

# 一个基于分类数据挖掘的汉语分词消歧模型

吕强\* 夏俊鸾 钱培德  
(苏州大学计算机科学与技术学院)  
(江苏省计算机信息处理技术重点实验室)

## A Model of Chinese Segmentation Disambiguation Based on Classification Rule Discovery

Lv Qiang Xia Junluan Qian Peide  
(School of Computer Science and Technology, Soochow Univeristy)  
(Jiangsu Provincial Key Lab of Computer Information Processing)

**Abstract:** After analyzing the current typical Chinese segmentation evaluation functions as well as factors in these evaluation functions, this paper points out that these approaches are simply naive and lack of objectivity and completeness. The paper presents a new approach to address the problem. We first extract the ambiguous data, which can completely describe the right segmentation of ambiguous words and its proper context, from the training set (tagged corpus). Then we mine the rules from these data with classification rule discovery. And we use these rules as the new evaluation function. Such evaluation function can show the better objectivity, completeness and extensibility over the previous ones. Finally we test our approach on the latest benchmarks with the comparative comments. Therefore we present a framework for making segmentation knowledge computable.

**Keywords:** Chinese segmentation; Disambiguation; Classification rule discovery

---

\*Email: qiang@suda.edu.cn Tel: 0512-67165762ext203 Addr: 江苏省苏州市十梓街1号158信箱邮编215006

**摘要:** 通过分析经典的分词评价函数以及在评价函数中常用的分词要素, 本文指出了这些评价函数在客观性和完备性方面都朴素简单的不足。本文的基本方法是从熟语料库中收集歧义数据, 这些数据可以详细刻画歧义的切分情况以及歧义所在的语境, 对这些数据进行分类数据挖掘, 抽象出其中的分类规则, 再用这些规则作为分词评价函数, 从而体现了客观性、完备性和可扩展性。最后利用本文的方法进行了分词测评, 给出了比较结果。本文提供了一个很好的将分词知识可计算化的框架。

**关键词:** 中文分词; 消歧; 分类规则挖掘

## 1 引言

汉语自动分词在各种信息系统中是极为重要的基本技术, 同时也是计算语言学界公认的难题。目前没有解决的两大难点就是歧义切分消解和未登陆词识别。针对于这两大难点已有不少研究者提出了很多的解决的方法, 目前主流的解决方案是语料库和统计技术为主。而最终实现分词过程的关键是可计算化的分词准则, 也就是分词的评价函数。一个句子或者短语如何切分是根据评价函数而来, 评价函数的好坏也就直接决定着分词结果的好坏。

本文从分析分词的评价函数入手, 这里的评价函数事实上也就是分词知识的一种表述。我们通过研究现有的评价函数, 从熟语料库中收集歧义数据, 这些数据可以详细刻画歧义的切分情况以及歧义所在的语境。对这些数据进行分类数据挖掘, 抽象出其中的规则, 用这些规则作为评价函数, 体现了客观性、完备性和可扩展性。

## 2 相关研究

记 $\mathcal{A}$ 为评价函数所使用的分词要素集合,  $a_i \in \mathcal{A}, i = 1, 2, \dots$ 。目前在统计语言模型中评价函数常用的要素有如下几种:

- $a_1$ 词频[1]: 一个词在大规模训练语料库中出现的次数;
- $a_2$ 局部词频[2]: 一个词在一个语言环境中出现的次数, 这个语言环境可以是一句话或者一个段落;
- $a_3$ 词性[1]: 一个词在大规模训练语料库中出现的词性, 通常不止一种;

- $a_4$ 字互信息[3, 4]: 对于汉字串 $xy$ , 互信息 $mi(x, y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$ , 其中 $p(x, y)$ 表示在大规模训练语料库中 $(x, y)$ 同现的概率,  $p(x)$ 和 $p(y)$ 分别表示 $x, y$ 出现的概率。 $mi(x, y)$ 表现出两个字 $x$ 和 $y$ 在整个语料库中结合的紧密程度;
- $a_5$ t-测试差[3, 4]: 对于汉字串 $vx yw$ ,  $x, y$ 之间的t-测试差 $dts(x, y) = t_{v,y}(x) - t_{x,w}(y)$ , 其中 $t_{v,y}(x) = \frac{p(z|y) - p(y|x)}{\sqrt{\sigma^2((p(z|y) + p(y|x)))}}$ ,  $t_{x,w}(y)$ 类推。t-测试差表示在一个语境中两个字 $x$ 和 $y$ 的结合程度;
- $a_6$ 局部字互信息[5]: 在一个句子或者段落的范围内两个字的互信息;
- $a_7$ 可计算化的语义信息[6]。

对于熟语料库处理和应用的流程如图1:

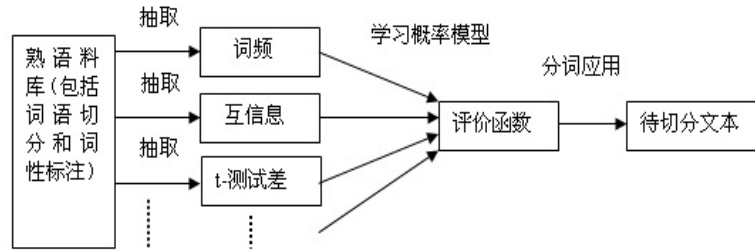


图 1: 基于统计的分词处理流程

从图1可以看出, 从语料库中抽取出来的分词要素之间是相互联系的, 将其中若干的要素组合就可以得出评价函数。在整个流程中学习这一步最重要, 直接决定了最后的分词应用结果。目前主要的评价函数有如下几种:

- $a_1$ : 一元词频法[1], 只考虑词频大小, 决定切分结果, 评价函数为

$$P(W) = \prod_{0 < i < c} P(w_i) \quad (1)$$

即一个句子中各个词概率的乘积。

- $a_4 + a_5$ : 二元汉字互信息模型[4], 评价函数为

$$md(x, y) = mi(x, y) + \alpha \cdot dts(x, y) \quad (2)$$

- $a_4 + a_6$ : 语言环境消歧模型[5], 评价函数为

$$I(x : y) = \alpha \cdot I_g(x : y) + (1 - \alpha)I_l(x : y) \quad (3)$$

- $a_1 + a_3$ : 隐马尔可夫模型[1], 评价函数为

$$P(W) = P(T)P(W|T)/P(T|W) \quad (4)$$

- $a_1 + a_2$ : 局部词频模型[2], 评价函数为

$$F(w, l) = \alpha \cdot f(w) + (1 - \alpha)f(w, l) \quad (5)$$

上面列举了目前比较常用的进行分词的一些评价函数。通过比较和分析, 不难发现其中存在着几个问题。首先我们可以看到公式(2)(3)(5)中将多个分词要素进行组合时, 无一例外的都会在评价函数中将各个分词要素进行线性叠加, 而其中的 $\alpha$ 值在各自的论文中都是通过做试验所得的经验值, 这个值的选取对最后的结果影响很大, 所以必须要具有客观性。其次对于各个分词要素的取舍组合也是比较朴素的, 在公式(2)(3)(5)中都是不超过两种分词要素的组合。为什么不可以有其它要素呢? 是否这些要素足够了? 这就是本文所说的完备性。所以一种良好的分词消歧模型应该可以更加丰富灵活地组合已有的或将有的分词要素, 这就是我们所说的可扩展性。

比较系统地提出用机器学习中的分类方法来处理交集型歧义是文献[7], 利用SVM(支持向量机)作为分类的实现手段, 设计并实现了一种特定的分词消歧。对于三字词的交集型歧义 $c_1c_2c_3$ , 其有两种切分方案, 即 $c_1/c_2c_3$ 和 $c_1c_2/c_3$ , 这两种切分方案分别称为正例和反例。利用 $c_1/c_2$ 和 $c_2/c_3$ 两个切分点处的互信息(分类特征)作为二维输入向量, 通过SVM的方法学习出最优分类面, 利用学习出来的最优分类面对待切分的交集型歧义进行切分。该文提出了一种很好的分词的思路, 但是利用SVM来实现, 有一些缺陷:

1. 分类特征选择问题: 如果固定的少数几个分类特征(即上文所说的分词要素)就能够解决分词消歧问题的话, 那么其它的优化算法也同样能够达到与SVM相同的准确率效果。事实是, 分词消歧不可能只用目前已经发现的分词要素就可以比较好地解决。
2. 高维问题: 如果选择多个分词要素, 那么, 训练SVM的输入向量的高维灾难问题将使SVM性能不如文献[7]所表述的。

3. 训练困难：由于SVM是借助二次规划来求解，而求解二次规划将涉及到 $m$ 阶矩阵的计算( $m$ 为样本的个数)，所以当训练样本数很大时，该矩阵的存储和计算将耗费大量的机器内存和运算时间。
4. 核函数选择问题：由于歧义抽取出来的要素数据是非线性可分的，所以要将非线性可分的数据映射到特征空间，在映射的过程中需要一个核函数，如何寻找针对于交集型歧义数据的核函数，目前没有可靠的理论依据。

针对于上面的问题，本文提出了利用分类数据挖掘的方式来解决上述问题。以类似于上述分词要素作为属性，以文本中字与字之间的“断”和“连”作为类别结果，从大规模的熟语料库中统计出字与字之间的属性值与字与字之间状态，通过分类数据挖掘的方法，抽象出切分的规则。最后将若干的规则作为评价函数，以体现其客观性、完备性和可扩展性。

### 3 数据记录抽取

本文主要解决的是歧义切分消解，所以本文要从大规模语料库中抽取出交集型歧义和组合型歧义，然后针对在歧义的分点处进行数据记录的抽取。上述的分词要素中 $a_1$ 和 $a_4$ 是全局量，而 $a_2, a_3, a_5, a_6$ 和 $a_7$ 为局部量。这个局部环境我们都规定为一个上下文窗口。这个窗口的大小在本文中定义为一个自然段。我们针对交集型歧义和组合型歧义的不同特点，抽取不同的用于分词的示范属性。

#### 3.1 交集型歧义

交集型歧义[8, 9]的特点是切分结果与所在语境的上下文关系比较密切，所以在属性中加入全局量的同时要尽可能加入局部量。除了上面介绍的局部量以外，再定义一个局部量属性“上下文相关度” $a_8$ 如下：

**定义 1 相对词频：**词 $B$ 相对于词 $A$ 的相对词频 $T(B, A)$ 是在所有出现词 $A$ 的自然段中出现词 $B$ 的次数总和。

**定义 2 覆盖度：**对于词 $A$ ，假设其所在的所有自然段中出现不同词的个数是 $m$ ，统计的整个语料库的不同词数量为 $n$ ，那么词 $A$ 的覆盖度 $C(A) = m/n$ 。

**定义 3** 上下文相关度: 定义词 $A$ 与词 $B$ 的相关度 $R(A, B) = T(B, A)/C(B)$ , 上下文相关度表示两个词在一个语言环境中的关联紧密程度。

在数据抽取的过程中, 数据记录的类别就是所考虑的切分点在熟语料库中是“断”还是“连”, 最终交集型歧义的属性向量是 $(a_1, a_2, a_4, a_5, a_6, a_8)$ 。那么在大规模熟语料库中的识别出所有交集型歧义后, 在每一个交集歧义的分点处都统计出属性向量和切分点状态, 最终形成了一个数据库。数据库的部分数据如下:

热电厂 供热/ 能力 八百, 7, 0, -0.0098642, -11.7119, 0, 0, 1

热电厂 供/热 能力 八百, 35, 2, 2.82271, 1.31898, 4.06044, 0, 0

热电厂 供热 能/力 八百, 2360, 0, 3.81534, 22.5175, 0, 0, 0

上面就是一个在数据库中歧义字段的真实数据。在上面的例子中的交集型歧义是“供热能力”, 其中交集的词语为“供热”、“热能”、“能力”, 在字段中的“/”表示当前考虑的切分点。接下来是六个属性量, 分别对应 $(a_1, a_2, a_4, a_5, a_6, a_8)$ 属性量的值。最后一个数据表示切分点状态, 1表示“断”, 0表示“连”。

### 3.2 组合型歧义

如果一个字串成为一个词, 且字串内部也可以拆分成多个词, 那么这个字串就被称为组合型歧义[10]。组合型歧义的切分结果与歧义字串所在语境的词法结构和词性有密切的联系。所以我们对组合型歧义只抽取组合型歧义的前驱词性和后继词性, 属性向量为 $(a_3, a'_3)$ , 即前驱词和后继词的 $a_3$ 属性值。切分点的类型与交集型歧义相同。那么在大规模语料库中统计出所有的组合型歧义的属性向量和类型, 也形成了一个数据库, 部分数据库的数据如下例:

干部/n 群众/n 一起/s 研究/v 扶贫/v 开发/v 的/u 路子/n ,n,v,0

公安局/n 破获/v 一/m 起/q 合伙/vd 抢劫/v 出租车/n 司机/n 案/Ng ,v,vd,1

组合型歧义的数据格式和交集型歧义类似。上面的例子中组合型歧义是“一起”, 在第一句句句中是连接在一起, 而在第二句中是分开的。在句子后的两个值分别表示前驱词 $a_3$ , 后继词 $a'_3$ , 最后的数量值表示切分点状态。

## 4 消歧规则挖掘

针对交集型歧义和组合型歧义的不同, 规则挖掘的方法也不相同。不同

交集型歧义之间的切分规律非常相近，所以我们把所有的数据都放在一个数据库中。然而组合切分规律却因词而异，所以对于每一个组合型歧义都只挖掘特定词的数据。

#### 4.1 数据离散化

在规则挖掘之前我们必须将连续型的数据进行离散化。我们可以看到在交集型歧义的数据中有整型和浮点型数据，所以首先要将其离散化。这里我们要用最小描述长度划分法MDLP[11]，根据信息熵的原理，对于连续型属性都需要离散化。我们以人民日报1998年1月到6月的熟语料库为例，从中搜索所有的交集型歧义，然后按照上文介绍的方法计算出所有的属性值。以t-测试差为例，离散化后的结果见表1：

表 1: t-测试差离散值表

离散化值	连续值范围
0	< -24.505
1	-24.505 ~ -8.165
2	-8.162 ~ -2.51
3	-2.51 ~ -0.52
4	-0.52 ~ 7.83
5	7.83 ~ 24.15
6	24.15 ~ 48.37
7	> 48.37

同样通过MDLP的离散化方法，最终得到各个属性离散化区间的个数如下：属性 $a_1$ 为10个区间，属性 $a_2$ 为6个区间，属性 $a_4$ 为12个区间，属性 $a_5$ 为8个区间，属性 $a_6$ 为6个区间，属性 $a_8$ 为15个区间。

#### 4.2 规则挖掘

数值离散化后就可以进行分类数据挖掘，在分类数据挖掘过程中结合中文分词的特点，需要保证挖掘规则的质量和速度。下面给出规则搜索空间和规则质量的有关定义：

**定义 4** 假设数据库中属性量个数为 $n$ ，对于每一个属性量 $a_i(0 \leq i < n)$ ，离散化后的离散值的个数为 $m_i$ ，那么完全搜索的规则空间为

$$2^n \prod_{i=0}^{n-1} m_i \quad (6)$$

按照在4.1节中我们提到的交集型歧义所有属性离散区间的个数,那么按照上述定义我们进行计算,这完全搜索的规则空间的大小为 $2^6 * (10 * 6 * 12 * 8 * 6 * 15) = 16588800$ ,搜索空间达到了千万级。由于本模型是一个开放性很强的模型,以后将会不断地添加新的属性,而且不同的离散化方法会导致不同的离散化区间个数,所以规则的搜索空间将会不断膨胀。当训练数据库更换以后要得出分词规则,利用完全搜索是不实际的。

**定义 5** 假设数据库上的一个规则 $R$ ,  $TP$ 表示被规则覆盖的记录中是符合规则类型的记录数,  $FP$ 表示被规则覆盖的记录中不符合规则类型的记录数,  $FN$ 表示不被规则覆盖的记录中符合规则类型的记录数,  $TN$ 表示不被规则覆盖的记录中不符合规则类型的记录数,那么规则 $R$ 质量定义为

$$Q = \frac{TP}{TP + FN} \cdot \frac{TN}{TN + FP} \quad (7)$$

该公式能够兼顾覆盖率和正确率,避免过分符合训练集所导致泛化的适用程度降低的问题。

根据上面的定义,传统的决策树[12]方法不但存在计算速度慢,而且会产生对训练集的过分符合,不利于规则运用到未来分词的待切分文本中。所以我们考虑到以上的两个原因,在分类数据挖掘中采用元启发式算法—蚂蚁算法[13, 14],进行数据挖掘。根据公式(6)可以看出搜索的空间比较大,在分词数据库中通常超过千万数量级,恰能发挥元启发式算法的优点,在有限的时间内找出比较好的解。利用公式(7)对搜索的规则质量进行衡量,使得规则具有较高的抽象性和概括性。

对人民日报的熟语料库利用蚂蚁算法进行分类数据挖掘,蚂蚁算法中用到的参数:蚁群中蚂蚁个数是所有属性量中的离散区间个数之和,每一个规则所必须覆盖的最低数据实例数取30,最大允许的规则可以不需要覆盖的数据实例数是全部实例数据的百分之五,蚂蚁算法的循环的代数(也就是数据挖掘的中止条件)设置为1000代。

以交集型歧义为例,挖掘出交集型消歧规则29条,对部分规则交集型歧义规则举例如下:

规则1. *If* ( $a_4 = 2$ ) *and* ( $a_2 = 1$ ) *Then Class*=“连” *Quality* 1

规则2. *If* ( $a_1 = 0$ ) *and* ( $a_8 = 0$ ) *Then Class*=“断” *Quality* 0.981481

对于每一个规则分为三个部分,If部分也就是条件部分,包含一个或者多个“属性=属性值”部分,这里的属性就是不同歧义所考虑的分词要素,而属性值都是属性经过离散化后的值或者属性的枚举量的值。Then部分是结

论部分,也就是某个切分点的状态,“连”或者“断”。Quality部分是该条规则按照公式(7)的计算值。

而组合型歧义是对每一个歧义字段进行挖掘,对于同样的语料库,我们发现了1251个组合型歧义,对于每一个歧义字段,有几条到数十条规则不等。

交集型消歧规则应用举例如下:

例1: 小明去买乒乓球拍, 乒乓球拍卖完了。

从中我们检测到交集型歧义“乒乓球拍卖”。此歧义有可能有两种切分可能,分别为“乒乓球拍/卖”和“乒乓球/拍卖”。那么我们将会两个切分可能点分别抽取该切分点的六个属性量值。经过离散化以后,“拍/卖”这个歧义点的数据分别是(0, 0, 1, 0, 0, 0),而“乒乓球/拍”这个歧义点的数据分别为(1, 1, 2, 1, 1, 0)。针对离散化后数据值匹配规则后,“拍/卖”这个歧义点匹配到规则2,所以切分开;而“乒乓球/拍”这个歧义点匹配到了规则1,将歧义点连接起来,所以最终的切分结果为“乒乓球拍/卖”。再看下例:

例2: 小明去拍卖行参观, 乒乓球拍卖完了。

在例2中也同样检测到交集型歧义“乒乓球拍卖”。同样通过数据抽取和数据离散化后,“拍/卖”这个歧义点的数据分别为(0, 1, 2, 1, 0, 1),匹配到规则1;而“乒乓球/拍”这个切分点的离散化后的数据是(0, 1, 1, 0, 0, 0),匹配到规则2,所以“乒乓球拍卖”这个歧义的最后切分情况为“乒乓球/拍卖”。

这里我们可以看到,与上下文相关的属性 $a_8$ 在这些消歧中起了决定性作用。

### 4.3 评价函数分析

上面挖掘出的规则就是本模型的最终评价函数,我们将从三个方面和第2节中所提到的评价函数进行分析和比较。

首先是客观性,也就是属性值之间函数关系。图2表示的一个训练集中交集歧义空间的分布图,

横坐标和纵坐标分别表示的属性 $a_4$ 和 $a_5$ ,图中的十字和三角表示空间中的一个歧义待切分点的切分实例,三角表示切分点为“连”,十字表示切分点为“断”。这里我们关注的就是当训练集中两个属性 $a_4$ 和 $a_5$ 满足什么样的函

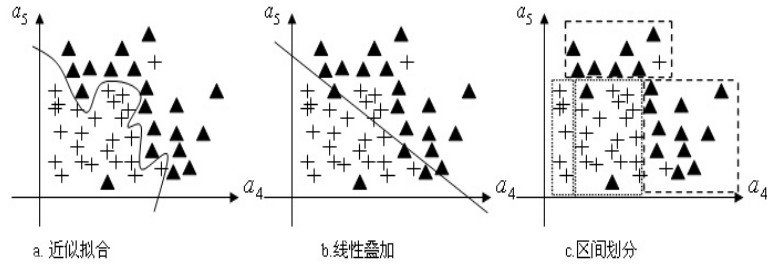


图 2: 交集歧义空间分布图

数关系时能够正确地识别切分点的切分状态。图2-a中显示的是一种理想的函数拟合,按照图中曲线的划分能够比较好地定义出 $a_4$ 和 $a_5$ 函数关系,得出比较理想的训练结果。但是这种函数利用数学的方法是基本上不可能找到的。图2-b是公式(2)的函数表示,希望将两种属性进行线性叠加,企图能够模拟图2-a中的函数曲线来划分出切分点的两种切分状态。线性叠加的方法过于简单且斜率 $\alpha$ 是不确定的,所以这种训练是很粗糙的。图2-c表示本文中所介绍的模型从六维向量 $(a_1, a_2, a_4, a_5, a_6, a_8)$ 降低到两维向量 $(a_4, a_5)$ 。将从训练集中挖掘的区间规则,是一种分段形式的函数关系。其实公式(2)也可以划归为规则区间的形式,当 $md(x, y)$ 取定一个阈值 $T$ 后,那么规则可以表示为

*If*  $(0 < a_4 < T)$  *and*  $(0 < a_5 < (T - a_4)/\alpha)$  *Then* *Class* = “断”

*If*  $(a_4 \geq T)$  *and*  $(a_5 \geq (T - a_4)/\alpha)$  *Then* *Class* = “连”

只是通过数据挖掘后得出的区间规则比公式(2)的准确度更高而已。公式(3)(5)与公式(2)类似,同样可以化为区间规则的形式。

其次是完备性,也就是考虑的属性量的个数,本文认为只要是对分词有影响的要素,不管这个影响的大小,都应该考虑。在公式(1)和(4)中是以整个句子作为考虑的对象。虽然不能划归为本模型的规则形式,但是其考虑的属性量偏少,是不可避免的缺陷。而在本模型示例中将六个属性都参与到规则挖掘中,所形成的评价规则将更加客观。

最后是可扩展性,目前的消歧问题没有彻底解决,从结果来看,说明我们并没有完全在计算机上实现分词知识。这里提出的基于分类数据挖掘的消歧模型,对于将来发现的分词属性,具有良好的扩充性。

综上所述,本文所提出的分词消歧模型可以涵盖目前所有的统计分词方法,同时也是一个良好的具有扩展性的模型。

## 5 测评与分析

本文主要处理中文分词歧义的消解问题，对于未登录词暂时不考虑。在对测试集进行分词测试之前，首先要从训练集抽取出相关的评价分词的要素，通过分类数据挖掘的手段得出规则知识。本文分词的主要流程如图3。

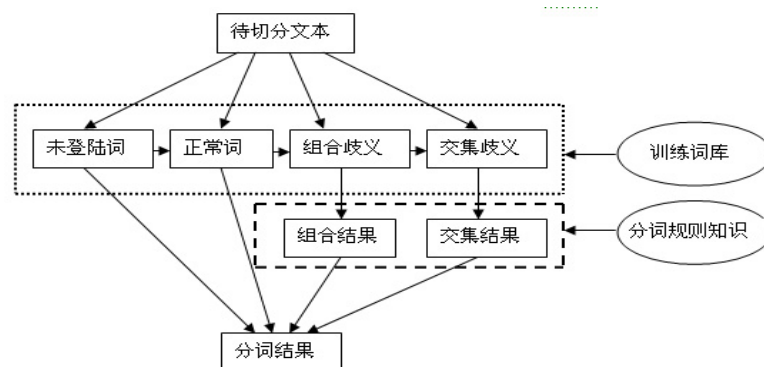


图 3: 分词流程图

对于一个待切分的文本来讲，采用正向最大匹配的方法结合训练词库以及在训练集中收集到的组合型歧义库对文本进行预处理。在文本中寻找出交集型歧义和组合型歧义。对于组合型歧义就利用在训练集中收集到组合型歧义进行实例匹配。在将待切分文本预处理成各种类型的短句后，将消歧规则应用到待切分文本的各个短句中，来确定在切分点处的最终的切分状态，最终所有切分好的短句就组成了分词结果。

我们利用两届Sighan国际分词测评比赛[15, 16]中采用的测试文本进行了测试。第一届我们测试了两个文本，PK(北京大学)，HK(香港城市大学)所提供的标准测试文本。第二届测试了四个文本，分别为PK(北京大学)，HK(香港城市大学)，AS(台湾中央研究院)，MSR(微软研究院)。首先利用训练集对测试文本进行了封闭测试，然后利用了人民日报1998年1月份到6月份的熟语料库作为开放测试的语料库，对测试文本进行开放测试。在Sighan的测评中召回率分两个部分进行比较，也就是在词库中的词语和未登录词两个部分。本文将只测试在词库中词的切分召回率 $R_{iv}$ ，不考虑未登录词的召回率 $R_{oov}$ 。

表 2: 第一届Sighan测评文本结果

Text	Word Number	OOV	OurRiv <sub>c</sub>	Rank <sub>c</sub>	OurRiv <sub>o</sub>	Rank <sub>o</sub>
PK	17194	0.069	0.983	1/8	0.984	1/6
HK	34955	0.071	0.981	2/4	0.980	1/3

表 3: 第二届Sighan测评文本结果

Text	Word Number	OOV	OurRiv <sub>c</sub>	Rank <sub>c</sub>	OurRiv <sub>o</sub>	Rank <sub>o</sub>
PK	104372	0.058	0.978	2/23	0.979	1/17
HK	40936	0.074	0.973	4/15	0.980	3/8
AS	122610	0.043	0.970	6/11	0.975	4/8
MSR	106873	0.026	0.987	6/29	0.988	2/19

在表2和表3中, Word Number表示在各自的文本中所含有的词数, OOV表示在测试文本中的词没有出现在训练文本中的比例, Our Riv<sub>c</sub>表示利用我们的方法进行封闭测试所得到分词召回率以及参加该文本封闭测试的程序个数, Rank<sub>c</sub>表示我们的召回率的在所有封闭测试程序中的排名。Our Riv<sub>o</sub>表示利用我们的方法进行开放测试所得到分词召回率以及参加该文本开放测试的程序个数, Rank<sub>o</sub>表示我们的召回率的在所有开放测试程序中的排名。

从表2和表3中可以看出, 在测试的所有的六个文本中, 在词库中的词的切分召回率都处于参加测试程序的前列, 其中我们可以看出:

1. 对于PK测试文本的开放测试召回率表现得比较突出。通过仔细分析可以看出PK的测试文本与人民日报的熟语料库的文章类型以及语言表达及其类似, 所以最后的切分结果比较好。这里可以看出, 训练文本对测试文本的最后切分结果的影响是很大的。
2. 在第二届的整个测试文本中包含两个繁体测试文本(HK和AS)和两个简体测试文本(PK和MSR)。从最终的开放测试数据来看, 我们可以发现简体测试文本的召回率基本上高于繁体测试文本的召回率。我们开放测试的语料库是简体文本, 从中可以推测出, 简体与繁体的表达和词语构成之间还是有比较大的差别。

当然在上面我们比较的只是在词库词的切分召回率, 而最终的召回率 $R = Roov * OOV + Riv * (1 - OOV)$ 。召回率分为两个部分, 一个是未登陆词的召回率, 另外一个是在词库词的召回率。两者乘以各自的分布比例, 得到最终的召回率。本文的模型中虽然没有涉及未登陆词的处理, 但是我们在处理两种歧义词的过程中并没有破坏未登陆词, 所以加上一个未登陆词的识别算法, 可以得出一个比较理想最终切分召回率, 为获取理想的F值打下一个

良好的基础。

另外值得重点指出的是，本文的这些结果，完全是基于封闭训练语料库或人民日报半年熟语料库，没有引入任何其它枚举性或者生成性的分词知识。

## 6 结语

本文的分词消歧模型是对多种当前比较经典的模型进行分析和总结后提出的，主要关心的是在中文分词中分词的评价函数，因为评价函数或者说评价的准则也就最终决定了分词的结果。而当前的分词评价函数在客观性、完备性和可扩展性方面都存在问题，主要表现在对分词要素之间的权重以及构成模式完全是经验性的，缺少科学的依据。而本文是通过数据挖掘的方法，从训练文本中提取的大量数据找出其中的规律性准则，充分发挥了分词要素之间相互关联的特点。将挖掘出的规则运用到不同文本的切分中，取得了比较好的结果。另外本模型的可扩展性很强，对于新发现的与分词有关联的分词要素能比较容易地加入到本模型中。

本模型也存在一些问题有待改进。在测试文本切分后，对切分错误的词语进行初步分析，发现比较多的情况还是组合性歧义切分错误。因为在待切分文本中统计词性信息具有一定的不确定性，且每一个词还具有多个词性，因此对于组合歧义的消歧来说，还需要引入更多的基于语义的分词属性。虽然我们目前还不知道这些分词属性是什么，但是，本文的模型完全可以为将来的进一步研究提供了一个很好的将分词知识可计算化的框架。

## 参考文献

- [1] 沈达阳, 孙茂松. 基于统计的汉语分词模型及实现方法. 中文信息学报, 15(2):96-98, 1998.
- [2] 金瑜, 陆启名, 高峰. 基于上下文相关的最大概率汉语自动分词算法. 计算机工程, 30(16):146-148, 2004.
- [3] 孙茂松, 黄昌宁, 邹嘉彦, 沈达阳, 陆芳. 利用汉字二元语法关系解决汉语自动分词中的交集型歧义. 计算机研究与发展, 34(5):332-339, 1997.
- [4] 孙茂松, 肖明, 邹嘉彦. 基于无指导学习策略的无词表条件下的汉语自动分词. 计算机学报, 27(6):736-742, 2004.
- [5] Zhang Mao-Yuan, Lu Zheng-ding, and Zou Chun-Yan. A chinese word segmentation based on language situation in processing ambiguous words. *Information Sciences, Elsevier*, 162(3-4):275-285, 2004.
- [6] 郭宏蕾, 姚天顺. 自然语言中时间信息的模型化. 软件学报, 8(6):432-440, 1997.

- [7] 李蓉, 刘少辉, 叶世伟, 史忠植. 基于svm和k-nn结合的汉语交集型歧义切分方法. 中文信息学报, 15(6):13-18, 2001.
- [8] 孙茂松, 左正平, 邹嘉彦. 高频最大交集型歧义切分字段在汉语自动分词中的作用. 中文信息学报, 13(1):27-34, 1999.
- [9] 闰引堂, 周晓强. 交集型歧义字段切分方法研究. 情报学报, 19(6):637-643, 2000.
- [10] Xiao Luo, Sun Mao-Song, and Tsou B K. Covering ambiguity resolution in chinese word segmentation based on contextual information. In *Proceedings of The 19th COLING*, 2002.
- [11] U.M. Fayyad and K.B Irani. Multi-interval discretization of continuous-valued attribute for classification learning. In *Proceedings of The 13th IJCAI*, pages 1022-1027, 1993.
- [12] J. R. Quinlan. Combining instance-based and model-based leaning. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 236-243, San Mateo, CA, 1993. Morgan Kaufmann Publishers.
- [13] M. Dorigo, V. Maniezzo, and A. Colorni. The Ant System: Optimization by a colony of co-operating agents. *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics*, 26(1):1-13, 1996.
- [14] R.S. Parepinelli, H.S. Lopes, and A.A. Freitas. Data mining with an ant colony optimization algorithm. *IEEE transactions on Evolutionary Computing*, 6(4):321-332, August 2002.
- [15] Richard Sproat and Thomas Emerson. The first international chinese word segmentation bake-off. [http://www.sighan.org/bakeoff2003/bakeoff\\_instr.html](http://www.sighan.org/bakeoff2003/bakeoff_instr.html), 2003.
- [16] Thomas Emerson. Second international chinese word segmentation bakeoff. <http://www.sighan.org/bakeoff2005/results.php>, 2005.