文章编号：1003-0077(2016)02-0026-06

一种基于特征映射的中文专家消歧方法

潘 霄[1,2]，余正涛[1,2]，郭剑毅[1,2]，毛存礼[1,2]，杨秀贞[1]

(1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 昆明理工大学 智能信息处理重点实验室，云南 昆明 650500)

摘 要：针对中文专家页面特点，以及用于消歧的基准专家页面中信息涵盖不全的问题，该文提出一种基于特征映射的中文专家消歧方法。首先，采用条件随机场模型，从基准专家页面和待消歧页面中提取出所定义的 12 维人物属性特征，并利用最大熵分类模型，结合已有消歧结果训练出各属性特征的权重；然后，针对某个专家的基准页面，计算待消歧页面与该页面的相似度，根据设定的阈值判断该页面是否单独成类，若不是单独成类，则利用特征映射，扩充该页面的属性特征，结合模糊聚类方法，得到与该页面为一类的页面。在“自然语言处理”及“机器学习”领域进行中文专家消歧实验，结果表明提出的方法能有效对中文专家页面进行消歧。

关键词：中文专家消歧；属性特征；特征映射；模糊聚类

中图分类号：TP391 **文献标识码**：A

A Chinese Expert Disambiguation Method Based on Feature Mapping

PAN Xiao[1,2], YU Zhengtao[1,2], GUO Jianyi[1,2], MAO Cunli[1,2], YANG Xiuzhen[1]

(1. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunan 650500, China;
2. Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, Yunan 650500, China)

Abstract: A Chinese expert page disambiguation method based on feature mapping is proposed according to the characteristics of the Chinese expert page. Firstly, with the help of CRFs model, 12 predefined character attributes are extracted from the standard and the candidate page, and their weights are decided by a ME classifier. Then, the page similarity is calculated to decide if the candidate page attributes should be appended Experiments on NLP and ML expert pages show the effectiveness of the proposed method in disambiguation.

Key words: Chinese experts page disambiguation; attributive character; feature mapping; fuzzy clustering

1 引言

由于专家重名和表示方式多样性的问题，导致以某一专家姓名进行检索将返回多个不属于该专家的页面，为准确区分出该专家的专家页面，须对获取到的页面进行专家消歧。通常专家消歧可以转化成专家页面的聚类问题进行解决。当前的专家消歧方法主要有以下几类：一是基于特征向量相似度的聚类消歧方法，如 Wang[1] 利用网页内容向量空间模型对专家页面进行聚类消歧，Bollegala[2] 提出利用上下文中的关键性短语相似度实现专家聚类消歧；二是基于属性相似度的聚类消歧方法，如 Cohen[3] 提出通过计算属性对间相似度实现专家聚类消歧，周晓等[4] 针对人名消歧的任务，提出基于人物属性互斥与非互斥的两阶段人名消歧的方法；三是基于特定关联关系的聚类消歧方法，如郎君[5] 提出的基于社会网络的人名重名消解，利用页面标题和上下文片断中人名的共现关系构建社会网络，并通过聚类的方法实现消歧。Tang[6] 提出的结合专家论文

收稿日期：2013-01-08 定稿日期：2014-01-05

基金项目：国家自然科学基金(61175068)；云南省软件工程重点实验室开放性基金(2011SE14)；国家教育部留学回国人员科研启动基金。

属性和论文合作关系的聚类消歧方法，选取文章标题、摘要、作者等作为特征，结合发表论文合作关系，通过基于 HMRF(Hidden Markov Random Field)的聚类方法，进行专家聚类消歧。

采用聚类的方法进行专家消歧，通常是以某个确定属于专家的页面作为基准页面，通过聚类，将与该基准页面聚为一类的页面挑选出来，作为专家页面。因此，消歧的正确性很大程度上依赖于基准页面中的信息，然而，由于页面信息量的限制以及信息更新速度较快，导致基准页面对专家信息涵盖不全，从而影响消歧的准确率。现有方法没有充分考虑基准页面的信息扩充，为解决这一问题，本文提出一种基于特征映射的中文专家消歧方法。

2　基于特征映射的中文专家消歧方法思想

基于特征映射的中文人名消歧方法的主要思想是先从基准页面和待消歧页面中提取出用于表征基准页面和待消歧页面的特征，并通过已有消歧结果得到各维特征的权重，然后，针对基准页面，利用基准页面属性与待消歧页面属性的相关性将基准页面与待消歧页面用带权重的特征表征成向量，计算待消歧页面与该基准页面的相似度，根据设定的阈值判断该基准页面是否单独成类，若不是单独成类，则寻找与该基准页面相似度最大的页面，利用特征映射的方法扩充该基准页面的特征向量，并将此页面归入该基准页面类，重复这一扩充过程，直至基准页面的特征向量不再被扩充为止，则将该基准页面与剩余的页面进行聚类，得到和该基准页面为一类的页面。该方法具体流程描述如下：

输入：基准页面为 ω，召回页面集为 $\theta=\{\theta_1,\theta_2,\cdots,\theta_n\}$，$n$ 为页面集中的总页面数，阈值 ε

(1) 分别对 ω 和 $\theta=\{\theta_1,\theta_2,\cdots,\theta_n\}$ 进行特征提取；

(2) 计算 θ_i 与 ω 的相似度 σ_i，$i=1,2,\cdots,n$；

(3) 得到 $\sigma^*=\max\limits_{1\leqslant i\leqslant n}\{\sigma_i\}$，若 $\sigma^*<\varepsilon$，则 ω 单独成一类，流程结束，若 $\sigma^*\geqslant\varepsilon$，进入下一步；

(4) 利用 σ^* 对应的召回页面 θ_k 的特征扩充 ω 的特征，扩充后的基准页面为 ω^*；

(5) $\omega=\omega^*$，$\theta=\theta-\theta_k$，θ_k 归为 ω 类，判断此时的 ω 是否还能再扩充，若能扩充，则转步骤(2)，若不能，则进入下一步；

(6) 将 ω 和 θ 进行聚类，得到 θ 中和 ω 聚为一类的页面，流程结束。

3　基于特征映射的中文专家消歧方法

3.1　特征提取与特征加权

由于中文专家页面信息中所包含的内容主要是对人物的描述，因此，选取人物相关属性作为表征基准页面与待消歧页面的特征，12 维人物属性特征定义如下，分别为人名、地名、组织机构名、职称、性别、民族、学历、毕业院校、出生日期、研究方向、获奖荣誉、承担项目。提取这些人物属性实际上是一个人物属性实体的提取问题，由于条件随机场模型[7]不需要很严格的独立性假设，可以融入丰富的特征，故其在实体抽取中被广泛运用且具有较高的准确率，因此，本文采用条件随机场模型进行人物属性实体的提取。然而每维属性特征所起的作用是不同的，还需要得到各维特征的权重，本文利用已有消歧结果，将各维特征作为分类模型的特征函数，对已知消歧结果的页面进行所属专家标记，训练出分类模型特征函数的权重，从而得到各维特征的权重，由于最大熵模型[8]可以任意加入对最终分类有用的特征，而不用顾及它们之间的相互影响，并且最大熵模型能够较为容易地对多分类问题进行建模，基于以上优点，本文使用最大熵模型训练各维特征的权重。

3.2　基准页面与待消歧页面的向量表征

在获得各维属性特征的权重后，为将基准页面与待消歧页面用向量表征出来，则需利用基准页面的属性与待消歧页面属性的相关性，也即以某个基准页面为基础，将待消歧页面的属性与该基准页面对应维的属性进行匹配，若某一维匹配成功，则该维的值为所匹配的属性的权重值，若匹配不成功，则该维的值为 0，各个待消歧页面的 12 维属性依次与基准页面进行匹配，直至把所有待消歧页面都表征为匹配结果对应的向量；基准页面的向量表征，则是根据其提取属性特征的情况而定，对于提取不到的属性特征，则对应维度的值为 0，对于能够提取出的属性特征则其对应维度的值为该属性的权重值。针对属性特征的匹配，本文采用基于《知网》的词语相似度计算方法进行匹配，参照刘群在“基于《知网》的词汇语义相似度的计算”中提出的方法[9]，综合考虑节点的共性信息和个性信息，给出如式(1)所示的义原语义相似度计算公式：

$$Sim(p_i,p_j)=\frac{Height(pnode)}{Length(p_i,p_j)+Height(pnode)} \tag{1}$$

其中，$Height(pnode)$ 表示义原 p_i，p_j 共同父节点的绝对高度。有了义原相似度计算公式后，两个词语 W_1 和 W_2 的相似度为各个概念的相似度的最大值，计算公式如式(2)所示。

$$Sim(W_1,W_2)=\max_{i=1,\cdots,n,j=1,\cdots,m} Sim(S_{1i},S_{2j}) \quad (2)$$

其中，$S_{11},S_{12},\cdots,S_{1n}$ 为 W_1 的 n 个概念，$S_{21},S_{22},\cdots,S_{2m}$ 为 W_2 的 m 个概念。两个概念语义表达式的整体相似度为式(3)。

$$Sim(S_1,S_2)=\sum_{i=1}^{3}\beta_i Sim_i(S_1,S_2) \quad (3)$$

其中，$Sim_1(S_1,S_2)$ 为“上下位”义原关系相似度，$Sim_2(S_1,S_2)$ 为其他独立义原相似度，$Sim_3(S_1,S_2)$ 为“符号”义原相似度。通过大量统计发现相似度为 0.7 以上的两个词语在词义上较为接近，因此定义 $Sim(W_1,W_2)\geqslant 0.7$ 时，认为对应维度上的属性匹配成功。

3.3 基准页面特征映射

在将基准页面和待消歧页面表征成属性权重值构成的向量后，需要通过特征映射的方法，借助待消歧页面属性特征对基准页面的属性特征进行扩充。首先是计算所有待消歧页面与基准页面的相似度，本文通过常用的余弦相似度来进行相似度的计算，公式如式(4)所示。

$$Sim(a,b)=\frac{\sum_{i=1}^{12}x_{ai}x_{bi}}{\sqrt{\sum_{i=1}^{12}x_{ai}^{\ 2}}\sqrt{\sum_{i=1}^{12}x_{bi}^{\ 2}}} \quad (4)$$

其中，a 为基准页面向量，b 为待消歧页面向量，x_{ai} 与 x_{bi} 为向量中对应维度的值。然后找到所有相似度中的最大值，通过设定的阈值判断该基准页面是否单独成类。由大量统计得出当相似度大于 0.5 时，两个页面可以成一类，因此定义当 $\max_{1\leqslant i\leqslant n}Sim_i(a,b)\geqslant 0.5$（$n$ 为页面数量）时，认为该基准页面不是单独成类。若不是单独成类，则找出对应最大相似度的待消歧页面，将基准页面与该待消歧页面的对应维度的属性特征进行比较。如果出现待消歧页面具有而基准页面不具有的属性特征，则用该待消歧页面对应维度的权重值替换基准页面对应维度的 0，扩充基准页面的特征，特征映射完成后，就将该待消歧页面归入此基准页面类，并从待消歧页面集中移除，使用映射后的基准页面向量与其余的待消歧页面计算相似度，再次寻找相似度最大的待消歧页面，采用同样的方法实现待消歧页面到基准页面的特征映射，直至基准页面的特征不再被扩充为止。

3.4 模糊聚类分析

3.4.1 模糊相似矩阵构建

经特征映射后的基准页面向量与各待消歧页面匹配特征后形成的向量作为行向量，则形成初始矩阵。为得到与基准页面为一类的文本，本文采用模糊聚类[10]的方法解决这一问题，首先是进行模糊相似矩阵的构建，基于已有的初始矩阵，设论域 $U=\{x_1,x_2,\ldots,x_n\}$，其中行向量 $x_i=\{x_{i1},x_{i2},\ldots,x_{im}\}(m=12)$，模糊相似矩阵中的各元素为 $r_{ij}=R(x_i,x_j)$，表示 x_i 和 x_j 的相似程度，如式(5)所示。

$$r_{ij}=\exp\left\{-\sum_{k=1}^{m}|x_{ik}-x_{jk}|\right\} \quad (5)$$

3.4.2 确定最佳聚类阈值

得到模糊相似矩阵后，就可基于模糊相似矩阵进行模糊聚类，在模糊聚类分析中，对于各个不同的阈值 $\lambda\in[0,1]$，对应着不同的聚类结果，找出最佳聚类阈值 λ，此时 λ 对应的聚类结果就是最合理的聚类结果。本文采用 F 统计量确定 λ 的最佳值。

设 $U=\{x_1,x_2,\ldots,x_n\}$ 为页面全体，$x_j=(x_{j1},x_{j2},\cdots,x_{jm})$，其中 x_{jk} 为描述页面 x_j 的第 k 个特征的数据 $(k=1,2,\cdots,m)$。设 r 为对应于 λ 值的类数，n_i 为第 i 类页面的个数，记 $\overline{x_{ik}}=\frac{1}{n_i}\sum_{j=1}^{n_i}x_{jk}(k=1,2,\ldots,m)$ 为第 i 类页面的第 k 个特征的平均值。记 $\overline{x_k}=\frac{1}{n}\sum_{j=1}^{n}x_{jk}(k=1,2,\ldots,m)$ 为全体样本第 k 个特征的平均值。

引入 F 统计量，如式(6)所示。

$$F=\frac{\sum_{i=1}^{r}n_i\sum_{k=1}^{m}(\overline{x_{ik}}-\overline{x_k})^2/(r-1)}{\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{m}(x_{ik}-\overline{x_{jk}})^2/(n-r)}\sim F(r-1,n-r) \quad (6)$$

它服从自由度为 $r-1$、$n-r$ 的 F 分布，其分子表征类与类间的距离，分母表征类内元素间的距离。因此 F 值越大，说明类与类之间的距离越大，则聚类就越好。如果 $F>F_\alpha(r-1,n-r)(\alpha=0.05)$，则根据数理统计方差分析理论可知类与类之间差异显著，说明聚类比较合理；再在满足 $F>F_\alpha(r-1,n-r)$ 的所有情形中，取差值 $F-F_\alpha$ 最大者的 F 所对应的 λ 作为最佳 λ 值，其所对应的聚类即为最佳聚类。

4　实验结果及分析

4.1　专家消歧数据集准备

对于实验数据集的准备，本文采用以下方式进行：首先从万方平台及与“自然语言处理”和“机器学习”领域相关的会议网站选取“自然语言处理”和“机器学习”领域专家各 150 人，利用 Google API 通过检索专家的姓名收集搜索引擎返回的前 10 个页面形成实验数据集，并选择 10 个页面中检索排序位于第一的页面作为该专家的基准页面。数据集基本情况如表 1 所示。

表 1　专家消歧实验数据集

领域	专家数	召回页面总数	拥有超过一个专家页面的专家数
自然语言处理	150	1 500	112
机器学习	150	1 500	97
	专家页面数	召回页面平均含有词语数	基准页面平均含有词语数
自然语言处理	452	437	425
机器学习	433	425	408

由表 1 中可以看出，以专家姓名进行检索所召回的页面中，有一半以上的页面并不属于该专家，通过对数据集的分析，发现这些不属于专家的页面中，一部分是属于与专家同名的人，一部分是与专家不相关且非描述人物信息的页面，可见在通过搜索引擎返回专家页面的过程中进行专家消歧具有重要的意义。同时，在数据集中，基准页面平均含有词语数略低于召回页面平均含有词语数，说明在包含的信息量上，有的基准页面可能要比召回页面少。为进一步说明基准页面涵盖信息不全的问题，本文对基准页面和召回页面中能够提取出各维特征的页面占各自页面集总数的比例分别进行了统计，结果如表 2 所示。

表 2　含有各维特征页面所占比例

页面类型	人名	地名	组织机构名	职业
基准页面/%	98.7	61.3	92.6	91.7
召回页面/%	98.2	87.5	85.1	80.4
	职务	获奖荣誉	性别	民族
基准页面/%	93.1	57.6	75.2	78.6
召回页面/%	95.2	88.3	89.6	90.7

续表

	学历	毕业院校	出生日期	承担项目
基准页面/%	95.5	95.1	71.3	96.3
召回页面/%	84.3	82.8	85.4	76.4

由表 2 中可以看出，对于人名和职务这两类属性特征，基准页面和召回页面大多都能涵盖，而对于组织机构名、职业、学历、毕业院校和承担项目这五类属性特征则基准页面涵盖面更广一些，但是，对于地名、获奖荣誉、性别、民族、出生日期和承担项目这六类属性特征的涵盖面基准页面却不如召回页面，属性特征涵盖面的不足很可能导致消歧错误的产生。

4.2　不同特征专家消歧对比实验

为验证利用提出的 12 维人物属性特征进行专家消歧的效果，在不进行基准页面特征映射的条件下，实验将使用 12 维属性特征作为特征进行聚类的方法与文献[2]中利用关键词相似度实现专家聚类消歧的方法进行了对比，但在这一实验中先忽略各维特征的重要程度，即不赋权重，实验的评价指标为召回率(R)，准确率(P)和 F 值(F)，公式如(7)～(9)所示。

$$R=\frac{\text{正确消歧的页面数}}{\text{实际的专家页面数}} \tag{7}$$

$$P=\frac{\text{正确消歧的页面数}}{\text{归类出的专家页面数}} \tag{8}$$

$$F=\frac{2PR}{P+R} \tag{9}$$

实验结果如表 3 所示。

表 3　不同特征消歧对比实验

关键词			人物属性特征(不赋权)		
R/%	P/%	F/%	R/%	P/%	F/%
82.2	80.1	81.2	85.2	83.5	84.3

从表 3 中数据可以看出，使用人物属性作为特征的聚类效果优于使用词频作为特征的聚类效果，所以，本文定义的 12 维人物属性特征能有效进行专家消歧。

以上实验是在各维属性特征等权重的条件下进行的，也即忽略了各维属性特征对消歧效果产生影响的程度不同，为证明对各维特征赋权重后的效果，实验将利用已知消歧结果得到的各维特征权重赋予各维特征，并和等权重的效果进行对比，对比结果如

表4所示。

表4 权重因素对比实验

人物属性特征(不赋权)			人物属性特征(赋权)		
R/%	P/%	F/%	R/%	P/%	F/%
85.2	83.5	84.3	87.7	85.8	86.8

从表4中数据可以看出,对属性特征赋权重后的效果优于不赋权重的效果,可见考虑各维特征的对消歧的不同影响程度能有效提高消歧的召回率,准确率和F值。

4.3 特征映射对比实验

为验证特征映射方法的效果,将本文提出的基于特征映射的方法与文献[4]中的两阶段人名消歧方法和文献[6]中的基于HMRF的聚类消歧方法进行了对比,实验结果如表5所示。

表5 特征映射对比实验

两阶段消歧方法			HMRF			特征映射方法		
R/%	P/%	F/%	R/%	P/%	F/%	R/%	P/%	F/%
88.2	86.9	87.5	90.7	88.2	89.4	92.1	90.8	91.5

从表5中数据可以看出,相比于不进行特征映射的方法,基于特征映射的方法使得召回率、准确率和F值均有提高。

为验证数据集规模对消歧效果的影响,将本文提出的基于特征映射的方法与两阶段人名消歧方法和基于HMRF的聚类消歧方法在不同规模数据集上达到的F值进行了对比,实验结果如图1所示。

从图1中可以看出,随着数据集规模的不断扩大,特征映射方法的F值在0.915附近波动,未呈现出下降趋势,且在不同的数据集规模下,特征映射方法的F值都高于其他两种方法,说明在不同的数据集规模下特征映射方法都能取得较好的效果,但其他两种不进行特征映射的方法的F值却随着数据集规模的扩大而下降,这是因为数据集规模越大,其基准页面涵盖信息不全的问题就越凸显,所得到的消歧效果就会越差。

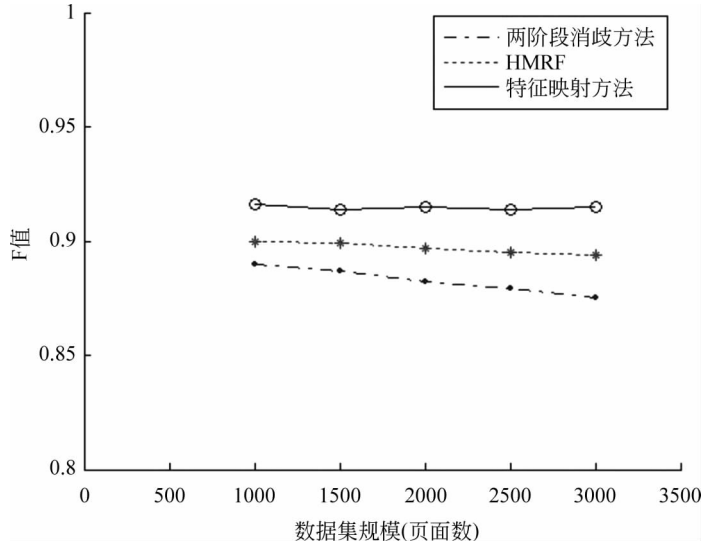

图1 不同数据集规模对比试验

5 结语

针对中文专家页面特点,以及用于消歧的基准专家页面中信息涵盖不全的问题,本文提出一种基于特征映射的中文专家消歧方法。该方法充分考虑了用于消歧的特征的选取,以及各维特征权重的确定,并且利用召回页面的特征对基准页面特征进行了扩充,实验证明所提出的方法取得了较好的消歧效果。下一步的工作,将考虑如何利用中文专家页面间的关联关系进行专家消歧,进一步提高消歧的效果。

参考文献

[1] Houfeng Wang, Zheng Mei. Chinese Multi-document Person Name Disambiguation [J]. High Technology Letters, 2005, 11(3): 280-283.

[2] Bollegala D, Matsuo Y, Ishizuka M. Disambiguating Personal Names on the Web Using Automatically Ex-

tracted Key Phrases[J]. Frontiers in Artificial Intelligence and Applications, 2006: 553-557.

[3] Cohen W, Ravikumar P, Fienberg S. A Comparison of String Distance Metrics for Name-matching Tasks [C]//Proceedings of the IJCAI Workshop on Information Integration on the Web, Acapulco, Mexico, 2003: 73-78.

[4] 周晓, 李超, 胡明涵, 等. 基于人物互斥属性的中文人名消歧[C]// 第六届全国信息检索学术会议, 2010.

[5] 郎君, 秦兵, 宋巍等. 基于社会网络的人名检索结果重名消解[J]. 计算机学报, 2009,(7): 1365-1375.

[6] Jie Tang, Limin Yao, Duo Zhang. A Combination Approach to Web User Profiling[J]. ACM Transactions on Knowledge Discovery from Data , 2010, 5(1): 2.

[7] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]// Proceedings of the 18th International Conference on Machine Learning, Wil-liamstown, USA, 2001: 282-289.

[8] Liyan Zhang. A Chinese Word Segmentation Algorithm Based on Maximum Entropy [C]// Machine Learning and Cybernetics (ICMLC), 2010 International Conference on. IEEE, 2010(3): 1264-1267.

[9] 刘群, 李素建. 基于《知网》的词汇语义相似度计算[J]. 中文计算语言学, 2002, 7(2): 59-76.

[10] Botía J F, Isaza C, Kempowsky T, et al. Automaton based on Fuzzy Clustering Methods for Monitoring Industrial Processes[J]. Engineering Applications of Artificial Intelligence, 2012, 4(26): 1211-1220.

潘霄(1988－),硕士,主要研究领域为信息检索、自然语言处理。
E-mail: 975178774@qq.com

郭剑毅(1964－),硕士,教授,硕士生导师,主要研究领域为自然语言处理、信息抽取、机器翻译。
Email:740446651@qq.com

余正涛(1970－),通信作者,博士,教授,博士生导师,主要研究领域为自然语言处理、信息检索、机器翻译。
Email: ztyu@hotmail.com

2016 中国中文信息学会战略研讨会在海口成功召开

当前已经进入以互联网和大数据为主要标志的海量信息时代。计算机和互联网技术的快速发展对中文信息处理技术提出了许多新的挑战。

继 2010、2012、2014 年学会战略研讨会之后，2016 年 5 月 5－6 日，中国中文信息学会在海口成功举办了 2016 中国中文信息学会战略研讨会，承办单位海南师范大学。本次会议邀请了来自 University of Illinois at Chicago 的 Liu Bing 教授、上海证券交易所总工程师白硕研究员及来自国内重点高校、研究所等十几位特邀讲者进行了大会发言，同时邀请了四十多位来自全国各地的中文信息处理领域的专家学者围绕我国中文信息处理事业当前发展状况和今后发展的战略机遇进行了深入的研讨。本次大会的主题是“语言智能处理技术发展趋势”。