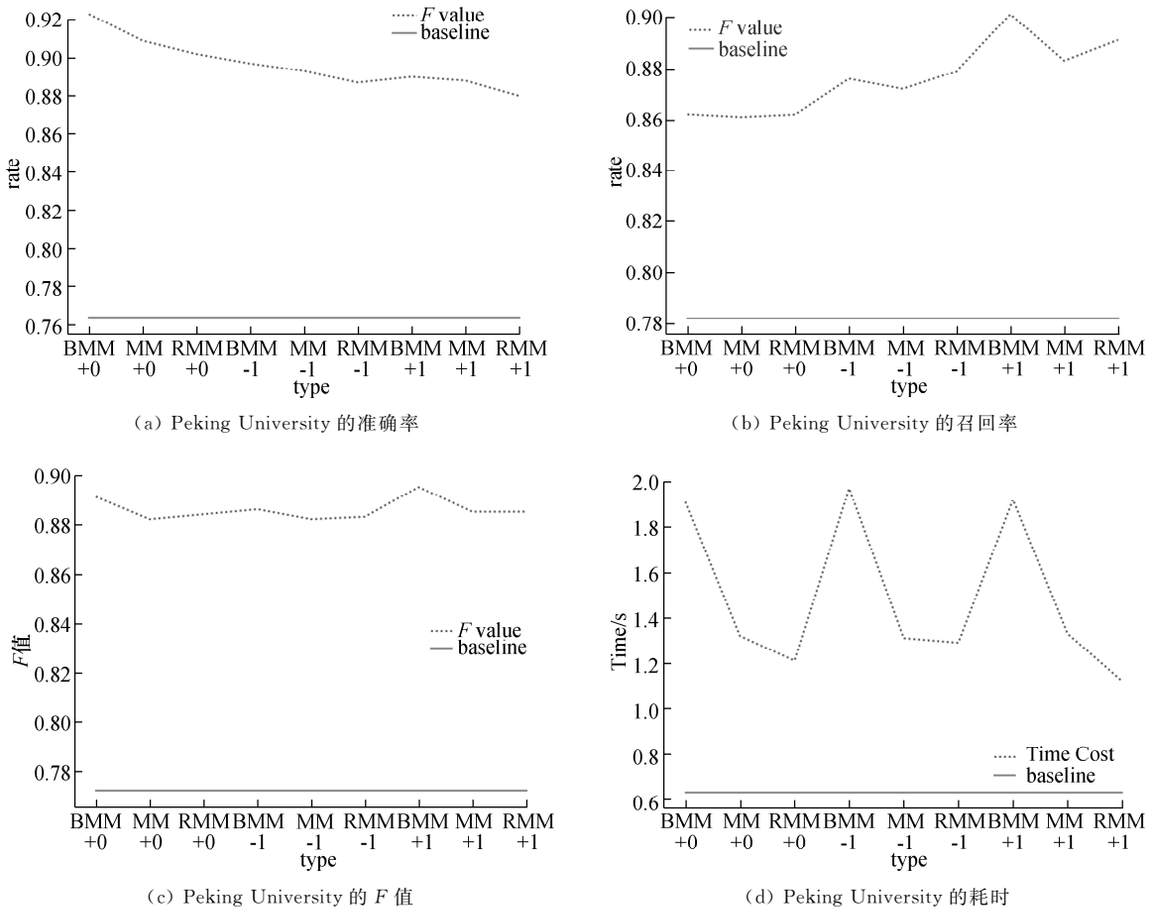



(c) Microsoft Research 的 $F$ 值




(d) Microsoft Research 的耗时

**图 1　Msr_test 测试结果**




(a) Peking University 的准确率　(b) Peking University 的召回率

(c) Peking University 的 $F$ 值　(d) Peking University 的耗时

**图 2　Pku_test 测试结果**

## 5　结　语

优化 HMM 算法可以明显提高中文分词的准确率、召回率和 $F$ 值。双向最大匹配(BMM)在改进中文分词的准确率、召回率和 $F$ 值效果较为明显，BMM＋0 算法在中文分词的准确最高和 BMM＋1 算法在中文分词召回率的改进效果最高。同时，双向词典匹配所影响的 $F$ 值提升与单向匹配所影响的 $F$ 值提升差异不大。

在后续的研究中，将相关算法应用于社交网络中的聊天记录分析，从而研究社交网络成语的喜好程度、决策偏好等信息。

**参考文献：**

[1] 付国宏，王晓龙. 汉语词语边界自动划分的模型与算法[J]. 计算机研究与发展，1999，36(9)：1144－1147

[2] 刘晨晖，张德生，胡钢. 基于 Kert 的中文主题关键短语提取算法[J]. 计算机应用，2019，39(S1)：245－249

[3] 刘勇，魏光泽. 基于双字哈希结构的最大匹配算法机制改进[J]. 电子设计工程，2017，25(16)：11－15

[4] Xue N，Shen L. Chinese word segmentation as LMR tagging[C]//Sighan workshop on Chinese language processing. association for computational linguistics. Philadelphia：University of Pennsywania Philadelphia，2003：176－179

[5] 黄昌宁，赵海. 中文分词十年回顾[J]. 中文信息学报，2007，21(3)：8－19

[6] 张春霞，郝天永. 汉语自动分词的研究现状与困难[J]. 系统仿真学报，2005，17(1)：138－143

[7] Lu X. Towards a Hybrid Model for Chinese Word Segmentation[J]. LDV Forum，2003，22：56－62

[8] 宫法明，朱朋海. 基于自适应隐马尔可夫模型的石油领域文档分词[J]. 计算机科学，2018，45(S1)：97－100

[9] 蒋卫丽，陈振华，邵党国，等. 基于领域词典的动态规划分词算法[J]. 南京理工大学学报，2019，43(1)：63－71

[10] 朱世猛. 中文分词算法的研究与实现[D]. 成都：电子科技大学，2011

[11] Rabiner L R. A tutorial on hidden Markov model and selected applications in speech recognition[J]. Readings in Speech Recognition，1989，77(2)：267－296

[12] Wen C M，Lian C L，An Y C. A comparative study on Chinese word segmentation using statistical models[J]. Proceedings 2010 IEEE International Conference on：Software Engineering and Service Sciences，2010(3)：482－486

[13] Gilhousen K S，Jacobs I M，Padovani R，et al. System and method for generating signal waveforms in a CDMA cellular telephone system[J]. NI Signal Generator，2015，33：92－100

[14] Meijering E. A chronology of interpolation：from ancient astronomy to modern signal and image processing[J]. Proceedings of the IEEE，2002，90(3)：319－342

（责任编辑：谭彩霞）