硕士学位论文

结合共指消解的跨文档中文人名消歧研究

THE STUDY ON CROSS-DOCUMENT CHINESE PERSON NAME DISAMBIGUATION WITH COREFERENCE RESOLUTION

刘杰

哈尔滨工业大学

2013 年 12 月

国内图书分类号：TP391.3　　学校代码：10213
国际图书分类号：621.3　　密级：公开

工学硕士学位论文

结合共指消解的跨文档中文人名消歧研究

硕士研究生：刘杰
导　　师：徐睿峰副教授
申请学位：工学硕士
学　　科：计算机科学与技术
所在单位：深圳研究生院
答辩日期：2013 年 12 月
授予学位单位：哈尔滨工业大学

Classified Index: TP391.3
U.D.C: 621.3

Dissertation for the Master Degree in Engineering

THE STUDY ON CROSS-DOCUMENT CHINESE PERSON NAME DISAMBIGUATION WITH COREFERENCE RESOLUTION

Candidate: Jie Liu
Supervisor: Associate Prof. RuiFeng Xu
Academic Degree Applied for: Master of Engineering
Speciality: Computer Science and Technology
Affiliation: Shenzhen Graduate School
Date of Defence: December, 2013
Degree-Conferring-Institution: Harbin Institute of Technology

摘 要

随着互联网的飞速发展，如何从爆炸式增长的信息中高效地找到自己所需信息成为信息检索研究的重要目标。其中，面向人名的检索有着非常广泛的应用。但是在中文互联网环境中，人名重名的现象非常严重，这给面向人名的检索带来了巨大困难。为此，人名消歧的研究近年来成为信息检索领域的重要课题。

分析显示，人名带来的歧义性既可能来自于文档内代词导致的共指歧义，也可能来自于多个文档之间对应于不同实际个体的重名歧义。因此，中文人名消歧包括文档内人名共指消解和跨文档人名重名消歧。共指消解的典型方法中，基于规则的方法可移植性比较差；而基于统计方法能够获得准确率和召回率的平衡，但对训练数据依赖很大。在跨文档重名消歧研究中，基于人名上下文词语特征的方法因缺乏消歧需要的知识而遇到消歧性能的瓶颈；而利用社会网络等外部知识的方法则受到所使用外部知识的限制而很难进一步提升消歧性能。

为此，本文进行了以下研究。第一、本文研究了通过结合人名构成规则和人名出现的特点改善人名识别结果的方法。第二、针对文档内的共指消解问题，设计实现了一种结合汉语语言规则和统计学习的方法，对候选名词短语对是否存在共指关系进行判定，实现文档内共指消解。该方法在CoNLL2012共指消解中文数据集上达到评价指标 0.651 的成绩。第三、在应用共指消解方法确定人名准确上下文的基础上，提出了一种结合百科知识和利用互联网检索验证的跨文档人名消歧方法。该方法在 CIPS-SIGHAN2012 中文人名消歧数据集上达到准确率 82.4%，召回率 83.4%的性能。

本文的贡献主要包括：第一，本文设计实现了一种有效地结合规则和统计的共指消解方法，该方法在 2012 年的 CoNLL 中文共指消解国际评测中获得国际第四和国内第二的成绩；第二，本文提出的利用百科知识的方法可以缓解实体信息不完整的问题，能够更加精确地衡量实体相似度，提高人名消歧的准确率，而利用互联网验证的方法则缓解了知识短缺问题，提高了人名消歧的召回率；第三、本文提出的结合共指消解的跨文档人名消歧方法能够更好地消除人名歧义。

关键词：中文人名消歧；共指消解；百科知识；互联网验证

Abstract

With the rapid development of the Internet, how to obtain the needed information effectively from explosive growing information has become the goal of information retrieval study, while the person name oriented searching plays an important role. However, in Chinese Internet environment, the phenomenon of person name repetition is serious, which brings many difficulties to person name oriented searching. Therefore, the study on person name disambiguation has become an important topic in information retrieval research.

The observations show that the person name ambiguity attributes to in-document co-reference ambiguity and cross-document entities ambiguity. Thus, the Chinese person name disambiguation includes in-document co-reference resolution and cross-document person name disambiguation. The existing works on Chinese co-reference resolution are camped into rule-based and statistic-based approach. In which, rule-based approach achieves good precision but its portability is unsatisfactory. The statistic-based approach normally achieves good balance between precision and recall, but its performance highly relies on the training data. In the cross-document person name disambiguation approaches, the one based on person name contextual features has the performance bottleneck attribute to the lack of needed knowledge while the one based on external knowledge, such as social network, has the difficulty to further improve the disambiguation performance since the limitation of external knowledge.

This study investigated several techniques to improve the performance of Chinese person name disambiguation. Firstly, the method for improving the Chinese person name recognition is developed which incorporates the constitutive rules and occurrence features of person names. Secondly, target to in-document co-reference resolution, the method incorporates the Chinese linguistic rules and machine learning is investigated to determine whether the candidate noun phrase pair has the co-reference relationship. The method achieved official score of 0.651 on the CoNLL2012 Chinese co-reference resolution dataset. Thirdly, after applying the co-reference resolution method to identify the accurate context of person names, a cross-document Chinese person name disambiguation method which leverages the encyclopedia knowledge and uses the internet verification is proposed. This method achieved 82.4% precision and 83.4% recall on CIPS-SIGHAN 2012 Chinese person name disambiguation dataset.

The contributions of this study may be summarized as below. Firstly, a Chinese co-reference resolution method incorporating rule-based and

statistic-based techniques is developed. This method is ranked 4th in the world and 2th in China in the CoNLL 2012 Chinese co-reference resolution evaluation. Second, a cross-document Chinese person name disambiguation method is investigated in which the encyclopedia knowledge is leveraged to solve the problem of incomplete description information for estimating the similarity between entities accurately. This is helpful to improve the precision of person name disambiguation. Furthermore, the internet verification is adopted to decrease the influence of the lack of the entity's information. As the result, the recall is improved. Thirdly, the proposed incorporation of in-document co-reference resolution and cross-document person name disambiguation has shown its good performance.

Keywords: person name disambiguation, co-reference resolution, encyclopedia knowledge, Internet verification

目 录

摘 要.....I
ABSTRACT.....II
第 1 章 绪 论.....1
1.1 课题背景及研究的目的和意义.....1
1.2 国内外相关技术发展现状.....2
1.2.1 中文人名识别.....3
1.2.2 共指消解.....4
1.2.3 人名消歧.....5
1.3 本文的主要研究内容和组织结构.....8
第 2 章 中文人名识别的改进方法.....10
2.1 引言.....10
2.2 基于后处理规则的中文人名识别.....10
2.3 文档内识别结果一致化方法.....14
2.4 实验结果和分析.....15
2.5 本章小结.....16
第 3 章 文档内中文共指消解.....17
3.1 引言.....17
3.2 候选名词短语提取方法.....17
3.3 共指关系确定方法.....18
3.3.1 基于统计的共指关系确定方法.....19
3.3.2 基于规则的共指关系确定方法.....21
3.4 共指链生成方法.....24
3.5 实验结果和分析.....25
3.5.1 候选名词短语提取性能实验.....26
3.5.2 共指关系判断实验.....26
3.6 本章小结.....29
第 4 章 跨文档中文人名消歧.....30
4.1 引言.....30
4.2 跨文档人名消歧问题分析.....31
4.3 基于互联网的跨文档人名消歧的框架.....33

4.4 基于百科知识的中文人名消歧 35
4.4.1 百科数据获取与整理 35
4.4.2 利用百科知识的人名消歧 36
4.5 结合互联网验证的中文人名消歧 37
4.5.1 查询关键词构造及检索结果处理 38
4.5.2 基于互联网检索结果的人名消歧验证 38
4.6 实验结果及分析 40
4.7 本章小结 45
结　论 46
参考文献 48
附录一 53
攻读硕士学位期间发表的论文及其它成果 55
哈尔滨工业大学学位论文原创性声明和使用权限 56
致　谢 57

第 1 章 绪　论

1.1 课题背景及研究的目的和意义

随着计算机科学技术的发展和信息化程度的提高，人们面临一个如何从爆炸式增长的信息洪流中高效地找到自己想要的有用信息的难题。人名作为人的标示，常常作为信息查找的关键点，因此在信息查找中有着极其重要的作用。人名重名现象严重，给使用人名查找信息带来了非常大的困难，如何消除人名的歧义成为一个非常值得研究的课题。

如何快速和准确地找到想要的信息，一直是研究热点。搜索引擎使得人们能够通过检索关键词，将要查找的信息的范围缩小到与该检索关键词相关的文档集中，有效地提高了人名查找信息的效率。人名作为人的一个符号，常常被作为查询关键词来查找具有该人名的个体的信息。人名重名现象非常严重，使得人名的歧义性非常大，给基于人名的信息查找带来了极大的挑战。因此，如何有效地消除人名的歧义就成为了一个值得着重研究的课题。

中文人名用字比较集中，众多的人们共用着比其小几个数量级的人名，因此，中文人名有着极高的歧义性，给基于人名查找相关个体信息带来了非常大的困难。在百度中检索“赵磊”，前 40 个结果中包括 10 个不同的实体：中国首席男模、华谊兄弟时尚掌门人、西南政法大学民商法学院副教授、记者、北医三院、清华大学公共管理学院博士后、二手房中介、作家、北京市发改委副主任、国家拳跆中心副主任。由公安部全国公民身份号码查询服务中心提供的数据，有 290, 607 个“张伟”。

搜索引擎允许用户使用相关关键词来查找想要的信息，在使用人名作为检索关键词中，搜索引擎并不对检索结果中的人名进行细致的分析，只是返回与检索关键词最相关的若干项。由于每个实体在现实中具有不同的知名度，那么，同名个体在互联网上出现的频繁程度差别极大。这样，搜索引擎基于词语相关度的信息度量策略，将使得知名度略低的实体信息淹没于海量的与其同名且知名度更高的实体的信息中。例如，在百度中检索“张家辉”，前 5 页都是指向实体“香港演员张家辉”，而与其同名的“西南大学讲师张家辉”就很少出现。

本课题研究的目的就是要通过充分利用各种特征和知识，包括合理利用文档内特征和外部知识来缓解中文人名消歧过程中的知识短缺问题，使得人名消歧的性能得到提升。

本课题的研究在理论和应用上都有重要的意义。在理论上，一方面，本文将共指消解技术应用到人名消歧中，对准确定位人名上下文起了非常大的作用；另一方面，本文提出的利用互联网来缓解人名消歧过程中知识短缺问题的思路对自然语言处理其它相关问题的解决有借鉴意义。在应用上，本文提出的中文人名消歧方法提升了人名消歧的性能，减轻了中文文本分析中人名带来的分析困难。一方面，中文人名的歧义问题很好地解决，有助于信息抽取和知识挖掘等应用取得更好的结果；另一方面，随着社会化媒体和社交网络的兴起，面向人和人之间的关系的研究有着巨大的商业价值，人名消歧性能的提升有助于在这些应用中去除重名带来的干扰，从而更加准确地确定相关的人物个体及其相关的信息。

1.2 国内外相关技术发展现状

在进行人名消歧之前，必须先完成人名识别，人名识别是人名消歧的基础，其性能对人名消歧影响巨大。如：“金振民院士和高山教授应邀来我院作学术指导”，如果不能识别出其中的“高山”为人名就会在对人名“高山”的人名消歧时将其漏掉。人名作为命名实体的一种，其识别包括确定人名边界和人名判定。确定人名边界就是确定人名是由哪几个字词组成。如：“朱芳勇闯三关”，必须根据上下文判断其中人名是“朱芳”还是“朱芳勇”。人名判定就是判定某个字串是否为人名。如：“他不幸从高山上摔下”，在确定词边界“高山”后，还要判断其是否为人名。另外，人名的识别是信息处理的一项关键基础技术，在信息检索、信息抽取和机器翻译等领域广泛应用。

进行人名消歧目的是要将分析同样的人名字串是否指向相同的个体，但是，相同的个体也可以使用不同的字符串来表示。在句子“第三纵队纵队长韦拔群紧紧握着邓小平的手，感动地说：‘邓政委, 辛苦了。’”中，“邓小平”和“邓政委”虽然不是相同的字串，但却指向相同的个体，即这两个名词短语存在共指关系。为了能够消除指向人的名词短语之间的歧义，必须将文档内的名词短语之间的共指关系分析清楚。

互联网飞速发展使其成为人们获取信息的一个非常重要的途径。在互联网上不仅有海量的网页，而且有大量的人工标注和组织的信息。这些信息都提供了非常丰富的知识，挖掘和利用相关的知识将有助于人名消歧的研究。如：“两会代表黄河语录”，如果能够从互联网中挖掘到知识“两会”是“全国人民代表大会”和“中国人民政治协商会议”的简称，那么将容易做

出判断“全国人大代表黄河”与上面提到的黄河很可能是指同一个人。

由前面的背景介绍可知，人名引起的歧义既可以出现在文档内也可以出现在包含相同人名字串的文档之间。因此，人名消歧包括文档内的共指消解和跨文档的人名消歧，而这两个子问题都离不开人名识别。围绕人名消歧，本节将针对这三个问题，分别介绍国内外技术发展现状，并对其做简要的分析。

1.2.1 中文人名识别

人名是命名实体的一种，除人名外，命名实体还包括地名、组织机构名、时间、地点、日期、货币和百分比。相关的评测有 MUC（Message Understanding Conference）和 IREX（Information Retrieval and Extraction Exercise）等。

目前命名实体识别的方法主要包括：基于规则的方法、基于统计的方法和规则与统计相结合的方法。

基于规则的方法通常利用标注的语料，使用特定的方法，自动或手动提取规则，用这些规则来过滤待分析的字串，对于满足规则的字串，将其识别为人名。孙茂松等人提出了利用人名组成规则以及人名旁边常出现的指界动词等对人名进行自动识别的方法，实验表明，人名识别召回率达到了 99.77%[1]。张俊盛等人对人名识别实验中被遗漏和误判的原因进行了深入的分析，对进一步提升人名识别性能的研究给出了建议[2]。基于规则的人名识别方法的性能主要依赖于识别时使用规则的数量和质量。数量越大的规则集合能够覆盖人名出现的情况越多，使得人名识别的召回率越高。质量越好的规则集合表示集合中规则的准确率越高，使得利用这些规则的人名识别准确率越高。规则的数量和质量都直接关系到基于规则的人名识别方法的性能。而总结出满足需求的这样的人名识别中规则难度往往很大：一方面，由于规则需要直接或间接的人工参与，很难对所有的规则进行总结，使得实际总结的规则对人名出现的情况的覆盖度不够高，进而利用此规则集的人名识别召回率不够高；另一方面，为了提高规则的质量，规则的复杂度就会很高，给总结规则带来了很大的挑战。总之，单纯利用规则的方法进行人名识别很难达到很好的性能。

基于统计的方法通常需要对语言建立相应的模型，利用机器学习技术，来做出更准确的判断。黄德根等人通过衡量候选人名与其周围信息以及候选人名中用字内部这两方面的关联程度来做人名识别，准确率和召回率都得到了提高，进而提升了人名识别性能[3]。Li 等人提出了一种集成两种不同的条

件随机场模型为一种混合模型的中文人名识别方法[4]。基于统计的人名识别方法由于很少的人工参与，使得人名消歧系统能够快速地构建。然而，基于统计的人名识别方法受到训练语料的影响很大，而且学习到的人名识别模型对于与训练语料相似的情况能够取得很好地人名识别性能，对于与训练语料相似度很低的情况则往往性能不佳。

目前主流的方法是用二者结合的策略，通过统计方法进行基本的识别，再用规则对特殊情况作一些矫正，使提高整体识别性能。Wu 等人提出了一种将现有的人工总结的知识与基于统计的命名实体模型相结合的混合方法，在该方法中，人工总结的知识的引入，不仅提高了人名识别的性能，而且缓解了可用于训练的人工标注数据有限的问题。其中，命名实体的特征词被用来对候选的命名实体词语进行过滤，同义词词林中的同义词用来平滑统计模型中因数据稀疏带来的概率为零的概率[5]。Chen 等人提出了一种结合词表和 N 元语法模型的中文人名识别方法[6]。

1.2.2 共指消解

两个名词短语指向相同的实体的语言现象叫做共指。共指的使用可以简化表达，使听者或读者能够快速捕获到更加重要的信息。但是，共指的存在给计算机实体识别和处理带来了很大困难。为此，共指消解研究的目的就是要对文本中对应于一个实体的各种表达和指代进行分析。目前，共指消解的方法主要有两种：基于规则的方法和基于统计的方法。

基于规则的方法主要是利用不同的信息，经过人工整理和提取规则，利用规则对共指现象进行判定。Hobbs 利用句法分析树，提出了一种基于句法分析树的共指消解方法，该方法通过在句法分析树上查找代词的先行词来找出共指的名词短语对[7]。Brennan 等人提出了一种中心方法的形式化来对对话中的注意点结构进行建模，并将其用作链接对话上下文和对应的代词的基础[8]。Mitkov 等人针对共指消解系统总是使用外部知识的情况，提出了一种鲁棒性较好的、知识需求量少的共指消解方法[9]。

基于统计的方法主要是利用人工标注的语料，从中人工地或是自动的总结或学习出消解共指的模型。有的研究将共指消解问题看作是名词短语的聚类问题，共指的名词短语被聚到同一个簇中，从而达到消解共指的目的；也有的研究将共指消解问题看作是名词短语对的共指关系分类问题，该方法以候选的名词短语对作为对象，训练好的分类器对候选名词短语对之间的共指关系进行二分类，即判断共指或不共指。

基于聚类的统计共指消解方法，将候选的名词短语表示成特征向量，利用特征衡量候选名词短语共指的可能性，基于候选名词短语之间的共指可能性对它们进行聚类。Cardie 等人为了进行共指消解，以每个候选名词短语作为待聚类点，提取特征将名词短语表示成向量，对向量进行聚类[10]。Lee 等人针对采用特征向量表示名词短语的聚类方法容易使高准确率的特征淹没于低准确率的特征中，从而使消解性能变差的问题，提出了一种逐层过滤的方法[11]。

基于聚类的方法需要对名词短语做特征表示，由于各个名词短语所处的上下文不同，要反应的特点也不相同，为了能够衡量任何两个名词短语之间的共指关系，需要将所有的名词短语使用相同的特征，为了能够覆盖所有名词短语的特点需要很多的特征。但是，对于某一个名词短语或者某两个名词短语之间，可能仅仅需要少数的特征就能够很好的完成共指消解，在众多的特征下，原本的共指关系可能被弱化了，给共指关系的判断带来困难。因此有人提出了将共指消解问题看作候选名词短语对的分类问题来减轻这种困难。McCarthy 等人提取了特征，使用决策树对候选对进行分类[12]。

此外，Hamidreza 等人从未标注文档中学习词语的相关度，然后在共指消解过程中加入了词语相关度信息，得到了比较好的共指消解性能[13]。Rahman 等人提出了一种混合共指消解方法，该方法将基于对名词短语排序的方法与名词短语对共指关系判定的方法结合起来形成的[14]。

1.2.3 人名消歧

重名使得人名的出现带有歧义，人名消歧就是要消除文档之间人名的歧义。对于需要消歧的人名，称其为待消歧人名。文档可以被看作是字串，待消歧人名也是一个字串，一篇文档，如果其以待消歧人名字串为子串，就称该文档中出现了待消歧人名字串。如果一篇文档，其中不仅出现了待消歧人名字串，而且该字串在该文档的上下文环境下表示一个完整的人名，就称该文档包含待消歧人名。基于上述的定义，人名消歧问题可以被这样描述，对于待消歧人名，给定包含该待消歧人名字串的若干篇文档，从这些文档中找出包含该待消歧人名的文档，并依据文档中待消歧人名指向的实体是否相同对找出的这些文档进行聚类，使得聚到同一个簇中的文档中的待消歧人名指向相同的实体。因此，本质上人名消歧是一个聚类的问题。目前，针对跨文档的人名消歧研究比较多，可分为基于人名上下文词语特征的方法和利用社会网络等外部知识的方法。

基于人名上下文词语特征的方法，从人名的上下文中提取特征，将文档表示成向量，定义向量上的相似度，使用聚类算法进行聚类。此类方法的研究主要集中在特征的选择和聚类算法的改进两方面。

在特征的选择方面，很多研究者提取待消歧人名所在文档中的显式信息。Wang 等人选择某些信息量大的词性的词和人名前后的词作为特征[15]。而 Xu 等人则选择关键短语作为特征[16]，Yoshida 使用命名实体、复合关键词（compound key word）和 URLs 为特征[17]。Xu 等对其之前提出的使用关键短语做特征的方法进行改进，将人名的上下文表示成关键短语的序列[18]。Chen 等人对特征加权方法进行了改进[15]。Bekkerman 等人在人名消歧中使用了网页之间的链接信息[19]。Han 等利用 Wikipedia 构建了一个大规模的语义网络，将共现的人名表示成向量，使用向量相似度计算方法计算其相似度，使用凝聚层次聚类算法聚类[20]。当文档的主题与其中待消歧实体的主题一致时，文档中的显式特征能够很好地反映待消歧实体的特点。但是，并不是所有的文档都满足上述的假设。例如：文档中出现并列介绍多个内容时，文档的主题是所有内容的综合，此时文档的主题与人名的主题是包含关系，因此文档的主题很可能与人名的主题并不相同。在人名消歧时，待消歧人名的主题容易被其文档主题误导。针对这个问题，如何提取能够更好地体现待消歧实体特点的特征引起了关注。

有的研究者从如何获得文档中待消歧实体的准确上下文入手，希望通过对上下文的精准定位来提高特征的质量。Bagga 等利用文档内共指消解，获得实体共指链，提取包含与该人名共指的实体所在的句子，使用向量空间模型，设定相似阈值，实现对人名的跨文档共指消解[21]。显然，该方法将查询人名的有效上下文限制在包含该人名相关共指链上实体的句子范围内。还有的研究者提出了提取文档中的隐式的特征来解决这个问题。Song 首先针对人名和词学习一个话题分布，将该分布作为特征使用凝聚层次聚类实现人名消歧[22]。Han 提出了一种使用异源知识构建待消歧人名字串的概率生成模型的方法，在该模型中，待消歧人名字串被看成是由一个三步采样生成的[23]。Han 等提出了一种同时考虑人名上下文和文档主题的人名消歧模型[24]。

很多研究是针对聚类算法展开的。Gong 针对使用凝聚层次聚类时无法确定聚类结束条件（即相似度阈值或簇个数）的问题，提出了一种从完全聚类树上确定最佳分裂点的方法，将从完全聚类树上确定最佳分裂点的问题看成一个二分类问题，利用分裂点的信息提取了特征，使用支持向量机的方法在训练集上训练，对新的聚类树上的分裂点做出判断，来提高聚类性能[25]。

Xu 先使用凝集层次聚类，然后使用基于中心的方法找出聚类结果簇中的离群点，使用基于短语的字符串核的支持向量机进行重新分类到更相似的簇中[16]。Wang 等针对常规的基于聚类的人名消歧方法，提出了基于两阶段策略的自适应共振理论来解决人名消歧的问题，该方法在第一阶段进行对待消歧人字串进行聚类，第二阶段合并相似的簇[26]。Tang 等针对人名消歧过程中很难捕获人名消歧过程中需要的所有信息以及聚类簇个数的确定问题，提出了一种对人名消歧的目标函数进行的两阶段的参数估计方法[27]。Yoshida 先使用强特征得到高准确低召回的聚类结果，利用聚类结果使用 Bootstrapping 的方法改善召回的情况，实验结果表明，性能好于 WePS 的最好结果 [17]。Chen 等人以人名消歧精度为指标来选择更合适的聚类停止点[15]。Tang 等人提出了一种能够动态评估聚类簇个数的方法[28]。Christof 等人使用标准的信息检索模型来对待消歧人名字串进行聚类[29]。Anderson 等人提出了一种自训练的方法进行人名消歧，在训练数据量比较小时仍然能够获得不错的性能[30]。

基于人名上下文词语特征的方法模型比较简单，容易达到一个不错的人名消歧性能，但文档中待消歧实体必须被表示成向量的形式，受到这个束缚，人名消歧的性能很难被进一步提高。因此，有的研究者提出了不将待消歧实体向量化的方法。这类的方法主要是根据采用的信息和构造的模型，使用相应的方式来完成待消歧实体的聚类，从而完成人名消歧。

有的研究者对待消歧实体的社会网络进行建模，在该模型上进行人名消歧。Tang 等人使用文档中待消歧人名共现的命名实体，构建该人名的社会网络，将该社会网络表示成二部图，还提出了一种基于二部图的社会网络相似度度量方法，基于此度量方法进行自底向上聚类，该方法在 WePS-2 人名消歧任务的测试集上取得了高于最好评测结果的好成绩[31]。郎君等人为了得到文档中待消歧人名的更加完整的社会网络，通过构造查询关键字，利用搜索引擎获得扩展待消歧人名社会网络的信息来完善其社会网络，最后利用所得的待消歧人名的社会网络进行人名消歧[32]。有的研究者利用外部的知识库，首先分析文档中待消歧实体在特定知识库体系下的情况，然后对处于相同情况的待消歧实体聚为一个簇。这种方法将本是聚类任务的人名消歧问题通过分类的方法进行解决。

对于分析和确定文档中待消歧实体在特定知识库体系下的情况的方法，包括学习排序和信息检索两大类。Liu[33]和 Li[34]都对学习排序做了很好研究。Geng 等人提出了一种对于给定查询人名，使用若干个最接近的查询人名以

及各自的查询文档的特征向量动态构造排序模型的方法[35]。Zhu 等人则对查询人名做了聚类，对不同簇中的查询人名使用特殊的排序模型[36]。

很多研究使用了不同的知识库。Zhu 等人利用大量的人物实体信息中，构建一棵人物本体树，然后通过提取有用的信息和改进的本体树匹配算法来有效地衡量两个人名字串之间的相似度[37]。Han 等人利用 Freebase，抽取到基于分类法的人名，通过网上检索获得人名属于某个分类或职业的证据，对于具有属于多分类或职业的人名利用网上检索的网页证据进行合并。然后利用这些证据，以证据网页的 Token、命名实体和 URL Token 为特征，使用 TF-IDF 加权，用 KNN 对另外网页中人名进行分类[38]。Chen 等人以社交网络为知识库，利用社交网络进行人名消歧[39]。Han 等人为了达到更好的人名消歧性能，综合了多个知识源来做人名消歧[40]。

1.3 本文的主要研究内容和组织结构

本课题的研究内容主要是围绕人名消歧问题，从几个方面展开：

对目前常用的中文人名识别方法在非规范句子中识别能力弱和对易与普通词混淆的人名的区分性能不好的问题，本文提出了在常用中文人名识别结果的基础上，利用规则进行校正的人名识别方法。该方法利用的规则主要包括人名构成规则和人名环境信息。实验表明，该方法为人名消歧提供了有力的保障。

针对中文文档内共指消解问题，本文提出了一种结合统计和规则的方法。首先利用文档的句法信息，提取候选的名词短语；然后将两个候选的名词短语结合为名词短语对，作为共指关系判断的对象；最后识别出的共指名词短语对，形成共指链，将整个文档中的共指关系得到消解。在 CoNLL2012 共指消解中文数据集上的实验表明，该方法能够很好地消解中文文档内的共指。

针对跨文档的中文人名消歧问题，本文提出了一种结合互联网百科知识和互联网验证的中文人名消歧方法。针对待消解人名对应的实体信息短缺问题，通过使用互联网百科知识丰富人名个体的信息以提高系统精确率，然后使用互联网验证对消歧结果进行验证，来提高系统召回率，从而使人名消歧的性能得以提升。在 CIPS-SIGHAN2012 人名消歧数据集上的实验表明，该方法能够明显地提高人名消歧的性能。

本论文分四个章节，各章节组织如下：

第一章绪论介绍了人名消歧的背景以及研究的目的和意义，并介绍了人名识别、文档内共指消解和文档间的人名消歧的国内外研究现状和分析，最

后对本课题的主要研究内容做了概述。

第二章介绍了本文采用的利用规则对人名识别结果进行校正的中文人名识别方法。

第三章介绍了本文提出的一种结合统计和规则的中文文档内共指消解方法，并对实验结果进行了分析。

第四章介绍了了本文提出的跨文档中文人名消歧方法，并对该方法在CIPS-SIGHAN2012 数据集上进行了评估和分析。

最后对全文工作进行了总结。

第 2 章 中文人名识别的改进方法

2.1 引言

人名作为一种命名实体，携带着非常丰富的信息，在信息查找和信息处理过程中起着极为重要的作用。人名识别是人名消歧的基础，对人名消歧的性能有着决定性的影响。对于一般场景下人名的识别，现成的方法比较成熟，且能够取得了相当不错的性能。但是，对于特殊场景下的人名，常见方法的性能往往不能满足需求。这些特殊场景有很多，本章主要关注两种情况：人名处于非完整句子的情况和待识别的人名易与普通词混淆的情况。常见的人名识别方法都是使用规则的完整的句子作为训练集，得到的人名识别模型通常利用了句子的信息，即候选人名字串的上下文信息。而在非完整句子的情况下，使用在完整句子上训练得到的人名识别模型很难做出正确的识别。此外，对于易与普通词混淆的人名，由于普通词出现次数远高于人名出现次数，基于统计的方法更倾向于将其识别为普通词，而不是人名。为了改善上述这两种情况下人名识别性能不好的问题，本章采用了一种对人名识别结果进行校正的混合人名识别方法，即先对文档中的人名进行识别，基于该识别结果，利用规则对相应的情况进行检查，对于与规则相冲突的识别结果进行校正，以此来提高特殊情况下人名识别的性能。

2.2 基于后处理规则的中文人名识别

作为命名实体的一种，人名如其他命名实体相同，具有很强的规则性，因此，可以通过使用这些人名的构成规则进行人名识别。中文人名有着很强的名族特色，很多少数名族有着独特的人名构成规则，由于其特殊性，往往能够被常规人名识别方法很好地解决。本课题的人名识别主要针对中文人名中占绝大数的简单的人名进行识别。

根据人名构成的特点，构造一个基于人名构成规则的人名判别器。完整的人名由姓和名两部分组成，其中姓的用字比较集中，是一个有限集合，而名的用字相对来说范围比较大，但大多数人名的名用字是比较集中的。中文人名中姓的种类分为：单姓、双姓和复姓三种，其详细介绍及举例如表 2-1 所示。名的种类分为单名和双名两种，其详细介绍及举例如表 2-2 所示。姓与名的组合得到中文姓名，其种类及举例如表 2-3 所示。

表 2-1 中文人名中姓的种类及举例

类别	说明	举例
单姓	姓由单个字构成	赵、李
双姓	姓由两个单姓组合而成，一般为：父亲姓+母亲姓	张包、李
复姓	姓由固定的两个字构成	诸葛、司

表 2-2 中文人名中名的种类及举例

类别	说明	举例
单名	名由单个字构成	亮、伟
双名	名由两个字构成	建国、晓明

表 2-3 中文人名种类及举例

类别	举例
单姓单名	李白、张伟
单姓双名	邓建国、李连杰
双姓单名	李张弘
双姓双名	张包子俊
复姓单名	诸葛亮、司马迁
复姓双名	欧阳鲲鹏、淳于珊珊

文档是一个字串，人名识别就是要找出其中所有是人名的子字串。在规范的语料中，常用的中文人名识别方法能够很好地对其中的人名进行识别，如对句子例 2-1 采用斯坦福分词系统 Stanford 和中科院分词系统 ICTCLAS 的人名识别，观察结果，其中“奥巴马”和“习近平”能够都被准确的识别到。斯坦福分词系统 Stanford 命名识别结果中标记的含义如表 2-4 所示。中科院分词系统 ICTCLAS 的命名实体识别结果中标记的含义如表 2-5 所示。

例 2-1 *据美国媒体报道，白宫在 4 日公布了美国总统奥巴马与中国国家主席习近平即将会面的细节*

Stanford 人名识别结果：*据/O 美国/GPE 媒体/O 报道/O ，/O 白宫/ORG 在/O 4/O 日/O 公布/O 了/O 美国/GPE 总统/O 奥巴马/PERSON 与/O 中国/GPE 国家/O 主席/O 习近平/PERSON 即将/O 会面/O 的/O 细节/O*

ICTCLAS 人名识别结果: *据/p 美国/ns 媒体/n 报道/v ，/w 白宫/n 在/p 4 日公/t 布/n 了/u 美国/ns 总统/n 奥/b 巴马/ns 与/p 中国/ns 国家/n 主席/n 习近平/nr 即将/d 会面/v 的/u 细节/n*

但是，在非规则语料，常见的人名识别方法对人名往往不能很好的识别

出。如对句子例 2-2 分别采用 Stanford 和 ICTCLAS 的人名识别方法的结果如下。Stanford 方法未能将人名“翁德云”识别为人名，而是错误的标记为“MISC”，ICTCLAS 则不能很好的判断人名“孙高明”的边界。

表 2-4 部分斯坦福系统命名实体识别标记含义说明

编号	标记	对应的英文名称	含义
1	GPE	Geographical Political Enti-ties	地理行政类型实体
2	ORG	Organization	组织、机构、团体
3	PERSON	Person	人名
4	-LRB-	Left Round Bracket	替换文中左圆括号
5	-RRB-	Right Round Bracket	替换文中右圆括号
6	MISC	Miscellaneous	其他杂项
7	O	Other	其他

表 2-5 部分中科院命名实体识别标记含义说明

编号	标记	含义
1	p	介词
2	ns	地名
3	n	普通名词
4	v	普通动词
5	w	标点符号
6	t	时间词
7	u	助词
8	b	区别词
9	nr	人名
10	d	副词
11	y	语气词
12	r	代词
13	x	字符串
14	g	语素字
15	vn	名动词
16	m	数词

例 2-2 *他还说:“三都澳一开放,那搬家就忙不过来了。”(《闽东报》翁德云、洪茂绪、孙高明)(“八里罕杯”短消息征文)*

Stanford 人名识别结果：*他/O 还/O 说/O :/O ``/O 三/MISC 都/O 澳/GPE 一/O 开放/O ,/O 那/O 搬家/O 就/O 忙/O 不/O 过来/O 了/O 。/O ''/O -LRB-/MISC 《 闽东报/MISC 》 翁德云/MISC 、/O 洪茂绪/PERSON 、/O 孙*

高明/PERSON -RRB-/MISC -LRB-/MISC ``/MISC 八里罕杯/MISC ''/MISC 短/MISC 消息/MISC 征文/MISC -RRB-/MISC

ICTCLAS 人名识别结果：*他/r 还/d 说/v :/x “/w 三/m 都/d 澳/b 一/m 开放/v ,/w 那/r 搬家/v 就/d 忙/v 不/d 过来/v 了/y 。/w ”/w (/w 《/w 闽东/ns 报/n 》/w 翁德云/nr 、/w 洪茂绪/nr 、/w 孙/nr 高明/a)/w (/w “/w 八/m 里/q 罕/g 杯/g ”/w 短消息/n 征文/n)/w*

另外，常见的人名识别方法对大部分的人名识别性能比较好，但对于特殊的人名，如容易与普通词混淆的人名，其识别性能就比较差了。如对句子例 2-3 分别采用 Stanford 和 ICTCLAS 人名识别方法，其结果如下。可以看出，由于句子中要识别的人名为“白云”，而白云既可以作为人名使用，也可以作为普通词使用，使得人名识别方法不能正确地将句子中的“白云”识别为人名。

例 2-3 *阳泉市长白云：发展非煤产业 推进城市转型.*

Stanford 的人名识别结果：*阳泉市/GPE 长/O 白云/O ：/O 发展/O 非/O 煤/O 产业/O 推进/O 城市/O 转型/O ./O*

ICTCLAS 的人名识别结果：*阳泉市/ns 长白/ns 云/n ：/w 发展/v 非/b 煤/n 产业/n 推进/v 城市/n 转型/vn ./w*

由上述的例子可以看出，现有的方法在规范语料中的人名识别性能比较好，而在非规范的语料中或是针对容易与普通词混淆的人名，人名识别性能很差。

针对这个问题，本文采用了一种对人名识别结果进行规则化校正的人名识别方法。校正之前的人名识别结果依赖现成的人名识别方法给出，如 Stanford 的人名识别方法。校正时，利用中文人名的构成规则，结合人名识别结果调整候选人名边界，然后根据候选人名的上下文，判断该候选人名是人名，还是普通词。

人名边界的调整主要针对两个情况进行矫正：情况一，利用文档内的特殊格式，确定人名的边界，如果符合人名的构成特点；情况二，对于使用常规方法进行人名识别后的结果，查看是否有人名的姓后面跟着单字词，而这些单字词与前面姓结合符合人名的构成特点。

候选人名的判断使用其上下文，主要是考察该候选人名左右是否有与人有关的特征词语，如果其左右出现了与人有关的词语，就将该候选人名识别为人名，更改分词结果和人名识别结果；否则，不做处理。其中与人有关的

特征词包括：人名、职业。

句子例 2-4 不是一个规范的句子，没有完整的句子结构，是介绍电视剧演员列表的一行，正如前面分析的那样，人名“白雪”没有被完全识别正确。由于有“----”和换行符将人名分开了，可以使用这些格式确定人名候选词的边界，即得到人名候选词“白雪”。利用前面的人名构成规则，查看这个人名候选词是否满足该规则，显然，两个人名候选词都满足规则，“白雪”属于单姓单名。最后通过检查候选人名左右是否有人名特征词，来决定是否将其识别为人名。在例 2-4 中，人名候选词周围有人名“王珠儿”（在 stanford 人名识别结果中，在 ICTCLAS 人名识别结果中为“王珠”），这样，人名候选词“白雪”被识别为人名。另外，对于 ICTCLAS 的人名识别结果中，人名“王珠儿”也可以通过确定正确的人名边界被修正而识别出来。

例 2-4 *王珠儿----白雪*

Stanford 的人名识别结果：*王珠儿/PERSON --/O--/O 白雪/O*

ICTCLAS 的人名识别结果：*王珠/nr 儿/g ----/m 白雪/n*

2.3 文档内识别结果一致化方法

人名的歧义存在于包含待消歧人名字串的上下文中，要消除待消歧人名的歧义，就必须分析所有包含待消歧人名字串的文档。但是，只有包含待消歧人名的上下文才需要进一步分析该人名指向的实体情况，对于那些出现待消歧人名字串但不包含待消歧人名的上下文，无需作其他处理，只需将其与包含待消歧人名的上下文区分开就消除了其中待消歧人名字串的歧义。因此，在人名消歧之前，需要判断文档内的人名字串是否是人名。

一般而言，在相同的上下文或文档范围内，一个字串或词语的角色是相同的。基于这样的观察，在通过规则对人名识别结果进行更正之后，还可以利用文档内多个相同字串的识别结果进行一致化，来提高对该字串的人名识别性能。

文档中，同一候选人名字串会出现多次，如果被识别为人名的次数超过该候选人名字串的 30%，就将该文档内的该候选人名字串的人名识别结果进行一致化，将其全部识别为人名。否则，将其一致化为普通词。

通过文档范围内相同人名候选词识别结果一致化，能够进一步更正未被识别为人名的候选人名字串。

2.4 实验结果和分析

本章的人名识别方法主要用来确定某一文档中某个字符串是否表示一个实体的人名，因此，设计了一个特殊的实验来评估本章采用的人名识别改进方法。实验采用了 SIGHAN-CIPS2012 的人名消歧数据集，该数据集包含 16 个人名，每个人名对应有若干篇文档，每篇文档中人名字串是否表示一个实体是确定的。该数据集的详细信息如表 2-6 所示。

表 2-6 SIGHAN-CIPS2012 数据集信息

人名	总文档数	字串是人名的文档数	字串是人名的文档数占比
丛林	84	68	0.810
华山	91	90	0.989
华明	47	46	0.979
方正	94	92	0.979
杜鹃	126	119	0.944
白云	195	181	0.928
白雪	94	85	0.904
胡琴	53	37	0.698
雷雨	61	47	0.771
高山	105	91	0.867
高峰	161	145	0.901
高明	165	147	0.891
高超	82	74	0.902
高雄	59	58	0.983
黄河	127	126	0.992
黄海	90	89	0.989
合计	1,634	1,495	0.915

使用本文的方法对人名对应的文档中的该人名的字串进行人名识别，如果文档中该人名字串是一个词，且表示一个实体人物，则将该文档分为“表示人名的文档”类；如果该人名字串不成词，或则虽然成词却是非人名的普通词，则将该文档分为“不表示人名的文档”类。因此，本实验是以文档为单位的。实验的性能采用“表示人名的文档”类的准确率、召回率以及 F 值来衡量。

本实验使用了 3 个其他方法与本文采用的方法进行比较，实验结果如表 2-7 所示。All_for_name 方法简单地将所有文档分为“表示人名的文档”类。

Stanford 的方法为使用斯坦福命名实体识别工具对文档进行人名识别，只要文档中出现对应的人名字串被识别为人名的情况，就将该文档分为“表示人名的文档”类。HLT_Rules 方法为不使用现有工具，直接使用前面介绍的方法。而 Stanford+HTL_Rules 方法为在 Stanford 人名识别结果的基础上再使用本章介绍的方法进行处理的方法。

表 2-7 人名识别结果

方法	P	R	F
All_for_name	0.915	1.000	0.956
Stanford	0.951	0.452	0.612
HLT_Rules	0.943	0.924	0.934
Stanford+HLT_Rules	0.970	0.980	0.975

实验结果显示，使用简单的 All_for_name 的方法就可以达到 95.6%的 F 值，而 Stanford 的方法和 HLT_Rules 的方法分别达到了 61.2%和 93.4%的 F 值，均低于 All_for_name 的结果，而 Stanford+HLT_Rules 方法获得了 97.5%的 F 值，超过了 All_for_name 方法 1.9%。

进一步分析可知，本数据集中的人名往往易与普通词混淆，存在歧义，如人名“胡琴”，除去可以是人名之外，还可以是一种乐器名称，是一个普通名词。基于规则语料训练而成的 Stanford 方法是对规则语料中的命名实体识别问题进行建模的，对于该数据集中的这些人名，在大范围的环境中，用作普通词的可能性比较大，因此，倾向于将其识别为普通词，从而导致识别中召回率非常低。针对这些特殊人名，本文采用的方法从人名的构成规则和人名的出现规律方面出发，使用比较宽松的限制条件达到了比较高的人名识别性能，为之后的人名消歧做好了准备。

2.5 本章小结

本章针对人名消歧问题中必须进行的人名识别模块，提出了一种结合人名构成和人名周围信息的方法对常用人名识别结果进行校正的方法，提高人名识别的性能。该方法主要是为了提高非规范的句子中人名识别和易与普通词混淆人名的识别的性能。此外，还讨论了通过对同一文档范围内的相同候选人名字串进行识别结果一致化来进一步改善人名识别性能的方法。

第 3 章 文档内中文共指消解

3.1 引言

文档内共指消解就是要找出文档内指向现实中相同个体的名词性短语。为了表达的需要，在文档内，同一个实体常常被用不同的名词性短语来表达，一方面是为了避免重复，使读者免于生厌，另一方面是为了简化表达。在句子例 3-1 中，“埃斯特拉达”与后面的两个“他”指向相同的实体，后面两个“他”的使用使得句子简短，利用读者或听者能够尽快的捕捉到整句话的整体含义。文档内共指消解就是要找出共指的名词短语，在例句中就是“埃斯特拉达”和后面的两个“他”，它们的共指链表示如图 3-1 所示。

例 3-1 *埃斯特拉达 表示，他 希望 上帝 能够 赐给 他 智慧*

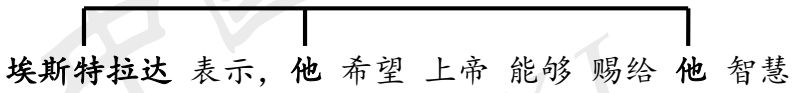

图 3-1 共指关系示例

本章将基于统计的方法和基于规则的方法相结合，首先使用统计的方法对共指对候选是否构成共指进行判定，再使用一些准确率非常高的规则对判定结果进行校正，在人名指代消解上得到了更高的性能。

文档内共指消解是对文档内的名词性短语之间的共指关系进行确定。共指的两个名词性短语，在现实中指向相同的实体。在例句“他是我的导师”中，“他”和“我的导师”作为两个名词性短语，虽然不能字串匹配，但却指向相同的实体。本章就基于统计和基于规则相结合的中文共指消解方法进行介绍，主要包括候选名词短语提取、共指关系确定和共指链生成三个步骤。

3.2 候选名词短语提取方法

共指消解需要考察的都是名词短语，为了消解文档内共指，在衡量它们之间的共指关系以及生成共指链之前，必须先提取需要考察的名词短语，作为候选名词短语。

由于共指消解需要考察的都是名词短语，因此最直接的方法就是将所有的名词短语提取出来，将它们都作为共指消解的考察对象，然而，这个是没

有必要的。观察发现，中文在表达一个名词短语时，中心词往往放在该名词短语的末尾，即中心词后置现象，如句子例 3-2 中，“加拿大历史上任职时间最长的总理”、“总理”都是名词短语，由于他们后缀匹配，中心词是相同的，因此它们一定指向相同的实体，其词性信息和句法分析结构如表 3-1 所示，词性标记含义见附录一。

例 3-2 *特鲁多 是 加拿大 历史 上 任职 时间 最 长 的 总理*

表 3-1 例 3-2 的词法信息和句法信息

编号	词	词性	句法结构
0	特鲁多	NR	(TOP(IP(NP*)
1	是	VC	(VP*
2	加拿大	NR	(NP(CP(CP(IP(LCP(NP(NP*)
3	历史	NN	(NP*))
4	上	LC	*)
5	任职	NN	(NP*
6	时间	NN	*)
7	最	AD	(VP(ADVP*)
8	长	VA	(VP*)))
9	的	DEC	*))
10	总理	NN	(NP*)))
11	。	PU	*))

为了减小问题的规模，本文在候选名词短语提取的过程做特殊的处理，只提取最长的名词短语。这样，在上述的例句中，只有“加拿大历史上任职 时间最长的总理”被提取为候选名词短语。

3.3 共指关系确定方法

在提取候选名词短语之后，要进行共指关系确定，就是要判定候选名词短语之间是否存在共指关系，这样的共指关系将用来在下一步生成共指链。本文以两个候选名词短语形成的候选名词短语对作为共指关系衡量的对象，通过确定形成的候选名词短语对是否共指，来确定整个文档内名词短语的共指情况。这里的候选名词短语对的共指关系确定采用了基于统计和基于规则相结合的方法。首先使用基于统计的方法，以候选名词短语为对象，使用机器学习方法训练得到的分类器对其进行分类，确定它们之间是否存在共指关系。然后利用规则的方法对分类器的分类结果进行校正。

3.3.1 基于统计的共指关系确定方法

为了确定候选名词短语对之间是否存在共指关系，本文采用机器学习的方法，训练一棵决策树分类器，对候选名词短语对是否具有共指关系进行判别。

对于候选名词短语对，通过提取特征，将其表示成向量，这样，每一个候选名词短语对向量作为一个实例。两个共指的候选名词短语对应的实例作为正例，两个不共指的候选名词短语对应的实例作为反例。

在一篇文档 D 中，假设提取出的候选名词短语集合为

$$SNP(D) = \{ np_1, \quad np_2 \ldots np_N \},$$

它们的共指关系为

$$R\text{-}Coreference(D) = \{ chain_1, chain_2 \ldots chain_L \},$$

对于 $chain_i$ 有

$$chain_i = \{ np_{i,1}, np_{i,2} \ldots np_{i,m(i)} \},$$

其中，$1 \le i \le L$, m(i)为共指链 $chain_i$ 包含的名词短语个数，L 为共指链数目。

对于形成实例的方法，常规的方法是不对候选名词短语对进行过滤，即将文档内任意两个名词短语组合作为一个实例。这种方法形成的实例数目庞大，而且正例和反例的比例相差比较大，正例远少于反例。为了得到更好的共指消解性能，本文对上述的实例形成方法做了改进，缩小正例和反例的数目差距。

名词短语间具有共指关系表示名词短语指向相同的实体，这种关系是可以传递的。考虑名词短语 np_1、np_2 和 np_3，如果 np_1 与 np_2 存在共指关系且 np_2 与 np_3 存在共指关系，那么，np_1 与 np_2 也必然存在共指关系。此外，在一篇文档中，被提到的某实体出现的位置总是比较集中，很少出现某一实体出现的位置跨越很大的距离。基于上述两个观察，本文对实例的形成方法做了调整，不是考虑文档内所有名词短语的组合，而是只用距离为 10 个句子距离的范围内的名词短语形成实例。一方面，这种形成实例的方法在大大地减少了形成实例的个数的同时，在一定的程度上缩小了正例和反例之间数目的差距；另一方面，共指的名词短语对之间的信息并没有被减少，可以利用这些信息还原整个共指链。

此外，本文对于实例的形成还做了一步改进，进一步缩小正例和反例之间的数量差别。对于正例的形成没有改变，仍然使用 10 句范围内的所有具有共指关系的名词短语对形成正例。对于反例的形成做了些调整，对 10 句范围内的所有名词短语，并不是所有的不存在共指关系名词短语对都用来形成反

例，而是要求形成反例的两个名词短语之间不得存在共指对。考虑 10 句范围内的名词短语集合为{np_1, np_2, np_3, np_4, np_5}，其中，np_2 与 np_4 之间存在共指关系，如图 3-2 所示。那么生成的实例包括{<np_1, np_2, - >, <np_1, np_3, - >, <np_2, np_3, - >, <np_2, np_4, + >, <np_3, np_4, - >, <np_3, np_5, - >, <np_4, np_5, - >}，而实例{<np_1, np_4, - >, <np_1, np_5, - >, <np_2, np_5, - >}采用这样的方法，将被过滤掉，不生成相应的实例。这是因为这些名词短语对之间存在共指的名词短语对，例如：<np_1, np_4>之间存在共指的名词短语对<np_2, np_4>。

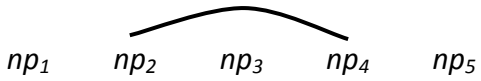

图 3-2 名词短语共指关系图

经过上述的方法，保留的名词短语对将用来形成实例，用于使用机器学习来训练分类器。本文针对共指消解的问题，设计了 61 个特征，包括 5 大类：基本信息、词性信息、句法信息、语义信息和文档环境信息，每个特征大类由包含若干子类，如表 3-2 所示。

表 3-2 特征类别信息

特征类别	特征子类
基本信息	名词短语匹配程度
	名词短语词距离
	名词短语句子距离
词性信息	名词短语是否是代词以及代词类别
	名词短语是否是专有名词以及专有名词的类别
句法信息	句法层级信息
	名词短语与所属子句的谓语的关系
	名词短语与连词的位置关系
语义信息	名词短语的单复数信息
	名词短语是否是人物词
	说话人信息
文档环境信息	名词短语所属句子的类别信息
	名词短语的位置信息

由每个名词短语形成的实例包括每个特征的特征值以及类别标签，是一个 62 维的向量

$$[\, t_1, t_2, \ldots t_i \ldots t_{61}, C], \quad 1 \le i \le 61$$

其中，t_i 表示实例对于第 i 个特征的特征值，C 为类标，$C \in \{+, -\}$，+

表示形成该实例的两个名词短语之间存在共指关系，即该实例为正例，- 表示形成该实例的两个名词短语之间不存在共指关系，即该实例为反例。本文采用决策树的机器学习方法，训练产生一棵决策树，用于对名词短语对形成的实例进行分类，从而对名词短语对之间的共指关系进行了判断。

3.3.2 基于规则的共指关系确定方法

上述基于统计的方法将名词短语对的共指判断问题映射到所选择的特征空间中，该特征空间和机器学习的方法共同决定了这种模型的假设空间。在理想的情况下，经过机器学习，可以得到该假设空间中的一个不错的假设，作为判定共指的分类器。但在训练集中，由于不能保证每一种类别的实例相对于训练来说足够多，再加上噪音数据的影响，使得学习到的模型可能对某些情况不能做出正确的判断。

为了弥补这方面的不足，提高共指消解方法的性能，本文提出了采用规则的方法，对符合规则的经过基于统计方法判定过的名词短语对进行过滤，校正那些分类错误的名词短语对。

规则的方法就是分析语料，结合自然语言的使用规律，总结出一些规则，使用这些规则对名词短语之间的共指关系进行判定。规则 R 是一个二元组 R=<Conditions, Operation>，其中 Conditions 是规则对应的条件集合，Operation 表示满足该规则的所有条件后要执行的操作。条件集 Conditions 由一个或一个以上的条件组成，用于定义规则使用的条件，Conditions={c_1,c_2 ... c_n}，其中 c_i 表示条件，1<= i <= n, n 为该规则中条件的个数。在基于名词短语对共指关系判断问题中，操作 Operation 只有两个取值：共指与不共指，即 Operation ∈{共指，不共指}。

在例 3-3 中，根据前面的候选名词短语提取方法，提取出的名词短语集合为{“高行健”， “记者”， “他”， “自己”，“这项桂冠殊荣”}。为了分析该句中名词短语对之间的关系，需要使用其词法分析和句法分析等信息，如表 3-3 所示。

例 3-3 *高行健 在 接受 记者 访问 时 表示，他 很 诧异 自己 能够 获得 这项 桂冠 殊荣*

在这例句中，“高行健”、“他”和“自己”都指向实体“高行健”，形成了一条共指链，如图 3-3 所示。

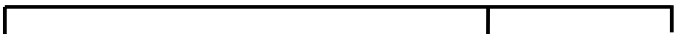

图 3-3 例句 3-3 中共指链

其中“高行健”和“他”共指，“他”和“自己”共指。这两对名词短语的共指判定都使用规则的方法。

表 3-3 例句 3-3 的词法分析和句法分析结果

编号	词	词性	句法信息	Tree-bank				
0	高行健	NR	(TOP(IP(NP*)	*	*	(ARG0*)	*	*
1	在	P	(VP(PP*	*	*	(ARGM-TMP*	*	*
2	接受	VV	(LCP(IP(VP*	(V*)	*	*	*	*
3	记者	NN	(IP(NP*)	(ARG1*	(ARG0*)	*	*	*
4	访问	VV	(VP*))))	*)	(V*)	*	*	*
5	时	LC	*))	*	*	*)	*	*
6	表示	VV	(VP*	*	*	(V*)	*	*
7	，	PU	*	*	*	*	*	*
8	他	PN	(IP(NP*)	*	*	(ARG1*	(ARG0*)	*
9	很	AD	(VP(ADVP*)	*	*	*	(ARGM-ADV*)	*
10	讶异	VV	(VP*	*	*	*	(V*)	*
11	自己	PN	(IP(NP*)	*	*	*	(ARG1*(ARG0*)	
12	能够	VV	(VP*	*	*	*	*	*
13	获得	VV	(VP*	*	*	*	*	(V*)
14	这	DT	(NP(DP*	*	*	*	*(ARG1*	
15	项	M	(CLP*))	*	*	*	*	*
16	桂冠	NN	(NP*	*	*	*	*	*
17	殊荣	NN	*))))))))))	*	*	*)	*)	*)
18	。	PU	*))	*	*	*	*	*

“高行健”和“他”共指判定使用的规则可以简单的描述为第三人称主句主语与重句主语共指，条件集合中条件的详细描述如表 3-4 所示，按照此规则，“高行健”和“他”组成的名词短语对正好满足该规则的要求，直接判定“高行健”和“他”共指。

该规则所描述的情况并不是极少数的情况。考虑句子例 3-4，其结构信息如表 3-5 所示，句中的“阿布拉莫夫”和“他”也完全满足上面规则提到的全部条件，可以直接判定名词短语“阿布拉莫夫”和“他”共指关系成立。

例 3-4 *目前 呢 ， 阿布拉莫夫 正 在 医院 接受 救治 ， 但是 他 已*

经 脱离 了生命 危险 。

表 3-4 第三人称主句主语与重句主语共指规则的条件描述

编号	条件描述
1	第一个名词短语为主句主语
2	第一个名词短语为第三人称（如：名称、第三人称代词）
3	第二个名词短语为主句的直接从句主语
4	第二个名词短语为第三人称代词
5	两个名词短语单复数相同

表 3-5 例句 3-4 词法分析和句法分析等信息

编号	词	词性	句法信息	命名实体	Tree-bank	
0	目前	NT	(TOP(IP(IP(NP*)	*	(ARGM-TMP*)	*
1	呢	SP	(FLR*)	*	*	*
2	，	PU	*	*	*	*
3	阿布拉莫	NR	(NP*)	(PERSON)	(ARG0*)	*
4	正	AD	(VP(ADVP*)	*	(ARGM-ADV*)	*
5	在	P	(PP*	*	(ARGM-LOC*	*
6	医院	NN	(NP*))	*	*)	*
7	接受	VV	(VP*	*	(v*)	*
8	救治	NN	(NP*))))	*	(ARG1*)	*
9	，	PU	*	*	*	*
10	但是	AD	(IP(ADVP*)	*	*	(ARGM-DIS*)
11	他	PN	(NP*)	*	*	(ARG0*)
12	已经	AD	(VP(ADVP*)	*	*	(ARGM-ADV*)
13	脱离	VV	(VP*	*	*	(V*)
14	了	AS	*	*	*	*
15	生命	NN	(NP*	*	*	(ARG1*
16	危险	NN	*))))	*	*	*)
17	。	PU	*))	*	*	*

句子例 3-1 中，“他”和“自己”共指关系也可以使用使用基于规则的方法进行判定。该规则的条件描述如表 3-6 所示。使用该规则，可以判定例句 3-1 中的“他”和“自己”共指。

基于规则的方法强制系统对某些情况做出相应的判断，相当于在系统训练时，添加了能够指导系统学习到此规则的大量的实例。一方面，规则是通过人工总结的，具有高度概括性，和很高的准确性；另一方面，规则的引入在一定程度上避免了系统由于某类型训练数据不充分而学习不到正确的分类

模型。基于上述两方面，采用基于规则的方法对基于统计的共指关系判定方法能够有很好的补充，提升系统整体性能。

表 3-6 “自己”规则的条件描述

编号	条件描述
1	第一个名词短语为主句主语
2	第二个名词短语为主句的直接从句主语
3	第二个名词短语为代词“自己”
4	两个名词短语单复数相同

3.4 共指链生成方法

经过共指关系判定，10 句范围内任意两个名词短语之间的共指关系就被确定了，为了消除文档内的共指，还需要将共指的名词短语对链接起来形成共指链。为了确定共指链，文档内所有共指的项必须被串起来，即同一条共指链上的任意两个名词短语之间都存在共指关系。

前面获得的 10 句范围内名词短语对的共指关系是一个无向图，其中名词短语是无向图中的点，共指关系是无向图中的边，两个名词短语之间存在共指关系表示无向图中两个点之间存在一条无向边。这样，由前面获得的 10 句范围内名词短语对的共指关系，生成共指链的过程，可以看作是对一个无向图求它的传递闭包，采用 Floyd 算法求无向图的传递闭包，如算法 3-1 所示。

算法 3-1 Floyd 算法伪代码

输入：矩阵 $\boldsymbol{W}=(w_{ij})$ 是 n x n 矩阵，$w_{ij}=1$ 表示名词短语 np_i 与名词短语 np_j 共指；$w_{ij}=0$ 表示名词短语 np_i 与名词短语 np_j 不共指

输出：矩阵 $\boldsymbol{W}$

```
Transitive-Closure(W)
1   for k = 1 to n
2       for i = 1 to n
3           for j = 1 to n
4               w_ij = w_ij ∨ ( w_ik ∧ w_kj)
5   return W
```

求得任意两个名词短语间是否共指的关系，就可以生成完整的共指链，完成文档内名词短语的共指消解。

3.5 实验结果和分析

本文使用 CoNLL 2012 共指消解国际评测中的中文数据集对本文提出的共指消解性能进行评估。该中文数据集包含 6 类语料，每种语料来源不同，具有不同的特点，其详细信息如表 3-7 所示。

表 3-7 CoNLL2012 共指消解中文数据集信息

语料类别	来源描述	特点
Broadcast Conversation(bc)	广播对话	口语化、代词多、人数多
Broadcase News(bn)	广播新闻	口语化、专有名词多
Magazine(mz)	杂志文摘	结构不统一
Newswire(nw)	新闻电讯	文章规范
Telephone Conversation(tc)	电话对话	代词多、口语化严重
Weblogs and Newsgroups(wb)	博客和新闻讨论组	口语化、结构不统一

该数据包含三部分：训练数据集、验证数据集、测试数据集，其详细情况如表 3-8 所示。

表 3-8 CoNLL2012 共指消解数据集信息

数据集	文件数	句子数	共指链数	共指的名词短语个数
训练数据集	1,391	36,487	28,257	102,854
验证数据集	172	6,083	3,875	14,383
测试数据集	166	4,472	n/a	n/a

对于共指消解的结果的评价方法，主要采用准确率、召回率及 F 值。其中准确率和召回率有不同的定义方法，目前主要有三种：MUC、BCUB 和 CEAFE。MUC 是 Vilain 提出的评价共指消解的方法，该方法中主要以名词短语对为考察对象，通过比较答案和系统产生的结果两者中共指链中共指对的准确率和召回率来计算共指消解的准确率和召回率[41]。B-CUBED 是 Bagga 针对 MUC 评价共指消解的方法中，将名词短语对之间的链接情况不加细分做出的改进，该方法以共指链中名词短语为考察对象，通过衡量每个名词短语在答案和系统产生的结果这两者中的与其共指的名词短语个数的情况来计算准确率和召回率，从而对错误的链接进行了区分，加大了对将两个较大的共指链错误地链接到一起的错误链接的惩罚，更好地衡量了共指消解的性能[42]。尽管 B-CUBED 方法能比较有效的衡量共指消解结果，但是，考虑到将共指消解的核心是链接，针对名词短语的方法会对部分链接进行重复计算，从而不能很好地衡量共指消解结果。针对这个问题，Luo 提出了 CEAFE 方法，

该方法通过使用求二分图的匹配来对答案和系统产生的结果的链接之间建立链接，从而计算准确率和召回率来衡量共指消解的性能[43]。

3.5.1 候选名词短语提取性能实验

候选名词短语的提取是共指消解的基础，其性能直接影响后续共指关系判断。如果候选名词短语提取性能很差，即使后续的共指关系判断性能很好，整体的共指消解性能将会是很差的，因此候选名词短语提取是共指消解的基础。为了衡量本文提出的候选名词短语提取方法的性能，分别在 gold 数据和 auto 数据进行了实验。其中 auto 数据中的词性标注和句法分析是采用自动化的工具进行的，而 gold 数据是在 auto 数据的基础上，针对词性标注和句法分析中的错误进行了人工校正。实验结果如表 3-9 所示。

表 3-9 候选名词短语提取实验数据

数据集	Precision	Recall	F
验证集 Gold	0.740	0.736	0.738
验证集 Auto	0.699	0.643	0.670
测试集	0.731	0.750	0.740

实验结果表明，本文提出的候选名词短语在 gold 数据上的性能要好于在 auto 数据上的性能。这应该是由于本文提出的候选名词短语提取方法主要是依靠句法分析信息来做的而造成的。本文的方法对句法分析信息依赖性很大，而句法分析的基础是词性标注，由于中文中词的词性情况非常复杂，使得词性标注的性能比较差，从而影响了之后的句法分析的性能，进而使得基于本文提出的候选名词短语提取方法在 auto 数据上的性能不如在 gold 数据上的性能。此外，语料中有一部分是比较口语化的语料，对于这些语料，由于词性标注性能比较差，这种情况更为明显。

3.5.2 共指关系判断实验

本文分别在验证集的 gold 数据集、验证集的 auto 数据集和测试集上分别做了实验，实验采用了本文提出的共指关系判断方法，实验结果如表 3-10、表 3-11 和表 3-12 所示。

由实验结果可以看出，在验证集 gold 数据集上的 F 值比在验证集 auto 数据集上的 F 值高了 4.1%。这表明在 gold 数据中准确的词性标注和句法分析对本文采用的方法影响很大，这主要有两方面原因。

表 3-10 共指消解在验证集 gold 数据集上的实验结果

Matrix	Precision	Recall	F
MUC	0.689	0.666	0.677
BCUB	0.755	0.741	0.748
CEAFE	0.488	0.511	0.499
OF-Develop-Gold			0.641

表 3-11 共指消解在验证集 auto 数据集上的实验结果

Matrix	Precision	Recall	F
MUC	0.654	0.564	0.606
BCUB	0.781	0.681	0.728
CEAFE	0.432	0.510	0.468
OF-Develop-Auto			0.600

表 3-12 共指消解在测试集上的实验结果

Matrix	Precision	Recall	F
MUC	0.693	0.660	0.676
BCUB	0.777	0.733	0.754
CEAFE	0.507	0.539	0.625
OF-Test			0.651

一方面，准确的词性标注和语法分析有助于前期候选共指名词短语的提取。候选名词短语提取是判断名词短语对之间共指关系的前提，只有将共指的名词短语对提取出来，作为候选名词短语，该名词短语与其他名词短语之间的共指关系才会被判断，对应的共指关系才有可能被正确判断出来，该名词短语对应的共指关系才能完全。

另一方面，准确的词性标注和句法分析为共指关系的判断提供了准确的特征，有助于名词短语之间的共指关系判断。从数据的角度来看，准确的词性标注和句法分析使得词性信息和句法信息更能反映数据内部的特点，关于词性和句法的数据噪音比较少，同样的特征提取方法，同样的训练模型，学习出的分类器能够更好地反映数据的特点，因此，在 gold 数据集上能够获得更好的共指消解性能。从模型的角度来看，准确的词性标注和句法分析的数据其内部的规律更为简单，同样的特征提取方法，同样的训练模型，总是能够得到更好的分类性能，因此，在 gold 数据集上的结果要好于在 auto 数据集上的结果。

数据集中包含了 6 种语料（广播对话 bc、广播新闻 bn、杂志文摘 mz、

新闻电讯 nw、电话对话 tc、博客与新闻讨论组 wb），这些语料分别来自不同的应用场景，语言的使用和句法结构差异很大，因此，这 6 种语料分别具有不同的特点。如果将这 6 种语料不加区分，使用所有的训练数据，使用机器学习的方法训练一个处理 6 种语料的分类器，这个分类器将不仅不能很好的利用候选名词短语对属于的语料类别，而且因数据中的情况过于复杂而影响分类性能。针对这个问题，本文对 6 种语料进行分别处理，对于每一种语料，使用该语料的训练数据来训练处理该语料的分类器，将 6 种语料的共指消解问题分解成为 6 个分问题。此外，由于实验中使用的数据集数量比较大，对于上述的分别训练的策略，依然能够保证训练充分。本文采用的共指消解方法在 6 种语料上的实验结果如表 3-13 所示，其中 all 表示采用这种 6 种语料分别对待的策略的综合结果。

表 3-13 在验证集中 6 种语料上的实验结果

语料	MUC	BCUB	CEAFE	OF
bc	0.633	0.664	0.377	0.558
bn	0.737	0.780	0.515	0.677
mz	0.472	0.682	0.462	0.539
nw	0.859	0.867	0.754	0.827
tc	0.793	0.769	0.462	0.675
wb	0.535	0.670	0.389	0.531
all	0.677	0.748	0.499	0.641

实验结果表明，在 6 种语料上的共指消歧的性能差别比较大，语料博客与新闻讨论组 wb 上达到了 CoNLL2012 评价指标 0.531，而在语料新闻电讯 nw 上达到了 CoNLL2012 评价指标 0.827，总的结果达到了 0.641。

对这 6 种语料的结果与它们的语料特点的对比表明，不同的语料由于其特点得到了不同的共指消解性能。对于语料博客和新闻讨论组 wb 和语料杂志文摘 mz，其共指消解性能最低，而这两种语料的共同特点是语料中的包含的句子结构差异性很大，句子结构复杂使得很难学习到性能特别好的分类器。在语料广播对话 bc、广播新闻 bn、电话对话 tc 和博客和新闻讨论组 wb 中，由于语料中句子的表达非常口语化，使得句子的结构不是特别完整且包含非常多的语气词，对这样的句子的句法分析性能比较差，由于这样的原因，学习到的分类器分类性能将受到一定的影响。语料广播对话 bc 和电话对话 tc 的共同特点是代词比较多，有实验分析可知，该方法能够比较好地处理代词之间的共指。语料新闻电讯 nw 中文章比较规范，其中的结构完整，用语比较

规范，本文提出的方法能够更好地提取有用的特征，共指消歧性能很好。

本文这部分的工作参加了 CoNLL-2012 中文指代消解评测。图 3-4 给出了本文系统性能与其他参加本次评测系统性能的比较。本文系统图中标为 Xu。在其他参评系统中，系统 Chen 结合了基于规则的方法和基于分类的机器学习方法，还使用了多遍过滤的方法进行过滤，CoNLL2012 评测指标达到 0.708[44]。Yuan 采用人工总结的规则构建的决策模型来判断候选的名词短语之间是否存在共指关系，该方法在 CoNLL2012 评测指标达到 0.664[45]。

实验结果可以看出，想要得到比较好的共指消解性能，不能单独使用基于统计的方法或者基于规则的方法，将两种方法结合起来可以得到更好的性能。Chen 和 Yuan 之所以能够达到好的结果，主要是因为系统中添加了更加丰富的规则。以规则为主的共指消解方法中利用了数量庞大的规则，一方面，规则的总结需要大量的人力，既耗时，实施难度大；另一方面，规则的总结对训练语料的依赖性很大，不同的语料，用语和表达可能差别很大，之前的规则可能就不能达到同样好的性能了。

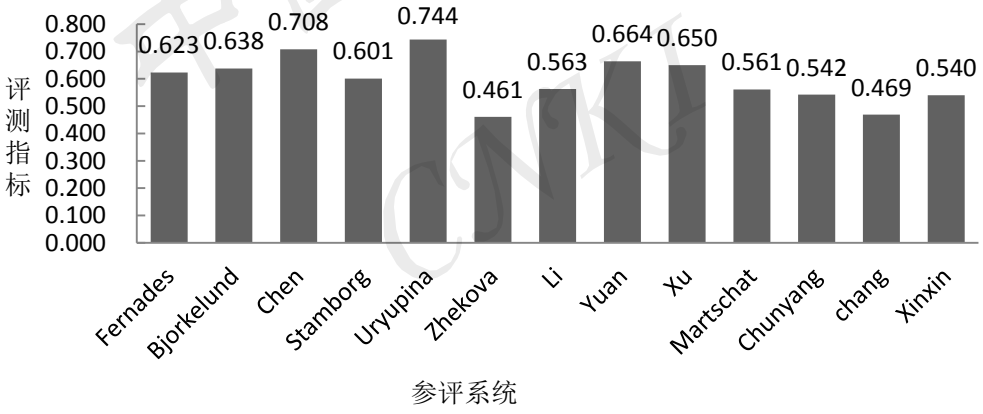

图 3-4 CoNLL2012 共指消解评测系统性能对比

3.6 本章小结

本章介绍了一种统计与规则相结合的中文文档内共指消解方法。首先抽取候选的名词短语，然后使用基于统计的方法，对名词短语对是否存在共指关系进行判定，再用人工总结的规则对符合规则的名词短语对的共指关系进行校正，最后形成共指链，完成文档内的中文共指消解。在 CoNLL2012 共指消解中文数据集上对该方法进行了评估

第 4 章 跨文档中文人名消歧

4.1 引言

待消歧人名字串出现在不同的文档中使得该待消歧人名在文档之间存在歧义，这就是跨文档的人名歧义。跨文档的中文人名消歧就是要针对待消歧人名，弄清楚包含该待消歧人名字串的文档之间，待消歧人名字串的含义。由于包含待消歧人名字串的文档之间公共的信息有限，想要分析清楚各个文档中待消歧人名的情况常常面临知识短缺问题。

人名消歧的早期研究主要是利用文档内词语作为特征度量文档中人名同实体的相似度，利用人名上下文中词语的差异实现同名人物的区分。但这种方法往往受到信息缺失问题的影响。一方面，某些人名的上下文中提供的信息量较少，影响到分类精度；另一方面，人往往具有多种身份，导致即使对于同一个人的上下文也可能出现显著差别。这就意味着单纯使用文档内词语的人名消岐方法存在一定的问题。针对这一问题，引用更多的补充信息有望缓解人名消歧面临的知识短缺问题，来提升人名消岐的性能。这里，观察包含“高峰”的文档例 4-1 和例 4-2：

例 4-1 *北京时间 12 月 29 日下午,中超北京国安举行建队 20 周年庆祝典礼。图为快马高峰现身典礼。*

例 4-2 *高峰那英之子高兴萌照曝光，已经八岁的他神情可爱，眉宇间神似浪子高峰。*

对应这些例子，首先可以考虑利用百科知识中提供的实体词条丰富实体信息。百度百科词条“高峰”中包含 58 个义项，前 10 个如表 4-1 所示。

文档例 4-1 中提到“北京国安”，而第 5 个义项“足球运动员”的描述中包含“北京国安”，显然可以确定例 4-1 中高峰指向义项 5。从这个例子可以看出是百科知识描述信息有助于人名消岐的实现。但是，文档例 4-2 未提及任何与足球相关的信息，在不使用任何外部资源的情况下，因缺乏“高峰”和“那英”相关的知识而无法将该文档中“高峰”链接到实体“足球运动员”，而如果我们在互联网以“高峰 那英”为关键词检索，就会看到很多结果网页中提及“高峰”的足球背景，图 4-1 为这些检索结果中的一条，利用这些互联网提供的知识，就能够将人名“高峰”链接到实体“足球运动员”，即使

在上下文中未出现任何“足球”相关词语。

表 4-1 百科知识词条“高峰”前 10 个义项

编号	标题	描述信息片段
1	相声演员	男，著名相声、快板演员，别号，高老板…
2	著名歌手	华语男歌手，1969 年出生于辽宁沈阳…
3	上海交通大学教授	国家杰出青年基金获得者…
4	中国矿业大学力学与建筑工程学院副院	（1965.6-）男，中国矿业大学教授…
5	足球运动员	主要效力于北京国安队。速度快…
6	影视演员	山西省大同市人，毕业于解放军艺术…
7	河北省女新闻工作者协会会长	笔名：戈红，女，1928 年生，汉族…
8	柔道运动员	出生于辽宁抚顺，2004 年雅典奥运会…
9	影视导演	《美丽家园》、《王长喜来了》…
10	云南省人民政府副省长	生于云南个旧，祖籍云南石屏…

针对知识短缺问题，本章提出了基于互联网的跨文档的中文人名消歧方法。利用互联网中丰富的文档，为人名消歧过程提供知识补充，从而达到更好的人名消歧性能。最后在 CIPS-SIGHAN2012 中文人名消歧的评测数据集进行了实验和评估。

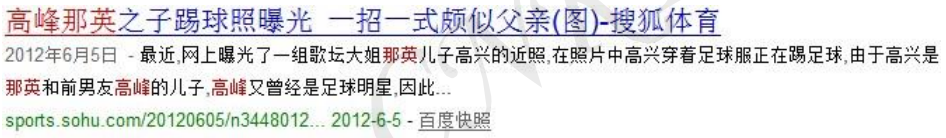
高峰那英之子踢球照曝光 一招一式颇似父亲(图)-搜狐体育
2012年6月5日 - 最近,网上曝光了一组歌坛大姐那英儿子高兴的近照,在照片中高兴穿着足球服正在踢足球,由于高兴是那英和前男友高峰的儿子,高峰又曾经是足球明星,因此...
sports.sohu.com/20120605/n3448012... 2012-6-5 - 百度快照

图 4-1 关键词“高峰 那英”的检索结果举例

4.2 跨文档人名消歧问题分析

对于一个待消歧人名，给定包含该待消歧人名字串的若干篇文档，要求消除这些文档之间待消歧人名字串的歧义。

待消歧人名字串在文档中的出现并不保证待消歧人名在该文档中出现，这个关系成立的条件是文档中出现的人名字串表示一个人名。待消歧人名字串在文档中却不表示人名的情况主要有两个。待消歧人名字串不以待消歧人名成词，对于待消歧人名“高峰”，句子“加拿大房价已达历史最高峰值”中，虽然出现了待消歧人名字串“高峰”，但其中“高峰”并不独立成词，其成词结构应为“加拿大 房价 已 达 历史 最高 峰值”。另一种情况是待消歧人名字串虽然独立成词，但该词不表示人名，而是普通词，这种情况主要是有些人名字串还有其他普通词含义，对于待消歧人名“高峰”，虽然句子“笔者认为，长期来看，高峰期限行应是较为合理的方案”中“高峰”独

立成词，但是，该词不表示人名，是一个普通词。

对于大多数名人，存在对该名人的介绍，此类介绍文档可以作为人名消歧的知识库。因此，跨文档的人名消歧就是要在人物实体知识库中，对待消歧文档中的人名找到相应的人物实体，将文档中人名与实体知识库中相应的实体之间建立链接。由于知识库对人物实体介绍往往依赖人工添加，而互联网上人名的出现比较随机化，知识库不可能同步更新所有互联网上出现的人物实体的信息。因此，知识库中的人物实体对文档中出现的人名的覆盖度有限，并不是所有的人物实体，在知识库中都存在与之对应的介绍文档。

对于文档中的人名所指向的人实体未被包含于知识库中的文档，其人名消歧就不能直接依赖知识库中人名对应的人物实体的指导了，为了消除这些文档之间的歧义，需要衡量待消歧人名在所在文档的上下文之间的关系，确定这些人名字串是否指向相同的实体。因此，要根据待消歧人名是否指向相同的关系对文档进行聚类，同一个簇中的文档中的人名指向相同的实体。

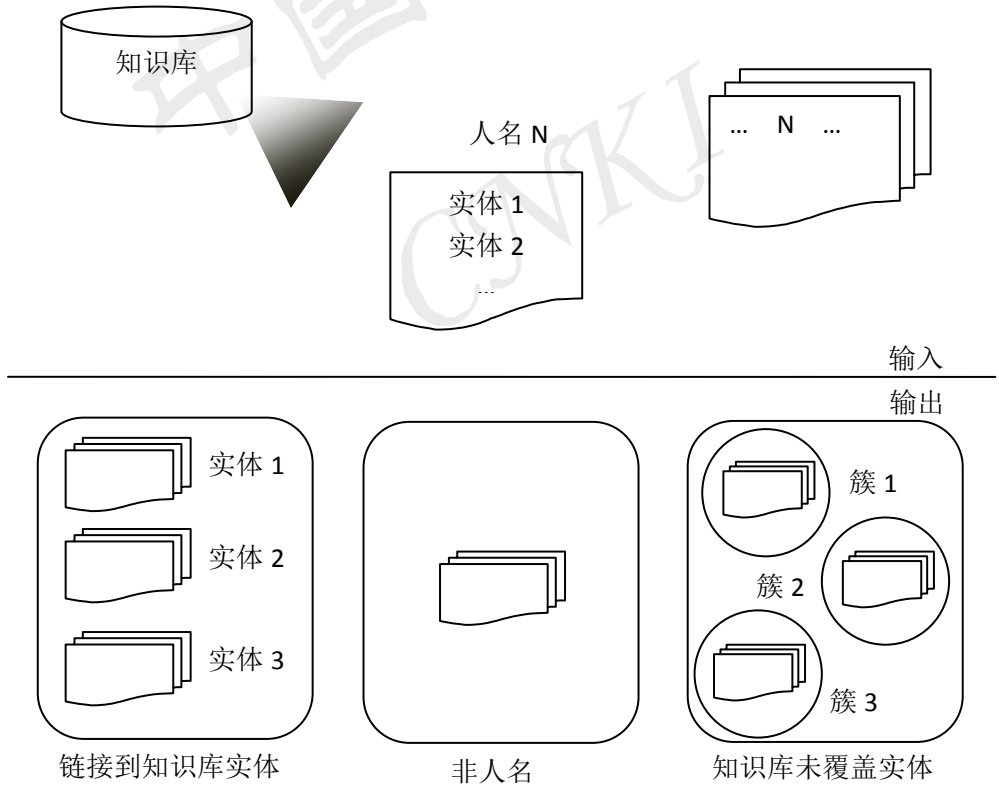

图 4-2 中文人名消歧输入输出

这样，跨文档的人名消歧问题可以表述为：对于一个中文人名 name，有一个包含叫这个人名的若干实体的描述信息的文档，这个文档被称作目标知

识库。对于该人名，有若干篇包含该人名字串的文档。人名消歧就是要消除这些包含该人名字串的文档中该人名字串的歧义，它们属于且仅属于三类中的一种。第一类，文档中出现的人名字串代表一个人物实体，且该人物实体的信息出现在目标知识库中。第二类，文档中出现的人名字串代表一个人物实体，但该人物实体的信息未被包含于目标知识库中。第三类，文档中出现的人名字串不代表一个人物实体，那么，该字串要么是一个普通词，要么不成词，是相邻若干词的片段。对于第一类文档，要确定其中的人名字串代表目标知识库中的哪个实体，将待消歧文档与知识库中对应的人物实体链接起来。对于第二类文档，要将待消歧人名代表相同实体的文档聚类到同一个簇中。对于第三类文档，将其做特殊标记，不做额外的处理。这样，跨文档的中文人名歧义就得到了消除。跨文档的中文人名消歧的输入输出如图 4-2 所示。

4.3 基于互联网的跨文档人名消歧的框架

针对前面介绍的跨文档人名消歧问题，本课题提出的方法按照如下流程来解决该问题：首先使用第二章介绍的人名消歧方法对待消歧文档中的待消解人名进行人名识别，被识别为普通词的待消解人名字串所在的文档，直接分为“非人名”类，不参与之后的处理，那么，剩余的文档中都包含待消歧人名；然后，对于每一篇剩余的文档，即待消歧人名字串表示人名的文档，衡量文档中待消歧人名指向的人物实体与目标知识库中的人物实体的相似度，相似度满足一定条件的，将该文档链接到目标知识库中的相应的实体，即被分为“实体 X”的文档；最后对未链接到知识库中实体的待消解文档，根据待消歧人名在它们中是否指向相同的实体对它们进行聚类，同一个簇中的文档中的待消歧人名指向相同的实体。人名消歧的整体框架如图 4-3 所示。

对于某一待消歧人名，给定的文档都是包含待消歧人名字串的，为了消除文档之间该待消歧人名字串的歧义，必须找出包含待消歧人名的文档，去掉那些待消歧人名字串出现但不表示人名的文档。这里采用本文第二章提出了基于规则后处理的人名识别方法对文档中的待消歧人名进行识别，被识别为待消歧人名的字串所在的文档将被保留，未被识别为待消歧人名的字串所在的文档将被标记为“非人名”类别，不参与之后的处理。

经过上一步的待消歧人名识别，找出了文档集中包含待消歧人名的文档，这些文档中的待消歧人名都指向某个现实中的人物实体，通过衡量文档中待消歧人名指向的实体与目标知识库中每个实体的相似度来将文档链接到相应

的目标知识库中的实体。文档中待消歧实体的出现包括两种形式：待消歧人名和其他词语指代。为了更好的确定人名上下文，先采用第 4 章的共指消解方法对文档进行处理，提取待消歧人名在该文档中的共指链，并以此将文档中该实体的指代都还原为该待消歧人名。为了衡量实体之间的相似度，使用百科知识丰富知识库中的实体介绍，然后分别使用不同类型的特征，得到不同的链接结果，训练分类器将结果进行合并，该方法将在 4.3 节中详细介绍。然后针对链接结果，分别从文档和实体描述中抽取关键词，再结合待消歧人名一起组成查询关键词，使用搜索引擎进行检索，通过对检索结果进行分析，对链接结果进行验证，该方法将在 4.4 节中详细介绍。

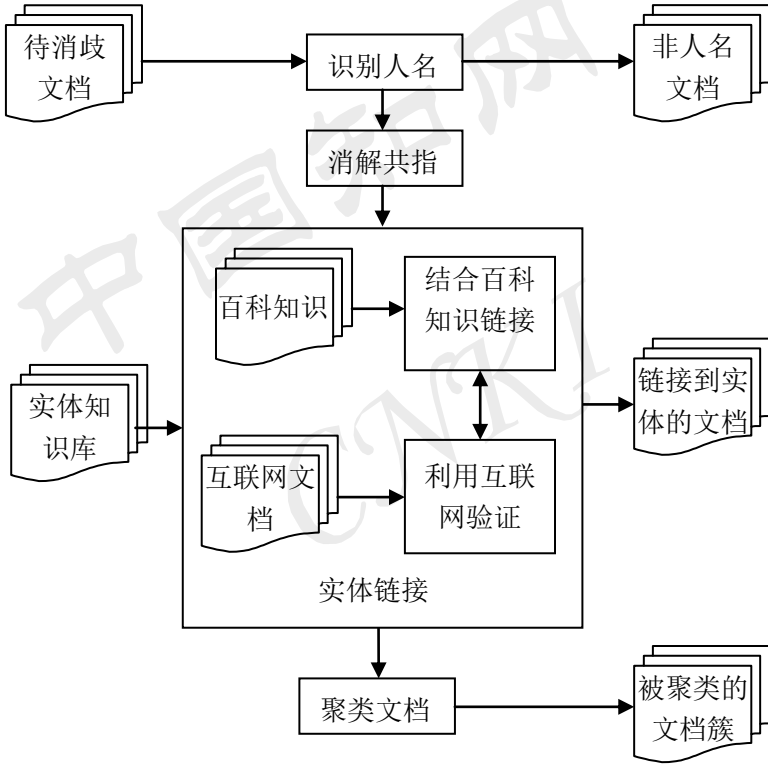

图 4-3 人名消歧整体框架

经过上述的实体链接后，还是有文档未被链接到目标知识库中的实体，这主要是因为目标知识库中实体有限，不能覆盖所有的实体。对于这些文档，根据其中待消解人名指向实体是否相同进行聚类。假设：对于同一实体，在互联网中出现的包含该实体的众多文本中，在实体附近几乎总是共现很多命名实体。基于这个假设，待聚类的文档中待消解人名附近的命名实体被抽取出来，根据文档间是否有相同的命名实体决定文档是否被分到一个簇中。这样，知识库中未包含的实体也被聚类了。

4.4 基于百科知识的中文人名消歧

人物知识库中包含对于某一实体人物的简短介绍，对于某一人名，在知识库中可以找到该人名所有实体的介绍文档。通过衡量待消解文档中人名的上下文与知识库中该人名个实体介绍的相似度，可以对文档中的该人名进行分类，将其链接到相应的实体，从而消解了人名歧义。由于知识库中的人物介绍普遍比较简短，包含的信息非常少，与之不同的是，待消解文档中提到该人物实体信息范围比较广。这样，知识库中人物介绍信息相对于待消解文档中的信息的覆盖度很低，给人名消歧带来很大的困难。

互联网上存在着丰富的信息，找到需要的信息，添加到人名消歧的过程中可以缓解知识短缺问题。为了解决知识库对于人物信息覆盖度低的问题，本文提出了使用互联网百科信息，来丰富知识库中人物信息。本节首先介绍了百科数据的获取和整理，然后介绍了利用百科进行人名消歧。

4.4.1 百科数据获取与整理

互联网上包含丰富的信息，可以给人名消歧提供很大的知识支持，一定程度上缓解人名消歧过程中知识短缺问题。互联网上存在着多个自由百科全书，这些百科全书由人们自由手动编辑，多个人反复修改和订正，已经形成了一个内容丰富、知识面广、非常可靠的电子百科全书。其中中文的百科全书有：维基百科中文版、百度百科等。其中百度百科结构清晰，内容丰富，可靠性较高。对于一个人名，百科里包含了几乎所有有名的叫这个名字的人物的信息，越是有名的人，关于该实体的信息越丰富。为了利用这些信息，编写了一个简易的爬虫，爬取百度百科的页面。使用这个方法，收集到了百度百科 4,000,000 个页面，约 120,000 个人名。

对于大多数待消歧的人名，在原有的知识库中存在一个该人名对应人物实体的介绍信息的列表，在百度百科中该人名的词条中也包含该人名对应人物实体的介绍信息列表。为了将百科数据作为外部知识来帮助人名消歧，需要从获取的百科数据中找到相应的人名页面之后，将该百科页面的人物实体与目标知识库中的人物实体间简历链接，一一对应起来。

为此，利用 Vector Space Model (VSM) 对两边的文档建模，使用 TF-IDF 加权，计算向量相似度，找到两组人物实体之间的最可能的链接。一方面在目标知识库中的实体需要找到尽可能相似的百科实体，另一方面，百科实体也要找到尽可能相似的目标知识库中的实体。由于目标知识库中实体的描述

信息大多比较简短，而百科中实体的信息篇幅却比较长，使用相似度计算并不能直接得到两者的对应关系。为了解决可能发生的冲突，在选择对应链接的实体时，采用按相似度从高到低的顺序来选择，一旦选到相应的实体，该实体和与该实体链接的实体都被移除出来，这样就得到了一个最可能的两组实体间的对应关系。这样就使用互联网百科信息对目标知识库进行了丰富。

4.4.2 利用百科知识的人名消歧

为了将待消解文档与目标知识库中的实体链接起来，衡量了文档中待消歧人名字串与目标知识库实体的相似度，为文档选择目标知识库中相关度最高的实体进行链接。为了达到更好的链接性能，运用了 3 类不同的信息来计算文档中待消解人名字串与目标知识库中实体的相关度。

第一类特征是词语特征。这类特征是使用待消歧人名所在文档与实体描述文档中的词语作为特征，进行实体相似度衡量的。为了使特征能够准确的体现待消歧人名的上下文信息，本文设计了一种基于人名的上下文匹配判别的加权 TF-IDF，记为 TF-IDFw，其计算方法如公式 4-1 所示。通过计算相似度，为每个待链接文本选择相似度最高的实体。

$$TF-IDF_w(t_i, name, d_j) = \frac{TF(t_i,d_j)\cdot IDF(t_i)}{\left(1+\alpha\cdot NotSameParagraph(t_i,name,d_j)\right)} \tag{4-1}$$

其中$TF(t_i, d_j)$表示词语t_i在文档d_j中的频率，$IDF(t_i)$表示词语t_i的倒文档频率，$NotSameParagraph(t_i, name, d_j)$反应词语$t_i$与人名 *name* 在文档$d_j$中是否在同一段落中出现过，如果出现过，该函数值为 0，否则为 1。α为权重调节参数，用于反映词语t_i与人名 *name* 之间的正相关程度，α值越大则代表 *name* 的上下文词语对链接到个体的判别权重越大。通过试验，α选择经验值 1.5。$TF-IDF_w$的值为[0,1]。该加权策略的目的是提高人名所在的同段落词语的权重，也就意味着同段落词语对人名消歧的决定性更大。

第二类特征使用作品名称信息。通过分析实体相关的作品实体在文档中的出现情况计算实体相似度。由于待消歧人名实体多为名人，提到该实体的文档中，往往有与之相关的作品实体。例如：艺术家有艺术作品，政府官员有刊物，歌手有歌曲，学者有论文和刊物会议等等。由于人的这些作品名称对于实体具有非常好的区分性，通过考察待消歧人名所在文档与知识库中实体描述信息中共有的作品名称情况，能够衡量该文档中人名是否指向该实体。

首先，从百科实体中提取出相应的作品实体名。因为作品实体名在人物

实体描述信息中多以书名号括起来，所以，作品名的提取只需将其中所有书名号中的短语提取出来。然后，在待消解的文本中找这些实体。如果待消解的文本中出现了这些作品实体，那么待消解的文本很有可能与出现的作品实体相对应的人物实体相关的。这样，待消解文本中作品实体名的出现将成为链接待消解文本和人物实体的证据。例如，对于待消歧人名“高峰”，存在一个实体为中央电视台副台长，从知识库中关于该实体的描述中可以知道，他导演了纪录片《闯江湖》。通过使用此类特征，如果“高峰”的上下文中出现了“《闯江湖》”，就可以将该上下文中的人名“高峰”链接到实体中央电视台副台长“高峰”。为了消除分词错误给寻找这些证据带来的消极影响，系统在未分词的待消解的文本中寻找这些提取的作品实体名。

考虑到作品名出现的上下文是否与待消解的人名有关，系统对寻找的范围进行了限制，这个范围设定为文本中待消解人名先后 40 个字的范围。这样处理背后的假设就是：离待消解人名越近的信息与该人名相关度越高，消解能力越大。证据最多的实体被链接到该待消解文本。

第三类特征使用百科词条中实体的标题信息。在爬取的百科人物页面中，对每一个人物实体都有一个标题，这个标题对该实体的身份打了标签。例如：“柔道运动员”、“南京大学副教授”等等。这个标题对消解人名歧义很有用。另外，在文本中，待消解的人名出现常常在其附近能够找到相应的身份词。如：“《恋爱中的宝贝》的演员班底：白雪、吴军、沈畅、黄觉等”，“作为国内著名的军旅歌手白雪”等等。基于这样的观察，系统分两个步骤利用身份词消歧：首先，从百科页面的人物实体列表的标题中提取每个人物实体的身份词；然后，在待消解的文本中寻找这些身份词。基于相同的考虑，寻找的范围被设定来确保寻找的范围属于待消解人名的上下文。

上述三类特征使用的信息不同，分别从不同的角度评估了文档中待消歧人名指向的实体与知识库中的实体之间的相似度，使用这三种特征分别可以得到了自己的链接结果。为了结合着三类特征的的链接结果，本文设计训练了一棵决策树分类器，用于将上述三种链接结果综合起来，达到更好的链接结果。此外，上述每种特征都使用独立阈值来使结果具有比较高的准确性。

4.5 结合互联网验证的中文人名消歧

考虑到百科知识的规模和覆盖度限制，在一些情况下人名消岐仍然面临信息缺乏的问题。利用互联网的资源补充知识和信息，有助于突破常规人名消歧方法的性能瓶颈，进一步解决人名歧义。观察发现互联网上相关的信息

可以经常在网页中共现，而不相关的信息则很少在网页中共现。例如：对于待消歧人名“高峰”有两个实体为“相声演员”和“足球运动员”，人名“那英”与实体“足球运动员”相关。如果在搜索引擎中检索“相声演员 高峰 那英”，前 10 个结果中，只有一个结果中同时覆盖了“高峰”和“那英”这两个词，进入该结果发现，“高峰”和“那英”来自该页面的不同板块。如果检索“足球运动员 高峰 那英”， 搜索引擎返回的前 10 个结果网页中，都包含了检索关键词中的那三个词，而且这三个词紧邻出现。这就意味着互联网验证可以降低人名歧义性。在这一方法中，为了提高人名消歧的性能，我们首先构造查询串，利用搜索引擎获得检索结果，然后利用训练得到的分类器对是否采纳义项候选进行决策。

4.5.1 查询关键词构造及检索结果处理

为了验证人名链接的结果，需要衡量给定文档中人名与被链接到的实体在互联网中是否真的存在其证据。为了检索到这些证据，检索关键词中应包含实体的信息和文档中人名的信息。针对两者的情况以及搜索引擎的特点，本文采用了细化的关键词提取方法，使检索关键词尽可能准确地、无歧义地反应实体和文档的信息。对于实体描述文本，有别名（原名、本名，又叫），提取其别名；有作品的，提取作品名；否则提取“机构+职业或头衔”。对于人名出现的文档，提取文档中的人名或作品名。在文档关键词提取过程中，考虑候选关键词与待消解人名的距离和频率等指标来对候选关键词的权重，选择 2 个权值最高的候选做关键词。实体关键词和文档关键词组成检索关键词。

使用构造的检索关键词，在搜索引擎获取前 20 个检索结果。我们称包含某一人名的多个实体的文档为总结式文档，如各种百科页面以及包含这些页面内容的页面。如果检索结果包含这类文档，验证的结果会很差。因此，必须从原始检索结果中过滤掉这些总结式页面。这里使用从知识库中实体提取出的关键词，用这些关键词进行组合作为检索关键词，搜索引擎返回的检索结果中就包含大量的这样的总结式页面，检索结果中检索关键词的覆盖程度高的页面即可以认为是总结式页面。这样，我们获得了总结式页面的 URL 列表，利用此列表对的验证过程中的检索结果进行过滤，得到最终的检索结果。

4.5.2 基于互联网检索结果的人名消歧验证

检索结果是将文档中待消解人名指向的实体链接到知识库中某实体的证据，利用这些证据，可以衡量链接结果的可信度。从检索结果的分析来看，

有两个特点：一、在实体是文档中人名对应的实体情况下，检索结果对来自两方的关键词覆盖率都很高；二、在实体不是文档中人名对应的实体情况下，检索结果对双方的关键词覆盖性差。基于上述的观察，可以使用检索结果对链接结果进行验证，来提高链接的性能。 为了更好地利用这些证据来给出准确的验证结果，本文设计和训练了分类器，来对结果做出是否采纳的判断。

对于待消歧人名 *name* 和百科中的第 *i* 个实体e_i，分别从百科中获取实体描述关键词 *keywords*(e_i)，从待消歧文本获取上下文关键词 *keywords*(doc_j)。利用 *name*、*keywords*(e_i)、*keywords*(doc_j)构造查询关键词，如果返回的有效检索结果$record_n$ (n=1…20)对实体描述关键词和文本上下文关键词均有一定覆盖度（依据实验设置经验值 0.75），则认为$record_n$为支持待消歧人名链接到当前实体的一个检索结果，那么，可以定义一个这些参数上的查询验证函数$IsApproved\left(record_n, keywords(doc_j), keywords(e_j)\right)$返回 1，否则返回 0。在此函数的基础上，检索结果中总支持数的计算方法如公式 4-2 所示。

$$Support(name, doc_j, e_j) = \sum_{n=1}^{20} IsApproved\left(record_n, keywords(doc_j), keywords(e_j)\right) \quad (4\text{-}2)$$

利用上述的针对检索结果的总支持数的计算方法，本文使用训练数据训练一个二值分类器，用于人名消歧结果的验证。分类器的使用的特征及其对应的取值范围如表 4-2 所示。该分类器使用训练数据和针对训练数据构造的检索关键词，还有相应的查询结果进行训练。

表 4-2　分类器特征及其值域

特征	特征值域
待验证消歧实体与检索结果支持数最高个体是否一致	0 / 1
待验证消歧实体在对应的检索结果中的支持比例	(0,1)
其他消歧实体在对应的检索结果中的最高支持比例	(0,1)
待验证消歧实体检索结果中个体描述关键词与上下文关键词满足最小距离阈值的比例	(0,1)
其他消歧实体检索结果中个体描述关键词与上下文关键词满足最小距离阈值最高的比例	(0,1)
待验证消歧实体对应检索结果中消歧人名与上下文关键词满足最小距离阈值的比例	(0,1)
其他消歧实体对应检索结果中消歧人名与上下文关键词满足最小距离阈值的最高比例	(0,1)
待验证消歧实体对应检索结果中消歧人名与实体描述关键词满足最小距离阈值的比例	(0,1)
其他消歧实体对应检索结果中消歧人名与实体描述关键词满足最小距离阈值的最高比例	(0,1)

使用训练数据训练后获得一个分类器，可以按照结合百科知识后的基线系统输出结果的置信度从高到低，依次对待验证消歧结果进行判别。该分类器为二值分类器，输出包含两种情况：如分类器输出为 1，表示将文档链接到该实体可行度比较高，采纳当前待验证消歧实体作为最终消歧结果输出，将相应的文档链接到该实体。如分类器输出为 0，表示将文档链接到该实体的可行度比较低，拒绝当前待验证消歧实体，并继续验证置信度次高的待消歧个体，直到发现分类器输出为 1 的个体作为消歧结果输出。如所有候选个体均无法通过验证，则输出结合百科知识后的基线系统输出的最高置信度个体作为最终消歧结果。

4.6 实验结果及分析

本文使用 CIPS‐SIGHAN 2012 人名消歧任务训练集为实验数据，该数据集包含 16 个待消歧人名。对于每个待消歧人名，数据包括一个该待消歧人名的实体列表，对其中的每个实体，都有一段简短的描述，对于每个待消歧人名，数据中有包含该待消歧人名的 47 到 195 篇文档，共 1,634 文档。

对于给定待消歧人名 $n \in N$（N 为人名集合），有包含该待消歧人名字串 n 的文档集 T，链接到同一实体的文档组成类“实体 X”（$1 \leqslant X \leqslant m$, m 为数据中待消歧人名 n 对应的已有的实体个数），文档中待消歧人名字串 n 不是实体而是普通词时，文档归为“非人名”类；文档中待消歧人名字串 n 是实体名但在给的实体列表中未定义，聚类为“簇 X”，X 为类编号，同一编号的文档中人名 n 表示同一实体。评价标准首先计算每个文档的准确率和召回率，然后对待消歧人名 n 的所有文档的准确率和召回率求平均得到待消歧人名 n 的准确率和召回率，如公式 4-3，4-4 所示，对待消歧人名集合 N 中的所有待消歧人名的准确率和召回率求平均得到总的准确率和召回率，如公式 4-5，4-6，总的 F 值计算按公式 4-7[46]。

$$Precision(n) = \frac{\sum_{t \in T} Precision(t)}{|T|} \tag{4-3}$$

$$Recall(n) = \frac{\sum_{t \in T} Recall(t)}{|T|} \tag{4-4}$$

$$Precision = \frac{\sum_{n \in N} Precision(n)}{|N|} \tag{4-5}$$

$$Recall = \frac{\sum_{n \in N} Recall(n)}{|N|} \tag{4-6}$$

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} \quad (4\text{-}7)$$

我们采用 8 个人名的数据做训练集，其余 8 个人名的做测试集。为了更好地衡量方法有效性，以 4 个人名为一组，每次选两组做训练集，另外的两组做测试集，交叉验证。

为了评估本文提出的方法的人名消歧性能，本文引入了一个基线人名消歧方法。基线方法采用向量空间模型，将待消歧人名所在的文档和知识库中的实体描述都表示成特征值的向量，其中特征采用名词、动词以及专有名词作为特征，特征值的计算方式使用 TF-IDF，相似度采用余弦相似度计算，之后采用层次聚类的方法对待消歧人名指向的实体未被知识库覆盖的文档进行聚类。基线方法在该数据集上的 Precision 为 73.3%，Recall 为 76.2%，F 值为 74.7%。

本文提出的使用百科作为外部知识的人名消歧方法，该方法在数据集上的结果与基线系统的结果对比如表 4-3 所示，其中还对比了使用共指消解前后性能的变化。

表 4-3　结合百科知识的人名消歧实验

实验系统	进行共指消解	Precision	Recall	F
基线系统	—	0.733	0.762	0.747
基线系统+百科知识	否	0.826	0.764	0.794
基线系统+百科知识	是	0.830	0.774	0.801

对比结果表明，在人名消歧任务中使用百科作为外部知识对结果提升明显。进一步观察可以发现，引入百科知识后，准确率提升了 9.7%，F 值提升了 5.4%。这说明利用百科知识的方法有效地提高了人名链接的精度。对比实验结果可知，使用共指消解使得准确率提升了约 1.4%，F 值提升了约 0.7%。

对共指消解使用实验进一步分析可知，本文提出的加权的 TF-IDF 的特征加权方法，对待消歧人名出现的段落的特征给予了比较高的权重，因此，在准确地对特征加权以便于更好地表示文档中待消歧实体的特点过程中，待消歧人名出现的判断起着非常大的作用。而共指消解将文档中对待消歧实体的指代都还原为了待消歧人名，从语义层面上得到了待消歧人名在该文档中的出现，从而能够更加精确地确定待消歧实体的上下文，提升人名消歧性能。

进一步分析表明，由于数据中知识库中实体的描述信息普遍比较简短，

包括的信息比较少，在衡量这些实体与文档中待消歧人名的相似度过程中，能够提取的有用特征比较少，因此不能很好地衡量他们之间的相似度，使得将文档中的待消歧人名链接到知识库中实体的准确度比较低。通过利用百科知识，对知识库中的实体描述信息进行了扩充，在链接过程中，能够提取到的有用特征增加，能够更好地衡量文档中人名与各个实体之间的相似度，因此，链接的准确率得到提高，从而使得人名消歧总的准确率得到提升。

为了评估利用互联网进行人名消歧验证方法的性能，设计了一组实验，其结果如表 4-4 所示。其中“基线系统+互联网验证”表示系统利用互联网验证的方法对基线系统的人名消歧结果进行了验证。为了考察总结式页面过滤对结果的影响，对有互联网验证方法参与的实验针对是否采用总结式页面过滤做了对比实验。另外，为了衡量验证分类器训练方法的选择，分别使用朴素贝叶斯（Naïve Bayesian, NB）、决策树（Decision Tree, DT）和支持向量机（Support Vector Machine, SVM）三种方法做了对比实验。

表 4-4　结合互联网验证的人名消歧实验

实验系统	检索结果处理	分类器	进行共指消解	Precision	Recall	F
基线系统	—	—	—	0.733	0.762	0.747
基线系统+互联网验证	直接检索结果	NB	否	0.710	0.729	0.719
			是	0.711	0.729	0.720
		DT	否	0.717	0.734	0.726
			是	0.717	0.734	0.726
		SVM	否	0.715	0.732	0.723
			是	0.714	0.731	0.722
	总结式页面过滤后结果	NB	否	0.739	0.816	0.775
			是	0.739	0.817	0.776
		DT	否	0.744	0.820	0.780
			是	0.743	0.818	0.779
		SVM	否	0.745	0.821	0.781
			是	0.753	0.826	0.788

实验结果显示，引入互联网验证的方法对人名消歧结果有了提升，准确率提升约 2.0%，召回率提升约 6.4%，使得 F 值提升约 4.1%。对消歧结果进行进一步观察发现，本文提出的互联网验证的方法很好的召回了那些本应链接到实体却被用来聚类的文档。

这说明这基于阈值和相识度的方法中，有些文档因其中待消解人名与所属实体的相似度过低而被进行聚类。这可能是由于文档中虽然提及了待消解人名，但并非以人名为主要论述点，甚至只是对该人名有一次提及，而且文档的主题与该实体的主题不是词级别上相关的，导致文档中人名与该实体的相似度很低。而本文提出的互联网验证的方法在一定程度上为这种情况补充了其缺乏的知识。实验结果显示，只采用互联网验证方法前后，F 值提升 3.4%，进行共指消解后，再采用互联网验证方法前后，F 值提升了 4.1%。这表明此方法可以稳定提升人名消歧的精度。

采用不同分类器的实验结果比较表明，支持向量机的方法取得了最好的人名消歧性能，高出了决策树方法与朴树贝叶斯方法的结果。

此外，分别用基线系统与“直接检索结果”和“总结式页面过滤后结果”的人名消歧结果做对比可知，“直接检索结果”的方法 F 值降低了约 2.4%，而“总结式页面过滤结果”的方法 F 值却提升了约 3.4%。实验表明如不对检索结果中的总结式页面进行过滤，不仅不能提升人名消歧的精度，反而使其降低。这一方面说明检索结果中大量存在的总结式页面对于利用互联网验证的人名消歧方法是很大的噪音；另一方面也说明本文提出的总结式页面过滤合理，能够有效地过滤掉噪音结果。

结合共指消解的互联网验证方法能够提升人名消歧性能，进一步分析表明，进行共指消解后，文档中对待消歧实体的指代被替换为待消歧人名，在提取关键字构成关键词时，能够更好地选择到更加重要的关键词，使验证更加准确。

最后做实验评估本文提出的在结合共指消解基础上，将基线系统、基于百科知识的方法、利用互联网验证的方法相结合的完整的人名消歧方法的性能，实验结果如表 4-5 所示。

实验结果显示，在对文档进行共指消解之后，利用百科知识的方法对人名消歧结果的准确率提升了 8.7%，F 值提升了 5.4%，再利用互联网进行验证，使得召回率提升了 6.9%，F 值提高了约 3.9%。利用互联网进行验证的方法中，直接使用检索结果仍然会使结果变差。总的来看，本文提出的结合共指消解的人名消歧方法相对于基线系统，人名消歧结果的准确率提高了 10.4%，召回率提升了 8.1%，F 值提升了 9.3%。

进一步的实验分析表明，在利用百科知识的方法的基础上再利用互联网进行验证能够达到更好的人名消歧性能，这说明利用百科知识的方法和利用互联网进行验证的方法能够很好的结合起来，共同提高人名消歧性能。

表 4-5 结合基线系统、百科知识、互联网验证实验结果

实验系统	检索结果处理	分类器	进行共指消解	Precision	Recall	F
基线系统	—	—	—	0.733	0.762	0.747
基线系统+百科知识	—	—	否	0.826	0.764	0.794
			是	0.830	0.774	0.801
	直接检索结果	NB	否	0.808	0.750	0.778
			是	0.8084	0.751	0.779
		DT	否	0.817	0.756	0.785
			是	0.809	0.752	0.780
		SVM	否	0.809	0.753	0.780
			是	0.810	0.755	0.782
基线系统+百科知识+互联网验证	过滤总结式页面后结果	NB	否	0.801	0.837	0.818
			是	0.803	0.838	0.820
		DT	否	0.826	0.831	0.828
			是	0.833	0.846	0.839
		SVM	否	0.824	0.834	0.829
			是	0.837	0.843	0.840

利用百科知识的方法通过提供更多的实体信息，在进行链接时，可以更加细致地刻画和表示实体，从而能够更好的衡量待消歧人名指向的实体与知识库中的实体的相似度，提高了待消歧人名文档链接到知识库中实体的准确率，从而提高人名消歧的准确率。

利用互联网百科知识验证的方法针对人名消歧中知识库中实体描述信息不能覆盖待消歧人名所在文档中的上下文情况，以互联网作为丰富的知识背景，从中寻找将待消歧人名链接到知识库中相应实体的证据，通过分析作为证据的检索结果，将文档中的待消歧人名链接到相应的实体，从而在知识库中实体描述信息缺乏对文档中待消歧实体的具体信息覆盖的情况下，通过检索结果进行知识外延，召回了本应该链接到知识库中实体却由于相似度低于阈值而被聚类的文档，提高了人名消歧的召回率。

对共指消解的作用进一步分析显示，对文档进行共指消解可以将文档中对待消歧实体的指代还原为待消歧人名，可以准确的确定待消歧人名的上下文，特别是召回那些因使用指代代替待消歧人名出现的上下文。准确的上下文可以更好地用于衡量实体之间的相似度，对人名消歧性能的提升有帮助。

本文方法与其他研究者的方法的比较结果如表 4-6 所示, 其中 Liu 为本文

的介绍的方法。Tian 等选择了 12 种与人相关的信息,采用信息抽取的方法从文档中查找这 12 种信息用于确定其指向的实体[47]。Peng 等针对每个待消歧人名，从训练集找出和互联网中整理该待消歧人名对应各个实体的文档,用于训练分类器进行实体链接[48]。Hao 等利用知识库中对实体描述的信息中提取表示实体的关键词用于实体链接[49]。Wang[50]和 Fan[51]都使用了信息抽取的方法提取属性。

表 4-6 不同研究者在 SIGHAN-CIPS2012 人名消歧数据集上实验结果

系统	Precision	Recall	F
Tian	0.926	0.903	0.914
Peng	0.833	0.879	0.856
Liu	0.837	0.843	0.840
Hao	0.686	0.860	0.763
Wang	0.744	0.691	0.717
Fan	0.640	0.680	0.659

实验结果显示，Tian 的方法取得了 F 值 91.4%，Peng 的方法获得了 F 值 85.5%。进一步分析可得，Tian 的方法中，使用的 12 种信息的确定对人名消歧的性能提升非常有利，但这 12 种信息的确定需要对待消歧人名有深入的分析，而分析过程中需要非常多的人工参与。Peng 的使用已经有明确实体指向的文档来丰富实体的描述信息，可以有效的缓解实体知识短缺问题，但文档的选择需要人工的标注。

4.7 本章小结

本章主要针对跨文档的中文人名消歧问题，在共指消解的基础上，提出了一种利用百科知识和互联网验证的人名消歧方法。实验结果表明，这两种方法都能够提升人名消歧性能，将这两种方法结合的人名消歧方法能够稳定地提升人名消歧性能，结合共指消解，能够使人名消歧性能获得进一步的提升。

结　论

本文针对中文人名消歧问题，从人名识别、共指消解和跨文档人名消歧几个方面展开了研究：

第一、对常用中文人名识别方法在非规范句子中和对易于与普通词混淆的人名的识别性能不好的问题，提出了在常用中文人名识别结果的基础上，利用人名构成规则和人名环境信息进行校正的人名识别方法。

第二、本课题针对文档内中文人名歧义问题，提出了一种结合基于统计的和基于规则的方法。该方法首先提取候选名词短语，针对中文表达中中心词后置现象提取共结尾的最长的名词短语作为候选的名词短语；然后利用候选名词短语形成名词短语对。这里，名词短语对的形成采用了两种平衡正例和反例数量悬殊的方法，使得生成的训练数据中正反例比例在可以接受的范围内。为了衡量名词短语之间共指的可能性，提取了多种特征，采用机器学习的训练一个分类器，用于判断名词短语对是否有共指关系；针对符合规则的名词短语对，利用人工总结的高准确率的规则对错误的共指关系判断进行纠正；最后，通过求传递闭包的方法将共指对链接成共指链，实现文档内的中文共指消解。该方法能够很好地消解中文文档内的共指，对消除文档内的人名歧义很有帮助。在CoNLL2012共指消解中文数据集上达到评价指标0.651的成绩。

第三、本课题针对跨文档的中文人名歧义，提出了一种结合百科知识和互联网结果验证的方法。这两种方法都是针对人名消歧过程中因知识短缺而面临性能无法进一步提升而提出的。引入百科知识的方法利用百科知识丰富的资源扩充了知识库中实体的知识，对待消歧人名所在的文档的情况的覆盖性有了提高。利用互联网验证的人名消歧方法则利用互联网海量的知识，通过衡量构造的查询关键词对应的检索结果的情况来衡量链接结果的执行度来做到对人名消歧结果的验证。该方法避免了传统的人名消歧方法引入静态的有限的知识对人名消歧性能提升有限的问题，为人名消歧过程中提供了需要的信息。此外，结合共指消解，可以更加准确地确定待消歧人名的上下文。实验结果表明，该混合方法能够稳定地提高人名消歧性能。在 CIPS-SIGHAN 2012 中文人名消歧数据集上达到 F 值 84.0%的性能。

本文的贡献包括：一方面，通过结合基于统计和基于规则的方法有效的进行了文档内的共指消解，并将其应用到人名消歧中，准确地确定人名的上下文；另一方面，在使用共指消解确定待消歧人名准确上下文的基础上，提

出了结合利用百科知识和利用互联网验证的跨文档人名消歧方法，有效地缓解了人名消歧过程中知识短缺问题，稳定地提高了人名消歧性能。

本文提出的人名消歧方法虽然在一定程度上缓解了人名消歧面临的知识短缺问题，并提升了人名消歧性能，但就人名消歧问题本身，没有能够给出一种更加系统化的、本质的解决消歧中知识短缺问题的方法，该问题还有待今后进一步研究。

参考文献

[1] 孙茂松，高海燕. 中文姓名自动辨识[J]. 中文信息学报. 1995. 9(2)：16-27.

[2] 张俊盛，陈舜德，郑萦等. 对语料库作法之中文姓名辨识[J]. 中文信息学报. 1992. 6(8)：7-15.

[3] 黄德根，杨元生，王省等. 基于统计方法的中文姓名识别[J]. 中文信息学报，2001, 15 (2)：31-37.

[4] Li L S, Huang D G, Li D. Recognizing Chinese Person Names Based on Hybrid Models[J]. International Journal of Advanced Intelligence, 2011, 3(2): 219-228.

[5] Wu Y, Zhao J, Xu B. Chinese Named Entity Recognition Combining A Statistical Model with Human Knowledge[C]//Proceedings of the ACL 2003 workshop on Multilingual and Mixed-language Named Entity Recognition-Volume 15. Association for Computational Linguistics, 2003: 65-72.

[6] Chen L, Zhang H, Li Z A. Chinese Personal Name Recognition Using N-gram Model and Rules[C]//Proceedings of the 7th International Conference on Computing and Convergence Technology, 2012: 450–453.

[7] Hobbs J R. Resolving Pronoun References[J]. Lingua, 1978, 44(4): 311-338.

[8] Brennan S E, Friedman M W, Pollard C J. A Centering Approach to Pronouns[C]//Proceedings of the 25th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1987: 155-162.

[9] Mitkov R. Robust Pronoun Resolution with Limited Knowledge[C]//Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2. Association for Computational Linguistics, 1998: 869-875.

[10] Cardie C, Wagstaff K. Noun Phrase Coreference as Clustering[C]// Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 1999: 82-89.

[11] Lee H, Peirsman Y, Chang A, et al. Stanford's Multi-pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task[C]//Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011: 28-34.

[12] McCarthy J F, Lehnert W G. Using Decision Trees for Coreference Resolution[C]//Proceedings of the 14th International Joint Conference on Artigicial Intelligence, 1995: 1050-1055.

[13] Kobdani H, Schütze H, Schiehlen M, et al. Bootstrapping Coreference Resolution Using Word Associations[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 2011: 783-792.

[14] Rahman A, Ng V. Supervised Models for Coreference Resolution[C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. Association for Computational Linguistics, 2009: 968-977.

[15] Chen C, Junfeng H, Houfeng W. Clustering Technique in Multi-document Personal Name Disambiguation[C]//Proceedings of the ACL-IJCNLP 2009 Student Research Workshop. Association for Computational Linguistics, 2009: 88-95.

[16] Xu J, Lu Q, Liu Z Z. Combining Classification With Clustering for Web Person Disambiguation[C]//Proceedings of the 21st International Conference Companion on World Wide Web, 2012: 637-638.

[17] Yoshida M, Ikeda M, Ono S, et al. Person Name Disambiguation by Bootstrapping[C]//Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010: 10-17.

[18] Xu J, Lu Q, Liu Z. Aggregating Skip Bigrams into Key Phrase-based Vector Space Model for Web Person Disambiguation[C]//Proceedings of KONVENS 2012 (Main track: oral presentations), 2012: 108-117.

[19] Bekkerman R, McCallum A. Disambiguating Web Appearances of People in A Social Network[C]//Proceedings of the 14th International Conference on World Wide Web. ACM, 2005: 463-470.

[20] Han X, Zhao J. Named Entity Disambiguation by Leveraging Wikipedia Semantic Knowledge[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management. ACM, 2009: 215-224.

[21] Bagga A, Baldwin B. Entity-based Cross-document Coreferencing Using the Vector Space Model[C]//Proceedings of the 17th International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, 1998: 79-85.

[22] Song Y, Huang J, Councill I G, et al. Efficient Topic-based Unsupervised Name Disambiguation[C]//Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, 2007: 342-351.

[23] Han X, Sun L. A Generative Entity-mention Model for Linking Entities with Knowledge Base[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 945-954.

[24] Han X, Sun L. An Entity-topic Model for Entity Linking[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012: 105-115.

[25] Gong J, Oard W D. Selecting Hierarchical Clustering Cut Points for Web Person-name Disambiguation [C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009: 778-779.

[26] Wang X, Liu Y, Wang X, et al. Adaptive Resonance Theory Based Two-stage Chinese Name Disambiguation[J]. International Journal, 2012, 2: 83-88.

[27] Tang J, Fong A C M, Wang B, et al. A Unified Probabilistic Framework for Name Disambiguation in Digital Library[J]. Knowledge and Data Engineering, IEEE Transactions on, 2012, 24(6): 975-987.

[28] Tang J, Zhang J, Zhang D, et al. A Unified Framework for Name Disambiguation[C]//Proceedings of the 17th International Conference on World Wide Web. ACM, 2008: 1205-1206.

[29] Monz C, Weerkamp W. A Comparison of Retrieval-based Hierarchical Clustering Approaches to Person Name Disambiguation[C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2009: 650-651.

[30] Ferreira A A, Veloso A, Gonçalves M A, et al. Effective Self-training Author Name Disambiguation in Scholarly Digital Libraries[C]//Proceedings of the 10th Annual Joint Conference on Digital Libraries. ACM, 2010: 39-48.

[31] Tang J T, Lu Q, Wang T, et al. A Bipartite Graph Based Social Network Splicing Method for Person Name Disambiguation [C]//Proceedings of the 34th International ACM SIGIR Conference On Research and Development in Information Retrieval, 2011: 1233-1234.

[32] 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，2009，32（7）：1365-1374.

[33] Liu T Y. Learning to Rank for Information Retrieval [J]. Foundations and Trends in Information Retrieval, 2009, 3(3): 225-331.

[34] Li H. Learning to Rank for Information Retrieval and Natural Language

Processing[J]. Synthesis Lectures on Human Language Technologies, 2011, 4(1): 1-113.

[35] Geng X B, Liu T Y, Qin T, et al. Query Dependent Ranking Using K-nearest Neighbor[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008: 115-122.

[36] Zhu Z A, Chen W, Wan T, et al. To Divide and Conquer Search Ranking by Learning Query Difficulty[C]//Proceeding of the 18th ACM Conference on Information and Knowledge Management, 2009: 1883-1886.

[37] Zhu X, Lu Z. A Novel Ontology-based Approach of Chinese Person Name Disambiguation[C]//Semantics, Knowledge and Grids (SKG), 2012 Eighth International Conference on. IEEE, 2012: 197-200.

[38] Han X P, Zhao J. Web Personal Name Disambiguation Based on Reference Entity Tables Mined From the Web[C]//Proceeding of the Eleventh International Workshop on Web Information and Data Management, 2009: 75-82.

[39] Chen K, Lu Z, Yin X, et al. Finding Web Appearances of Social Network Users via Latent Factor Model[C]//Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2012: 1077-1078.

[40] Han X, Zhao J. Structural Semantic Relatedness: A Knowledge-based Method to Named Entity Disambiguation[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 50-59.

[41] Vilain M, Burger J, Aberdeen J, et al. A Model-theoretic Coreference Scoring Scheme[C]//Proceedings of the 6th Conference on Message Understanding. Association for Computational Linguistics, 1995: 45-52.

[42] Bagga A, Baldwin B. Algorithms for Scoring Coreference Chains[C]//The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference. 1998, 1: 563-6.

[43] Luo X. On Coreference Resolution Performance Metrics[C]//Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005: 25-32.

[44] Chen C, Ng V. Combining the Best of Two Worlds: A Hybrid Approach to Multilingual Coreference Resolution[C]//Proceedings of the Joint Conference on EMNLP and CoNLL-Shared Task. Association for

Computational Linguistics, 2012: 56-63.

[45] Yuan B, Chen Q, Xiang Y, et al. A Mixed Deterministic Model for Coreference Resolution[C]//Proceedings of the Joint Conference on EMNLP and CoNLL-Shared Task. Association for Computational Linguistics, 2012: 76-82.

[46] He Z Y, Wang H F, Li S J. The Task 2 of CIPS-SIGHAN 2012 Named Entity Recognition and Disambiguation in Chinese Bakeoff. Proceeding of the Second CIPS-SIGHAN Joint Conference on Chinese Language, 2012: 108-114.

[47] Tian W, Pan X, Yu Z T, et al. Chinese Name Disambiguation Based on Adaptive Clustering with the Attribute Features. Proceeding of the Second CIPS-SIGHAN Joint Conference on Chinese Language, 2012: 132-137.

[48] Peng Z H, Sun L, Han X P. SIR-NERD: A Chinese Named Entity Recognition and Disambiguation System using a Two-Stage Method. Proceeding of the Second CIPS-SIGHAN Joint Conference on Chinese Language, 2012: 115-120.

[49] Hao Z, Wong D F, Chao L S. A Template Based Hybrid Model for Chinese Personal Name Disambiguation. Proceeding of the Second CIPS-SIGHAN Joint Conference on Chinese Language, 2012: 121-126.

[50] Wang L Y, Li S, Wong D F, et al. A Joint Chinese Named Entity Recognition and Disambiguation System. Proceeding of the Second CIPS-SIGHAN Joint Conference on Chinese Language, 2012: 146-151.

[51] Fan Q H, Zan H Y, Chai Y M, et al. Chinese Personal Name Disambiguation Based on Vector Space Model. Proceeding of the Second CIPS-SIGHAN Joint Conference on Chinese Language, 2012: 152-158.

附录一

本文中用到的词性标注标记的含义如下表所示。

编号	词性标记	含义
1	AD	专有名词
2	AS	动词“是”
3	BA	“把”
4	CC	口头名词
5	CD	许多、若干、,个把
6	CS	从属连接词
7	DEC	从句“的”
8	DEG	修饰“的”
9	DER	“得”
10	DEV	“地”
11	DT	限定词：“各”、“全”、“某”、“这”
12	ETC	“等”、“等等”
13	FW	外来词
14	IJ	感叹词
15	JJ	名词修饰语
16	LB	“被”、“给”
17	LC	方位词
18	M	量词
19	MSP	其他小品词：“所”
20	NN	口头名词
21	NR	专有名词
22	NT	时间名词
23	OD	序数
24	ON	拟声法
25	P	介词：“对”、“由于”、“因为”(除了 “把”和“被”)
26	PN	代词
27	PU	标定符号
28	SB	“被”、“给”
29	SP	句尾语气词
30	VA	表语形容词
31	VC	“是”

(续表)

编号	词性标记	含义
32	VE	表示存在的词：“有”、“有”、“无”、“没”
33	VV	情态动词、动词、“拥有”、“富有”、“具有”

攻读硕士学位期间发表的论文及其它成果

（一）发表的学术论文

[1] 徐睿峰，**刘杰**，桂林，徐建，陆勤.结合百科知识和互联网验证的中文人名消歧[J].(CCIR2013 会议已推荐《计算机学报》发表，学生第一作者)

[2] Ruifeng Xu, Jun Xu, **Jie Liu**, Chengxiang Liu, Chengtian Zou, Lin Gui, Yanzhen Zheng and Peng Qu. Incorporating Rule-based and Statistic-based Techniques for Coreference Resolution[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP-CoNLL2012), 2012: 107-112 (学生第一作者)

[3] **Jie Liu**, Ruifeng Xu, Qin Lu, Jian Xu. Explore Chinese Encyclopedic Knowledge to Disambiguate Person Names [C]//Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, 2012: 138–145 (第一作者)

[4] Jian Xu, Qin Lu, **Jie Liu**, Ruifeng Xu. 2012. NLP-Comp in TAC 2012 Entity Linking and Slot-Filling[C]//Proceedings of the Fourth Text Analysis Conference, 2012.

[5] Chengxiang Liu, Ruifeng Xu, **Jie Liu** et al. Comparative Opinion Sentences Identification and Elements Extraction[C]//Proceedings of the 2013 International Conference on Machine Learning and Cybernetics, 2013: 1886-1891.(EI Indexed)

[6] 王贺，刘呈祥，郑燕珍，**刘杰**，屈鹏，邹承天，徐睿峰. 否定句和比较句的情感倾向性分析[C].第四届中文倾向性分析评测会议, 2013.

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《结合共指消解的跨文档中文人名消歧研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：刘杰　　日期：2013年12月30日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：刘杰　　日期：2013年12月30日

导师签名：徐睿峰　日期：2013年12月30日

致　谢

本论文是在徐睿峰老师的耐心指导下完成的，是两年半的研究生工作的总结。在此衷心地感谢徐老师两年多以来在学习和生活上的教导和帮助，这些教导和教诲将会使我受益一生。在研究上，徐老师将我领入自然语言处理这个诱人的研究方向，教导我如何开展研究；在生活上，徐老师的关心和照顾都给我的研究生生活带来了很多方便和乐趣。感谢陆勤老师和徐军老师对我课题的指导，感谢陈建铭老师和林浚玮老师为我的论文提供了宝贵的建议。也感谢实验室其他老师们，特别是王晓龙老师为实验室营造了好的环境。

感谢徐建师兄和桂林师兄在课题研究中的帮助和建议。感谢实验室众多的兄弟姐妹们，和你们一起的学习和娱乐经历给我的研究生生活添加了很多难忘的时光。特别感谢张玥师兄和叶璐师姐对我找工作的帮助和指导。最后感谢室友周坤龙和肖坚同学，你俩对我生活上提供了很多的帮助。

最后感谢我的父母，是他们的支持让我一步步前进，是他们的鼓励给了我跨过困难，继续前进的动力。

阅读此文的还阅读了：

1. 中文指代消解名词短语的识别
2. 中文人名消歧算法研究
3. 基于属性信息的中文人名消歧研究
4. 结合Doc2Vec与改进聚类算法的中文单文档自动摘要方法研究
5. 基于句义结构分析的中文人名消歧
6. 中文人名跨文档指代消解研究
7. 基于层次聚类算法的中文人名消歧
8. 面向网络人物搜索的中文人名消歧
9. 基于规则的维吾尔人名智能消歧研究
10. 统计与规则相结合的中文指代消解
11. 中文跨文本人名同名同指消解研究
12. 共指消解研究方法综述
13. 语义Web中对象共指的消解研究
14. 结合共指消解的跨文档中文人名消歧研究
15. 中文词义消歧的方法研究
16. 基于融合特征的中文图书作者人名消歧方法研究
17. 利用优化的DBSCAN算法进行文献著者人名消歧
18. 基于人名消歧的自引统计研究
19. 浅谈高校课题研究中文档材料的运用
20. 面向体育新闻领域的中文简单名词短语共指消解
21. 中文跨文本人名同名同指消解研究
22. 基于《知网》的中文信息结构消歧研究
23. 基于ART网络的无指导中文共指消解方法
24. 一种基于谱聚类的共指消解方法
25. 中文跨文档指代消解的研究与实现

26. 面向产品评论的共指消解方法研究与实现

27. 面向人名消歧任务的人名识别系统

28. Web人名消歧方法的研究与实现

29. 基于关键证据与E2LSH的增量式人名聚类消歧方法

30. 异构Web对象共指冲突识别与消解研究

31. 基于层次聚类的跨文本中文人名消歧研究

32. 基于聚类集成的人名消歧算法

33. 中文个人名称规范文档的关联数据化研究

34. WPSOffice2009跨文档操作技巧

35. 基于决策树的中文指代消解

36. 中文系女生人名研究——以华中师大中文系2009、2010、2011级学生人名为例

37. 中文跨文本人名同名同指消解研究

38. 中文个人名称规范文档的关联数据化研究

39. 基于共指消解的实体搜索模型研究

40. 基于分步聚类的人名消歧算法

41. 词义消歧技术研究

42. 基于分类信心重排序的中文共指消解研究

43. 电子物证检验中文档的溯源

44. 基于有监督关联聚类的中文共指消解

45. 基于最大熵模型的共指消解研究

46. 一种基于特征映射的中文专家消歧方法

47. 基于查询扩展的人名消歧

48. 基于特征分选策略的中文共指消解方法

49. Vista中文档字体异常

50. WPS2009跨文档操作效率高