基于单文本指代消解的人物家庭网络构建研究*

顾静航[1,2]，朱苏阳[1,2]，钱龙华[1,2]，朱巧明[1,2]

（1. 苏州大学 自然语言处理实验室，江苏 苏州 215006；

2. 苏州大学 计算机科学与技术学院，江苏 苏州，215006）

摘要：人物家庭网络是社会关系网络中的一个重要组成部分，因此，如何高效准确地提取出人物的家庭网络具有重要研究意义。本文在前人工作的基础上提出一种基于单文本指代消解技术的人物家庭关系抽取方法，以此扩大人物家庭关系抽取的范围，进而提高人物家庭网络的召回性能。本文还提出了一种基于人物虚拟边的家庭网络评估指标,用于更合理地评价构建出的人物家庭网络的性能。在大规模中文语料 Gigaword 上的实验表明，本方法可以较为准确地抽取出人物的家庭关系，进而提高人物家庭网络的召回性能，从而为社会网络分析提供基础数据。

关键词：社会关系网络；家庭网络；单文本指代消解

中图分类号：TP391　　**文献标识码：**A

Research on Building Family Networks Based on Within-Document Coreference Resolution

GU Jinghang[1,2], ZHU Suyang[1,2], QIAN Longhua[1,2], ZHU Qiaoming[1,2]

(1. Natural Language Processing Lab, Soochow University, Suzhou, Jiangsu, 215006;

2. School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu, 215006)

Abstract: Personal Family Network is an important component of social networks, therefore, it is of great importance of how to extract personal family relationships. On the basis of previous research, we propose a novel method to extend the family relation extraction via Within-Document Coreference Resolution, improving the recall of family networks constructed. Meanwhile, a new evaluation metric is devised to evaluate the performance of personal family networks more reasonably. The experimental results on a large-scale corpus of Gigaword show that, our method can extract accurate family relations while increase the recall of family networks, thereby laying the foundation for social network analysis.

Key words: Social Network; Family Network; Within-Document Coreference Resolution

1 引言

人物的社会关系网络在当今信息化社会中具有重要的作用，对人物社会关系网络的分析和应用不仅可以提高人们的生活质量和生活效率，还可以衍生出巨大的商机。早期的社会网络构建主要依赖于大规模网页中的人名共现现象(Referral Web/Flink)[1, 2]进行构建，而并未深究其具体的关系类型；近年来的研究逐渐转向采用机器学习的方法，挖掘特定领域中的社会关系网络，如学术社会网络(ArnetMiner)[3]、文学作品中的社会网络[4, 5]和人物传记中的社会网络[6]等。

* **收稿日期：** 2014-6-14　**定稿日期：** 2014-7-27

基金资助：国家自然科学基金(61373096, 90920004)；江苏省高校自然科学重大项目(11KJA520003)。

作者简介：顾静航（1987——），男，硕士研究生，研究方向为自然语言处理，Email: gujinghang59420@sina.com；朱苏阳（1989——），男，硕士研究生，研究方向为信息抽取，Email: 20124227049@suda.edu.cn；钱龙华（1966——），通信作者，男，副教授，硕士生导师，研究方向为自然语言处理，Email: qianlonghua@suda.edu.cn；朱巧明（1963——），男，教授，博士生导师，研究方向为自然语言处理，Email: qmzhu@suda.edu.cn。

众所周知，家庭是人类社会最基本的组成单位，因而人物的家庭关系网络理应是社会关系网络中的核心部分。传统的社会关系网络分析(Social Network Analysis, SNA)往往着眼于以人作为个体，考察个体在网络中的作用，而忽略了家庭在社会网络中的核心地位，忽视了家庭作为一个整体对社会网络的影响；此外，其对社会网络中的人名歧义问题处理也比较简单，效果不太理想。针对以上不足，Gu 等(2013)[7]则对人物的家庭网络构建进行了研究，并通过相应的跨文本指代消解技术对人物的重名与多名问题进行了处理，虽然其构建出了较为准确的人物家庭，但却存在着家庭召回性能相对较低等问题。

本文的思想则是在 Gu 等(2013)[7]工作的基础上，融入单文本指代消解技术，提升人物家庭关系抽取的性能，在构建人物家庭网络的同时，进一步提高家庭网络的召回数量。本文在评估人物家庭网络的性能时，提出一种基于人物虚拟边的图检索 (Graph Retrieval Over Virtual Edges, GROVE)评价方法，用于更合理地评价人物家庭网络的性能。实验结果表明，本文提出的方法能够很好地提高从大规模的文本语料库中抽取出的人物家庭关系实例的数量，进一步提升人物家庭网络的召回性能。

本文分为以下几个部分：第 2 节介绍相关的研究工作；第 3 节提出基于单文本指代消解的人物家庭网络构建方法；第 4 节描述系统性能的评价方法；第 5 节给出实验结果和分析；第 6 节为总结和展望。

2 相关工作

社会关系网络构造的一个首要任务是人物关系挖掘，它是命名实体间语义关系抽取的一个特例，其任务是从自然文本中提取出人物之间所存在的语义关系。关系抽取研究大都采用基于机器学习的方法，根据其对标注语料库数量的需求，可以分为指导性学习[8]、弱指导学习[9]和无指导学习[10]等，语料标注的数量和质量通常决定了抽取性能的好坏。弱指导学习方法，由于其仅需要极少量的人工干预就可以自动地挖掘出大量的关系实例，从而避免了语料库的标注问题，因而被广泛地应用在多种关系抽取任务中。

在人物关系网络构建方面，早期的研究大都利用基于网页人名共现的方法，如 Kautz 等(1997)[1]提出的基于 Web 的社会网络系统 Referral Web 以及 Mika 等(2005)[2]开发的 Flink 系统等，它们都利用人名的共现次数来实现社会关系网络的挖掘。近些年的研究逐渐采用机器学习的方法，旨在挖掘更为丰富的人物社会关系。Tang 等(2008)[3]提出了 ArnetMiner 系统，它利用 SVM 和 CRF 等分类模型构建学术人物之间的关系网络。Zhu 等(2009)[11]所提出的 StatSnowball 系统，它通过自举学习进行人物社会关系的抽取，继而使用概率模型和马尔科夫逻辑网络等方法，在开放的 Web 环境下构建人物关系网络。Peng 等(2012)[12]采用基于树核函数的方法挖掘人物社会关系，并将其扩展为静态关系（如家庭关系和商业使用关系）和动态关系（如人物交互关系）。Elson 等(2010)[4]、Agarwal 等(2012)[5]和 Agarwal 等(2013)[13]对小说人物的社会关系网络进行了研究，提出了隐式社会关系的概念，即共同参与某一社会事件（如互动和观察等）的角色之间所存在的社会关系。Camp 和 Bosch(2011)[6]则从人物传记中提取带有情感极性的人物社会关系，利用 SVM 分类模型构建社会关系网络。

值得一提的是 Gu 等(2013)[7]的工作，他们以人物的家庭关系为核心，采用自举学习的方法对人物的家庭关系进行抽取，在对人物进行家庭网络融合时，采用一种简单且有效的跨文本指代消解方法解决人物的重名与多名问题，并构建出质量较为可靠的人物家庭网络。

3 基于单文本指代消解的人物家庭网络构建

本文以 Gu 等(2013)[7]的系统作为原型，首先，采用与之相同的自举学习方法进行人物

家庭关系抽取并习得相应的家庭关系模式；其次，在获得相应的家庭关系模式后，本文对这些模式进行泛化，使用新的模式对文本进行匹配，针对匹配到的句子本文提出一种基于单文本指代消解的人物家庭关系抽取方法，以拓展人物家庭关系的抽取范围；最后，本文对最终获得的人物家庭关系采用与 Gu 等(2013)[7]相同的跨文本指代消解方法进行家庭网络融合，构建出人物的家庭网络。具体的人物家庭网络构建流程如图 1 所示。

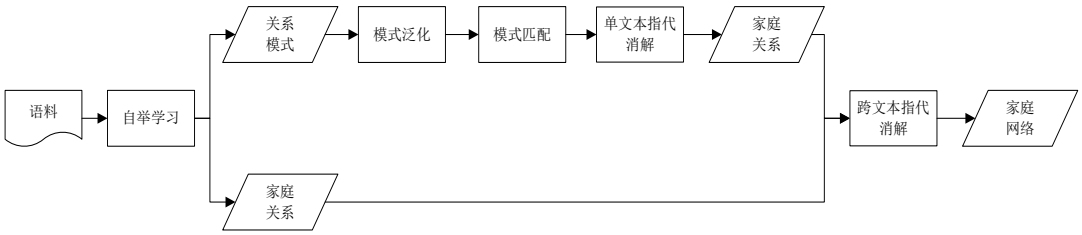

图 1 人物家庭网络构建流程

3.1 关系模式的泛化

自举学习方法可以自动地从文本中抽取出人物的家庭关系，但文本中仍然会存在有大量由自举学习方法无法捕获的人物家庭关系，考虑以下例子：

a) “*记者在女团比赛结束后拨通了**刘璇**家中的电话,她的父亲**刘锦成**在电话中连声说:“太意外了!””*

b) “*故事的主人公是华北油田勘探二公司 6022 钻井队劳动模范**袁永德**和他的妻子**凌金艳**。”*

其中，上述两个句子分别包含了一对父女“*刘璇 刘锦成*”和一对夫妻“*袁永德 凌金艳*”，且 a 句中包含了模式关键词“*的父亲*”，b 句中包含了模式关键词“*的妻子*”。然而，由于模式关键词的左右两侧并不都是人名，因而基于自举学习的方法无法精确匹配到上述例句，也就无法抽取出其中的人物关系实例。

人们在现实的语言表述中往往会应用代词来指代上文中出现过的人物，以使表达显得简洁且连贯。本文在自举学习的基础上，使用人称代词替代已习得关系模式中的位于关键词之前的人物，达到对模式进行泛化的目的。如模式“*<Child>的父亲<Parent>*”，用人称代词“*她*”进行泛化后可以得到新的模式“*她的父亲<Parent>*”。用泛化后的模式重新检索语料库，并对结果进行分词、词性标注和命名实体识别。在对模式进行泛化时，本文使用了常见的单数人称代词，包括“我、你、您、他、她、其、自己”等。

在模式泛化的基础上，本文提出一种基于单文本指代消解的人物家庭关系抽取方法以扩大关系实例的抽取规模，进而通过跨文本指代消解技术对抽取出的人物家庭关系实例进行家庭网络融合，从而构建出人物的家庭网络。

3.2 中心理论基础知识

代词的引入，可以良好的表达句意，使语句更为连贯，因而，对人称代词的消解，需要联系上下文环境，同时要兼顾语言的连贯性等问题。中心理论[14]正是基于语篇连贯性的特点，以中心焦点的方式阐述了在英语语篇中代词的分布规律及其实现所需的各种条件，它认为语段中出现的话语实体是语篇的中心，而这些中心在上下文中的突显程度以及实现它们的语言形式都会对整个语篇的连贯性产生影响。中心理论认为每一语段都应该包含有以下三种不同类型的中心：

(1) 前向中心(Forward-Looking Center, Cf)：是指语段中可能存在的会话焦点。它是与下文发生联系的枢纽，其可能包含一系列对象，这些对象以其突显度的强弱排列。

(2) 后向中心(Back-Looking Center, Cb)：是指语段当前的会话焦点。它只应包含一个对

象，该对象起到与先前语段相关联的作用。中心理论认为前一语段 Cf 的集合中，突显度最高的对象应为本句的 Cb。

(3) 优选中心(Preferred Center, Cp)：中心理论认为在前向中心里突显度最高的那个对象应该作为优选中心。

在中心理论模型中，根据前后两个语段(分别设为 U_{n-1} 和 U_n)的三种中心的变化，可以定义以下几种过渡类型，具体如表 1 所示。

表 1 中心理论中语段间的跳转类型

	$Cb(U_n)=Cb(U_{n-1})$或者无 $Cb(U_{n-1})$	$Cb(U_n)\neq Cb(U_{n-1})$
$Cb(U_n)=Cp(U_n)$	延续(Continue)	顺转(Smooth Shift)
$Cb(U_n)\neq Cp(U_n)$	保持(Retain)	硬转(Rough Shift)

其中，上述四种跳转类型分别表示了语段间不同的连贯程度，其连贯性由高到低依次是：Continue > Retain > Smooth Shift > Rough Shift。

3.3 中心理论在中文指代消解中的应用

(一)后向中心的选择

本文以中心理论的基本原则作为基础，结合中文自身的表述特点，对中心理论进行了一些改进和简化，对其在中文指代消解任务中的应用给出了如下的判断规则：

(1) 一个句子中，如果人名先于代词出现，那么该句的 Cb 应是本句 Cf 中突显度最高的人名；

(2) 一个句子中，如果某个人名先于代词出现，那么该句中的待消解代词应指向本句内的某个人物，即其符合句内指代消解情况；

(3) 一个句子中，如果代词先于人名出现，那么该句的 Cb 应是前一句 Cf 中突显度最高的人名；

(4) 一个句子中，如果某个代词先于人名出现，那么该句中的待消解代词应指向前一句中的某个人物，即其符合句间指代消解情况；

(5) 一个句子中，与待消解代词具有相同“表述形式”的代词，认为其与待消解代词指向同一人物，它们可以形成一条代词链；

(6) 一个句子中，如果既没有人名也没有代词，则该句的 Cb 应与前一句的 Cb 保持一致。

(二)人名突显度的判断

确定不同人名在 Cf 集合中的突显度对于 Cb 的选择有着重要意义。在本文的实验中，Cf 中各对象的突显度按照中文语法角色排列的顺序为：主语 > 宾语 > 其他。

本文在确定人名突显度时，结合了结构句法分析和依存句法分析结果，对句子中的主宾语成分进行判断。本文使用的句法分析器为 Stanford Parser。在进行依存句法分析时，句法分析工具定义了多种依存关系类型，如“nsubj”代表名词性主语，“dobj”代表直接宾语等，通过依存关系标签可以确定句子中的主语核心词汇以及宾语核心词汇，再依靠结构句法识别出核心词汇所在的名词短语。与主语核心词汇一同出现在名词短语中的人名即可认为是充当句子的主语成分；与宾语核心词汇一同出现在名词短语中的人名即可认为是充当句子的宾语成分；此外，则认为人名在句中充当其他成分。

图 1 和图 2 分别给出了例句“*江泽民总书记会见李政道夫妇*”的结构句法分析和依存句法分析结果。通过依存句法分析可以发现主语核心词汇是“*总书记*”，宾语核心词汇是“*夫妇*”。通过结构句法分析可以发现句子中存在包含主宾语核心词汇的两个名词性短语“*江泽民总书记*”和“*李政道夫妇*”，因而人名“*江泽民*”在句中充当主语成分，人名“*李政道*”在句

子充当宾语成分，人名突显度“*江泽民* > *李政道*”。通过结构语法判断句子内名词短语时，识别出句中最长的名词短语即可，即识别结构句法树中顶层的“*NP*”标签即可。处于同一成分内的多个人名，按其到依存根结点“*ROOT*”的词汇距离进行排序，距离越近，突显度越高。

```
(ROOT
    (IP
      (NP
        (NP (NR 江泽民))
        (NP (NN 总书记)))
      (VP (VV 会见)
        (NP
          (NP (NR 李政道))
          (NP (NN 夫妇))))))
```

图 1 结构句法分析

```
nn(总书记-2, 江泽民-1)
nsubj(会见-3, 总书记-2)
root(ROOT-0, 会见-3)
nn(夫妇-5, 李政道-4)
dobj(会见-3, 夫妇-5)
```

图 2 依存句法分析

3.4 特征选择

特征向量的选择对机器学习方法有着很重要的影响，好的特征向量应该不受限于某个固定的领域，而具有一定的通用性，且能如实地反映语言现象中的直观特征。本文参照前人对中英文指代消解的研究成果，结合自身对中心理论的理解，给出以下特征：

(1) Distance：候选人名到代词链中首个出现的代词之间的实体距离。如果人名是距离代词链中首个代词最近的人名，则距离取 1；若相差一个人名则距离取 2，以此类推；如果人名的位置在代词链中第一个代词之后，距离可以取负值。

(2) NameIsFocus：人名是否是其所在句子的 Cb，若是则取 1，否则取 0。

(3) AnyOtherFocus：人名与待消解代词之间是否含有其他 Cb，若有则取 1，否则为 0。

(4) InsideSentence：待消解代词适合何种消解方式，句内消解为 0，句间消解为 1。

(5) NameInPunc：人名是否出现在括号、书名号等特殊符号内，若是则取 1，否则取 0。

4 系统性能评价方法

本文在家庭网络构建中主要涉及单文本指代消解、人物家庭关系抽取以及家庭网络融合等过程，因此，为了更为全面地衡量系统性能，需要给出不同阶段的系统性能。

4.1 单文本指代消解性能评价方法

本文对单文本指代消解性能的评价采用了较为通用的准确率、召回率和 F1 指数作为评价指标。

4.2 人物关系抽取性能评价方法

由于在大规模语料中进行关系抽取时难以考察召回率，因而本文关注的重点是人物关系抽取的准确性，故选取准确率作为评价指标。

4.3 家庭网络性能评价方法

对于家庭网络性能的评估，目前还没有一个标准的方法。Gu 等(2013)[7]所采用的评价方法只考虑了家庭构成的总体情况，即要求所发现家庭必须在人物数量、人物指代链、人物间的关系类型完全正确时才可认为其是一个正确家庭，然而这样的评价方法并未考虑家庭内部的构成情况，大量挖掘出的家庭都存在着部分人物关系正确的情况。

为了更好的衡量所发现家庭内部的人物关系情况，本文提出一种基于家庭内部人物间虚拟边的图检索(Graph Retrieval Over Virtual Edges, GROVE)评价方法。

所谓虚拟边是指一个家庭内部具有直接关系的人物实体之间，通过使用各自指代链内的不同名称，进行关系组合后形成的边，如以下例子：

a) “*尼日利亚国家元首**阿巴查**的夫人**玛丽亚姆·阿巴查**26 日在接受本社记者……*”

b) “*尼日利亚国家元首**阿巴查**和夫人**玛利亚姆·阿巴查**、外交部长……*”

上述例句源自不同的文档，但人物来自于同一个家庭。该家庭内部的人物“*阿巴查*”和“*玛丽亚姆·阿巴查*”具有夫妻关系，同时妻子还具有别名“*玛利亚姆·阿巴查*”。这对夫妻通过别名可以形成 2 条虚拟边，包括“*阿巴查--玛丽亚姆·阿巴查*”以及“*阿巴查--玛利亚姆·阿巴查*”。

虚拟边的提出，使得在性能评价时，可以更好的兼顾家庭内部人物关系的情况。准确率和召回率的计算公式如下所示：

$$Precsion_i = \frac{\left|Right\left(Response(i)\right)\right|}{\left|Response(i)\right|} \quad \text{公式(5.1)}$$

$$Recall_i = \frac{\left|Right\left(Response(i)\right)\right|}{\left|Key(i)\right|} \quad \text{公式(5.2)}$$

其中，*Response(i)*是指机器识别的第 i 个家庭中全部的虚拟边个数，*Right(Response(i))*是指机器识别的家庭中正确的虚拟边个数，*Key(i)*指与机器识别家庭相对应的标准家庭中全部的虚拟边个数。系统整体性能的计算公式为：

$$Precsion = \sum_{i=1}^{N} w_i \cdot Precsion_i \quad \text{公式(5.3)}$$

$$Recall = \sum_{i=1}^{N} w_i \cdot Recall_i \quad \text{公式(5.4)}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad \text{公式(5.5)}$$

其中，w_i 为标准家庭在标准集中的权重，是标准家庭中的虚拟边数与标准集中虚拟边总数的比值，计算方法如下：

$$w_i = \frac{\left|Key(i)\right|}{\sum_{i=1}^{N}\left|Key(i)\right|} \quad \text{公式(5.6)}$$

5 实验

本节根据前文的描述，讨论了相关实验结果，依次分析了单文本指代消解的性能、人物关系抽取的性能以及人物家庭网络融合的性能。

5.1 实验语料及预处理

本文采用 Gigaword 中文语料库作为家庭网络构建的实验数据。该语料来源于新闻，共有 1,033,679 篇新闻报道，包含新华社新闻和早报新闻。

(一)单文本指代消解训练实例的生成

本文所采用的训练数据来源于 ACE2005 中文语料。ACE 语料库中标注了实体信息、关系信息和事件信息等，其中，实体信息的标注相当于指代关系，实体链也即是相应的指代链。本文在生成训练实例时，遵循以下几条原则：

(1) 只选用那些包含有单数人称代词的实体链；

(2) 在代词消解范围的选择上，本文的选取策略基于语言表述中的一个事实：先行语与

指示语的距离往往不会很远，如距离过远则会引起阅读困难。因而针对某一个代词的消解，本文取其上文中的 2 句与其所在句，共 3 句作为该代词的消解范围；

(3) 在生成训练实例的过程中，人称代词作为照应语，人名作为先行语，在消解范围，照应语之前的所有人名都是其潜在的先行语。照应语和距离最近且与其处于同一条实体链中的先行语组成正例，消解范围内的其他先行语则和照应语构成负例。

本文的实验共使用了 1014 个代词进行训练实例的生成，涵盖了 ACE2005 中的 276 篇文章，共产生 2643 对“*先行语——照应语*”指代对，其中正例 1014 对，负例 1629 对。

(二)单文本指代消解测试实例的生成

测试数据来源于使用泛化模式对 Gigaword 语料库进行匹配后得到的文本，针对人称代词而言，有效的候选先行语同样在 3 句范围以内。

测试语料由于没有实体链的标注信息，因而需要进行人工标注。标注时同样选取与代词指向同一实体且距离代词最近的人名组合成正例，其余为负例。在通过泛化模式进行匹配而得到的句子中，与前文具有现实指代关系的人称代词有 1258 个，在消解范围内将这些代词与人名组成指代对，共有 3560 对，其中正例 1258 对，负例 2302 对。

5.2 单文本指代消解性能

(一)单文本指代消解总体性能

由于本文的指代消解过程是针对泛化模式所匹配到的文本中的单数人称代词，故在以往指代消解任务中较为通用的特征，如同位语特征、先行语类别特征、照应语类别特征以及单复数一致性等特征并不能满足本文指代消解任务的需求。因而，本文选用基于就近指代原则进行指代消解的方法作为基准系统。就近指代原则是指，在进行指代消解时选取位于代词前且距离代词最近的人名作为消解结果。本文的实验使用 SVM-light 分类器，实验中采用了径向基函数作为分类器的核函数。实验结果如表 2 所示。

表 2 单文本指代消解性能统计

指代分类	代词分布	基准系统			中心理论		
		P (%)	R (%)	F1(%)	P (%)	R (%)	F1(%)
总体性能	1258	69.3	72.6	70.9	78.4	75.5	77.0
句内消解	712	79.6	83.4	81.5	83.2	83.4	83.3
句间消解	546	55.9	58.4	57.1	71.6	65.2	68.3

从表中可以发现，应用了中心理论后，性能有了很好的提升。这主要是由于中心理论考虑了语篇的上下文连贯性，可以较好地解释句间指代现象，因而对需要进行句间消解的代词具有良好的效果。从表中还可以看出：

- 针对人名与人称代词之间的指代消解问题，仅依靠就近指代原则已可以较好的解决大部分情况，说明人们在语言表达中，往往会使用代词去替代上文中刚刚出现的人名，这么做是为了不影响他人对句子的理解，同时也使句子更为简洁、连贯。
- 需要进行句内指代消解的代词和需要进行句间指代消解的代词在总体数量上的差别并不显著，而就近指代原则可以较好的解决有关句内指代消解的问题，但在面对句间指代消解问题时则略有不足。

本文对系统错误的指代消解情况进行了探究，分析主要原因有以下几种：

- 分词错误(约占错误比例的 65%)。分词错误主要指人名的切词错误以及词性标识错误，这是本系统错误中的主要原因。本文使用的分词工具为 ICTCLAS。使用自动分词工具会产生将人名切分错误或将非人名词语识别成人名等情况，从而影响系统性能。如例句：“*二 连 文书 蔡 报 罕 是 九江市 江 洲 镇 人 ,[他] 父亲 蔡*

灿光 得知 儿子 的 部队 也 来 家乡 抢险，跟着 运送 石料 的 船 来到 大坝，想 看 一 看 两年 没 见面 的 儿子 。”该例句中代词的先行语应为“*蔡报宇*”，但由于分词错误无法识别出人名。

- 句式表达较为复杂(约占错误比例的 21%)，本文的模型不足以覆盖。文本中会出现某些表达较为复杂的句子，如例句“*李鹏 请 韩升洲 转达 [他] 和 夫人 朱琳 对 金泳 三 总统 和 夫人 的 亲切 问候 和 谢意 。*”对于上述情况，本文的方法会倾向于选择最近的人名作为消解结果，因此，并不适合此类复杂的指代现象；
- 文本中出现普通名词表示人物的情况(约占错误比例的 8%)。文本中存在用普通名词来表述人物的情况，如例句“*据 报道 ， 男童 遇害 后 ， [他] 的 母亲 孙碧音 哀伤 地 接受 媒体 的 访问 ， 透露 死者 生前 懂事 乖巧 。*”该例句中代词的先行语应指向“*男童*”，但由于本文的指代消解过程只关注人名与代词之间的消解问题，并不适用针对普通名词的指代消解问题；
- 指代范围过小(约占错误比例的 6%)。本文在进行指代消解时，将先行语的范围限定在 3 句以内，但消解过程中会存在代词真正的先行语出现在消解范围之外的情况，针对这种情况，本文的方法并不能有效的进行指代消解。

(二)不同特征对性能的影响

本节采用特征分离方式来考察不同特征对系统性能贡献度的影响。所谓分离方式，就是在使用特征时，每次将一个特征分离出特征集而不予采用，以考察这个特征对系统性能的影响。本文将 Distance、NameIsFocus、AnyOtherFocus、InsideSentence、NameInPunc 等特征依次从系统的特征集合中分离出去，结果如表 3 所示。

表 3 特征分离实验

特征集	P (%)	R (%)	F1(%)
所有特征	78.4	75.5	77.0
-Distance	69.9	29.6	41.6
-NameIsFocus	72.2	71.9	72.1
-AnyOtherFocus	80.5	70.0	74.9
-InsideSentence	83.1	63.2	71.8
-NameInPunc	75.5	75.7	75.6

其中，“-”表示在使用所有特征的基础上分离该特征，从表可以看出：

- 距离特征的影响最为显著。将距离特征分离出去后，可以发现，指代消解的性能下降的最为明显，这是由于首先句内消解的情况比句间消解要多，同时这也印证了语言表述中，为了使表述简洁明了，代词往往是指代距离最近的人名；
- 代词信息对于指代的消解具有重要意义。当拥有明确的代词信息后，可以对指代消解适用句内情况还是句间情况给出很好地指导。中心理论的主要思想便是解决连贯语篇中的指代消解问题，因而代词适用的消解类型对指代消解十分重要；
- 在中心理论中，先行语的信息对指代消解有积极意义。在进行句间指代消解时，代词需要指向上文中的焦点人物，因而对人名是否是焦点的判断尤为重要。
- 在语言表述中，存在少数情况，人名会出现在如括号、书名号等特殊符号内的情况，此种情形下的人名往往不适合充当代词的先行语。

5.3 人物家庭关系抽取性能

本文的人物家庭关系抽取分别经历了基于自举学习的关系抽取阶段，以及在此基础上通过泛化模式后而引入的基于单文本指代消解的关系抽取阶段。在经过这两个阶段后，需要对

抽取出的人物实例对进行准确率的评价。本文采用抽样统计的方法，分别对人物关系对无放回取样 4 次，样本大小为 100，然后人工评判其正确性，关系抽取的性能如表 4 所示。

表 4 人物家庭关系抽取性能

关系抽取方法	发现的人物对数目	P(%)
自举学习	2167	91.5
单文本指代消解	1062	81.7
自举学习+单文本指代消解	2861	87.3

从表中可以看出，基于自举学习的方法可以抽取出大量的人物关系实例；单文本指代消解所抽取出的关系实例的准确度相对不足，但在引入单文本指代消解后，可以有效地挖掘出新的人物关系实例，从而提高关系抽取的召回性能。

两个阶段的关系抽取过程所发现的人物家庭关系存在部分重叠，需要对所获得的结果进行合并。在合并关系抽取结果时，如果两对人物对中的人名一致，且关系类型也相同，则可以直接合并，认为它们表示同一对人物关系，合并后共得到 2861 对人物对。

5.4 家庭网络融合性能

对构建出的人物家庭网络的性能进行评价时，需要对文本内的人物家庭进行人工标注。本文采用与 Gu 等(2013)[7]相同的方法对人物的家庭网络进行人工标注，得到的人工标注结果如表 5 所示。

表 5 人工标注结果

家庭总数	家庭中人物总数
192	660

表 6 反映出单文本指代消解后，通过对人物家庭关系抽取的拓展，在引入大量新的人物关系实例对的情况下，系统对于人物家庭网络构建的情况。

表 6 家庭网络性能

家庭网络构建方法	发现的家庭总数	正确的家庭数目	P (%)	R (%)	F1(%)
未引入单文本指代消解	176	81	84.7	64.1	73.0
引入单文本指代消解后	224	115	80.3	76.7	78.5

从表中可以发现，本文的方法可以发现更多的家庭，家庭的召回数量有显著提升。但使用单文本指代消解技术时，新发现的人物关系对的准确率相对较低，噪音的引入，对家庭网络的融合存在很大影响。如例句“*李鹏 请 韩升洲 转达 [他] 和 夫人 朱琳 对 金泳 三 总统 和 夫人 的 亲切 问候 和 谢意 。*”本文的方法在经过单文本指代消解后会形成一对具有夫妻关系的错误实例对“*韩升洲*”和“*朱琳*”，而这对错误实例对将在进行家庭网络融合时与正确的夫妻实例对“*李鹏*”和“*朱琳*”构建成一个具有 1 个妻子、2 个丈夫的错误家庭，这样的情况将影响家庭网络构建的准确度。

6 总结和展望

本文给出了一种基于单文本指代消解技术来构建人物家庭网络的方法。先通过自举技术学习进行人物关系抽取并习得家庭关系模式；再对模式进行泛化，用新的模式进行文本匹配后，之后对匹配到的句子进行单文本指代消解；最后使用跨文本指代消解技术将抽取出的人物家庭关系进行融合，形成家庭网络。实验结果显示，本文提出的单文本指代消解方法可以

有效地拓展人物家庭关系的抽取规模，从而提高所构建出的人物家庭网络的召回性能。

虽然目前通过单文本指代消解技术使得家庭网络的召回数目有所提升，但仍然存在着召回数量不足、家庭类型不够丰富、不同家庭之间没有联系等问题。下一步工作中，计划考虑更多家庭内部的人物关系类型，融入不同家庭之间存在的联系，同时考虑零指消解等进一步扩大家庭规模的方法，从而进一步丰富人物家庭关系网络。

参考文献

[1] Kautz H, Selman B, Shah M. ReferralWeb: Combining Social Networks and Collaborative Filtering[J]. Communications of the ACM, 1997, 40(3): 63-65.

[2] Mika P. Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2005, 3(2): 211-223.

[3] Tang J, Zhang J, Yao L, et al. ArnetMiner: Extraction and Mining of Academic Social Networks[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2008: 990-998.

[4] Elson D K, Dames N, McKeown K R. Extracting Social Networks from Literary Fiction[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 138-147.

[5] Agarwal A, Corvalan A, Jensen J, et al. Social Network Analysis of Alice in Wonderland[J]. NAACL-HLT 2012, 2012: 88-96.

[6] Van De Camp M, Van Den Bosch A. A Link to the Past: Constructing Historical Social Networks[C]//Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. Association for Computational Linguistics, 2011: 61-69.

[7] Gu J H, Hu Y N, Qian L H, et al. Research on Building Family Networks Based on Bootstrapping and Coreference Resolution[C]//Proceedings of the 2nd Natural Language Processing and Chinese Computing. Springer Berlin Heidelberg, 2013: 200-211.

[8] Zhou G D, Zhang M. Extracting Relation Information from Text Documents by Exploring Various Types of Knowledge[J]. Information Processing & Management, 2007, 43(4): 969-982.

[9] Oh J H, Uchimoto K, Torisawa K. Bilingual Co-Training for Monolingual Hyponymy-Relation Acquisition[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Association for Computational Linguistics, 2009: 432-440.

[10] Zhang M, Su J, Wang D, et al. Discovering relations between named entities from a large raw corpus using tree similarity-based clustering[M]//Proceedings of the 2nd International Joint Conference on Natural Language Processing(IJCNLP-2005). Springer Berlin Heidelberg, 2005: 378-389.

[11] Zhu J, Nie Z, Liu X, et al. StatSnowball: A Statistical Approach to Extracting Entity Relationships[C]//Proceedings of the 18th International Conference on World Wide Web. ACM, 2009: 101-110.

[12] Peng C, Gu J H, Qian L H. Research on Tree Kernel-Based Personal Relation Extraction[C]// Proceedings of the 1st Natural Language Processing and Chinese Computing. Springer Berlin Heidelberg, 2012: 225-236.

[13] Agarwal A, Kotalwar A, Rambow O. Automatic Extraction of Social Networks from Literary Text: A Case Study on Alice in Wonderland[C]//Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP-2013). 2013.

[14] Gordon P C, Grosz B J, Gilliom L A. Pronouns, Names, and the Centering of Attention in Discourse[J]. Cognitive Science, 1993, 17(3): 311-347.