句间回指模糊的多层次消解模型

宋培彦,刘宁静
SONG Pei-yan,LIU Ning-jing

北京师范大学 中文信息处理研究所,北京 100875
Institute of Chinese Information Processing,Beijing Normal University,Beijing 100875,China
E-mail:spy@mail.bnu.edu.cn

SONG Pei-yan,LIU Ning-jing.Multi-levels model for G-anaphora disambiguation between sentences.Computer Engineering and Applications,2010,46(5):24-27.

Abstract: G-anaphora disambiguation between sentences has become an important issue in Chinese information processing currently.Based on semantics and pragmatics,the anaphora phenomenon in legislation texts is formalized in the multi-levels model, which can be used in Chinese sentences group G-automatic processing.Antecedent and reference expression can be linked by concept primitive,57 basic sentence patterns provide framework to identify the relations,and context unit contains the macro knowledge for them.The analysis of corpora indicates that the model is efficient in G-anaphora disambiguation between sentences.
Key words: anaphora disambiguation;formalization of context;semantics chunks share;natural language understanding

摘 要:句间回指消解是当前中文信息处理的一个重要研究课题,直接从语义和语用入手,以法律文本为语料来源,对句间回指进行形式化描述和消解,服务于计算机句群自动理解。概念基元是"显微镜",看清指代语与先行语的微观语义联系;句类是"放大镜",将指代语和先行语纳入57组基本句类中进行关联;语境单元则是"望远镜",为指代语和先行语提供宏观的语境知识。语料考察结果表明,这一多层次的消解模型对实现句间回指消解是有效的。
关键词:回指消解;语境形式化;语义块共享;自然语言理解
DOI:10.3778/j.issn.1002-8331.2010.05.008 **文章编号**:1002-8331(2010)05-0024-04 **文献标识码**:A **中图分类号**:TP391

回指(Anaphora)作为指代(Reference)的一种常见形式,是实现语篇连贯的重要手段,成为当前自然语言理解领域的重要课题。

例句 1 一九一一年孙中山先生领导的辛亥革命,废除了封建帝制,创立了中华民国。

例句 2 单位$_1$犯罪的,对单位$_2$判处罚金,并对其$_3$直接负责的主管人员和其他直接责任人员判处刑罚。

例句1中共有两个小句,前一小句的"一九一一年孙中山先生领导的辛亥革命"作为语义块被后一小句共享,构成回指。例句2中的代词"其"指向前面的"单位$_1$"和"单位$_2$",用代词实现回指。

前指是语言中最常见的指代形式,但是后指现象也并不罕见,如:

例句 3 本法所称公民私人所有的财产,是指下列财产:(一)公民的合法收入、储蓄、房屋和其他生活资料;(二)依法归个人、家庭所有的生产资料;(三)个体户和私营企业的合法财产;(四)依法归个人所有的股份、股票、债券和其他财产。

例句3中的"下列财产"涵盖了(一)~(四)这四种情况,属于后指。后指现象在法律文本语料中屡见不鲜。

根据不同的视角,回指可以分为单文本回指和跨文本回指,也可以细分为句内回指和句间回指,主要关注单文本中的句间回指消解,包括前指和后指两种形式。

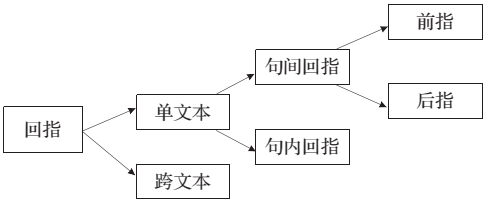

图1 回指类型划分

事实上,前指和后指只是语言表层结构的不同,在语义上并没有本质区别。将前指和后指统称为回指,英文翻译为G-Anaphora,是指语言结构中的某个成分指向先前表达过或随后出现的某个成分,指代语所指的内容称为先行语;把确定指代语的先行语的过程称为回指模糊消解(G-Anaphora Disambiguation),简称回指消解。

回指反映的是概念之间的关联,对计算机来说这种关联是一种指代模糊,回指消解是使这种指代模糊关系明确化,这对于机器翻译、信息抽取、自动文摘等具有非常重要的应用价

基金项目:国家"十一五"科技支撑计划重大项目资助(the National Great Project of Scientific and Technical Supporting Programs Funded by Ministry of Science & Technology of China During the 11th Five-year Plan. No.2007BAH05B01)。
作者简介:宋培彦(1980-),男,博士,主要研究方向:中文信息处理;刘宁静,女,博士,研究方向为中文信息处理。
收稿日期:2009-11-10 **修回日期**:2009-12-25

值(侯敏等,2004[1];Hobbs,1976[2];Yang X.F,Su.J and Tan C.L,2006[3]),是近年来自然语言理解领域的热点领域(周国栋等,2007[4])。MUC、ACE、TREC 等一系列评测活动大大推动了这一领域的研究。

黄曾阳先生创立的概念层次网络理论(Hierarchical Network of Concepts,以下简称 HNC 理论)(黄曾阳,1998[5];苗传江,2005[6];黄曾阳,2004[7]),试图以概念联想脉络为主线,深入自然语言的语义和语用层面,建立一种模拟大脑语言感知过程的自然语言表述模式和计算机理解处理模式,使计算机获得消解模糊的能力。回指现象属于 HNC 五重模糊中的"指代冗缺模糊",是 HNC 重要的研究内容。概念基元(HNC1)、句类理论(HNC2)和语境理论(HNC3)数学表示式的提出提供了良好的理论基础(黄曾阳,2004[7])。HNC 团队近年来在指代、省略、语义块共享、句群划分、知识库建设等领域(张全等,2007[8];贾宁、张全,2008[9]),取得了较大的进展。在这些成果的基础上对句群中的回指现象进行研究的。

汉语"重意合、轻形合",话题优先、对语境强烈依赖是其显著特点,墨守先句法、后语义的传统成规,将汉语强行纳入上下文无关文法 CFG 进行剖析,必定会得到大量"成分残缺不全"的句子,语义处理能力也难以提高;单纯地用目前流行的统计方法来获取深层的语义、语用知识,脱离语言学基础理论和知识库的支持,则无异于问道于盲。因此,在句群这一更大的视野内,以语义和语用等人类先验知识统摄全局,从概念、句类、语境 3 个层次上对句间回指关系进行消解,是一种可行的办法。多层次消解模型就是对回指消解的有益尝试。

1 回指现象分析

1.1 语料来源

法律语言学将法律文本分为立法语言、司法语言和执法语言 3 种类型。与司法语言、执法语言相比,立法语言的表述更为严谨、规范性更强:所用词汇均属于高频词,基本上都是陈述句,没有出现疑问句和感叹句,指代明确,语义严谨,法律领域特色明显,句子表述极为严密而不失灵活,反应了现代汉语的基本面貌,比较适合于计算机处理。选取的语料主要为中华人民共和国宪法(以下简称"宪法"),作为我国的根本大法,具有最高法律效力,经过若干次修订和完善,成为我国法律体系中的最具代表性的法律,集中体现了立法语言的上述特点。

1.2 语料标注

以宪法为语料来源,经过手工选取和标注,得到 210 个有代表性的句群,标注过程如图 2 所示。

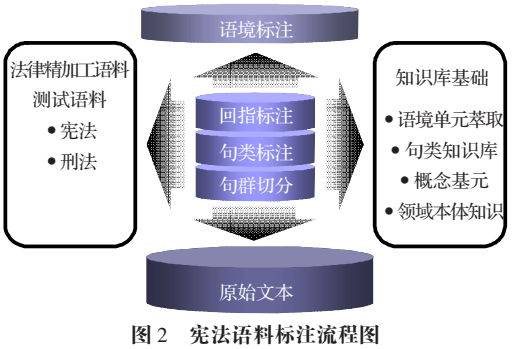

图 2 宪法语料标注流程图

语料标注的主要步骤和内容是:

步骤 1 句群切分。HNC 认为,"扣题就自然形成句群,句群就是围绕着一个特定概念展开的话题"。为了缩小研究范围并精确反映句间的回指关系,据此将句群进一步限定为:围绕同一话题,由两个或两个以上分句组成的语言单元。主要依据句号为标记切分出句群;当句群与语感不符时,以语感为准,力求形成较为完整、连贯的句群。

步骤 2 句群和回指基本标注。对句群中的每个小句逐句标注句类、格式代码、语义块、回指、句间关联语等信息。存储为 XML 格式,采用 XML Schema 验证,用于考查和验证概念基元和句类对回指消解的作用。

步骤 3 语境单元信息标注。对步骤 2 中得到的语料,抽取出若干有代表性的句群,标注语境单元知识,用于考查和验证语境对回指消解的作用。

1.3 语料分析

步骤 1 在收集到的 210 个句群中(句群的实际数量可能会略有变化,主要取决于划分句群的标准),共出现句间回指现象 612 处,根据指代语的类型,可以把这些回指现象分为代词回指(人称代词和指示代词)、名词回指(主要是 p、w 具体概念和 g、z、r 抽象概念)和语义块共享三种类型。

由代词作为指代语构成的回指在语料中共出现 46 次,约占总数的 7%。人称代词一般可以独立作为指代语,而指示代词则往往作为语义块的限定成分;没有出现"那"、"某些"、"这里"等语义指向不明的代词。HNC 映射符号的主体和概念类别符号分别是:

	HNC 符号	概念类别符号	词语
人称代词	p400m m=0-7	pl9	它;它们;他们
指示代词	l9yu OR lj9yu	lug9 OR lg9	这些;各种;其他

g 型、z 型和 r 型概念表示抽象概念,p、w 类概念可用于对具体事物的命名,这两类概念大致相当于一般所说的"名词",以此作为指代语构成的回指在语料中共出现 161 次,约占总数的 26%。p、w 概念包括泛指和特指两种类型,其中一些属于"未登录词",常用的有:

fpj2*	特指的国家、省、县
fpwj2*	特指的城市
fwj2-	特指的地区、地点
fwj2*k	特指的洲、洋、海、山、河、湖、岛、沙漠、平原等
fpemn*k	特指的政党、公司、组织、团体、家族
fpmn*k	特指的名人
fpwmn*k	特指的品牌

回指可以不出现指代语,通过语义块的共享实现补充,构成迭句、链句等句子形式,相当于语言学中所说的"零形回指"。这类回指出现次数多达 405 次,约占总数的 67%。详情见表 1。

表 1 回指类型及其分布

类型	概念类别	出现次数	所占比例/(%)
代词性回指	pl9、lug9、lg9	46	7
名词性回指	部分 g 型、z 型和 r 型,p、w 类概念	161	26
语义块共享	迭句、链句等句子形式	405	67

实现回指消解,首先要从当前词语中预先判断是否可能发生回指,称为回指判定;然后,根据某种策略,将候选指代语选择最恰当的指代语,与先行语建立关联,称为回指恢复。

2 消解策略

概念基元知识 HNC1、句类知识 HNC2 和语境单元萃取 HNC3

可以为回指判定和恢复提供支持,通过“领域句类表示式+HNC映射符号”获得预期知识。实现回指消解,主要取决于两类函数关系和领域知识本体。

2.1 情景 SIT 和事件背景 BAC 是领域的函数

也就是说,一旦确定了领域,情景和背景就是确定的。领域信息一般蕴含在扩展基元概念符号体系中,用于对人类活动进行描述。消解时应充分考虑概念基元所蕴含的领域信息。

表 2 领域概念总览表

类型	符号
心理、意志、行为	71,72,73
人类思维活动	8
第二类劳动	a
追求与理念活动	b,d
第一类劳动	q6
业余活动	q7
信仰活动	q8
本能活动	6m(m=0~5)
灾祸	3228α(α=8-b)
状态	503,50α(α=8-b)

语境单元 SGU 有限性的本质在于领域 DOM 类型的有限性、领域句类 SCD 数量的有限性和事件背景 BAC 类型的有限性。在这个三维空间中,领域 DOM 作为主轴,其情景 SIT 和事件背景 BAC 都是领域的函数。“宪法”属于第二类劳动 a5,文体为论述 Description 型,记为 SGUD,语境单元知识有以下线索可资利用:

(1)背景 BAC:基本背景包括法律文本涉及的立法机关、语言风格、日期等元数据信息,为静态信息;还包括句群中的方式 Ms、途径 Wy、起因 Pr 与目的 Rt 的描述,即辅语义块 fK;领域 DOM 也蕴含了某些特定的背景知识。

(2)领域 DOM:《HNC 理论全书》对法律领域知识设立了独立的概念节点 a5,对宪法所涉及的世界知识也进行了详尽的描述,有利于在此基础上设计独立的宪法领域句类表示式。

(3)情景 SIT:领域句类表示式以符号 SCD 表示,由 SIT=SCD(A,B,C)可知:A,B,C 是并列关系,其出现顺序取决于具体的领域句类表示式。

领域句类表示式相当于句群的“摘要”或“模板”,它规定了该领域的句群所应有的语义块及其概念类别,通过人工总结、先验地赋予计算机,使计算机获得世界知识。

2.2 语义块是句类的函数

句类表示式提供了一个句子必须有什么的前提,因此,它是语义块整体缺省判断的依据。同理,语义块构成表示式提供了语义块要素必须有什么的前提,因此它是语义块要素缺省判定和恢复的依据。

句类的数学表达式定义为:

SC=GBK1+EK+GBKm(m=2~4) (HNC2)

其含义是,用 57 个基本句类及其两相混合就可以描述汉语的全部句子;1 个句类最多包含 4 个语义块,GBK 是句子的广义特征语义块,EK 是句子的核心特征语义块。不同的句类中语义块出现的顺序和格式也不同。

2.3 法律领域知识本体

知识本体(Ontology)是对概念体系的明确的、形式化的、可共享的规范。法律领域知识本体(domain ontology)是对法律领域中的客体进行分析,反映了这些客体之间的关系,以明确的、形式化的、可共享的方式描述各个客体所代表的概念。法律领域知识本体表现为一个相互关联的概念词表,有助于对领域知识进行系统的分析,把领域知识形式化,减少语义的模糊性,为计算机处理特定领域的知识提供便利。

总之,HNC 是研究的主要理论基础。概念基元是“显微镜”,看清指代语与先行语的微观联系;句类是“放大镜”,将指代语和先行语纳入 57 组基本句类中进行关联;语境单元则是“望远镜”,为指代语和先行语提供宏观的语境知识。回指消解需要在这 3 个层面上进行研究。如图 3 所示。

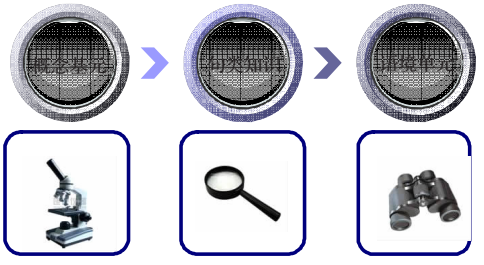

图 3 回指消解的 3 个层次

3 消解模型

对一个句子或句群的句类分析过程,HNC 的相应处理策略是:中间切入,上下并进。

具体操作为:

	中(切入点)	上	下
句子	语义块感知	句类代码认定	语义块构成分析
句群	领域句类表示式认定	句间关系认定	语义块转换、并合

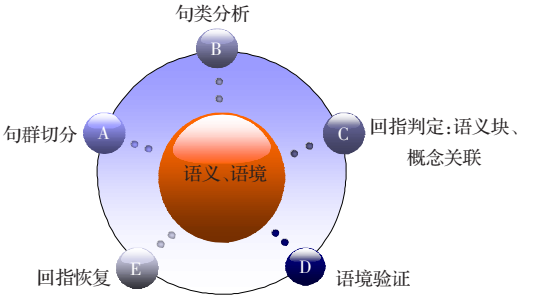

图 4 回指消解流程示意图

按照“中间切入,上下并进”的策略,从句类分析入手,底层依靠概念基元,高层依靠语境单元分,为多个阶段逐步进行,如图 5 所示。

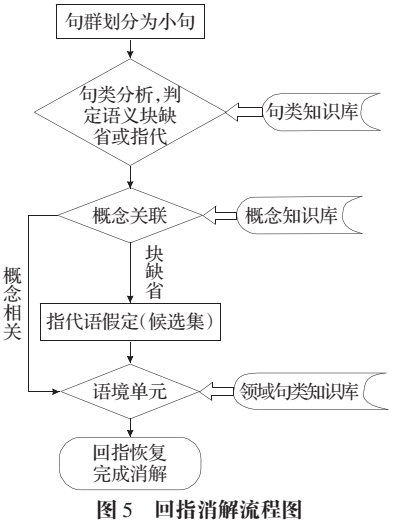

图 5 回指消解流程图

具体过程是:

(1)初始化阶段:对原始文本进行切分,以标点符号为依据得到若干句群,回指消解将在句群范围内部进行。

(2)句类分析阶段,获得各个小句的句类知识,如语句格式、语义块构成、句类转换和语义块变换等。对各个小句的词语序列 w_1~w_n,搜索 HNC 词语知识库,获取词语映射符号,考察词语之间的概念关联性;并激活人工预置的领域句类表示式 SCD。

(3)语境单元萃取阶段。通过步骤(2)领域句类表示式SCD,获得领域 DOM 信息。背景和情景除了来自于领域信息,还可以从语篇的元数据和句群中的辅块(条件、参照等)获得。语境单元知识可以对回指消解起到验证作用。

4 回指消解示例

例句 各少数民族聚居的地方实行区域自治,设立自治机关,行使自治权。各民族自治地方都是中华人民共和国不可分离的部分。——摘自《中华人民共和国宪法》

步骤 1 句群切分阶段。

句群的 4 个小句分别为:

小句 1:各少数民族聚居的地方实行区域自治,

小句 2:设立自治机关,

小句 3:行使自治权。

小句 4:各民族自治地方都是中华人民共和国不可分离的部分。

句群有两个主要特点:(1)虽然语料包含了两个句号,但两个句子在语义上联系非常紧密, 因此按照语感应视为一个句群。(2)存在多种类型的回指现象:小句 1 的 JK1"各少数民族聚居的地方"被小句 2 共享,构成迭句;小句 2 的 JK2"自治机关"被小句 3 共享,构成"链句";小句 1 的 JK1"各少数民族聚居的地方"与小句 4 的 JK1"各民族自治地方"构成同指关系。

步骤 2 句类分析阶段。对句群中的各个小句进行句类分析,获得句类代码和句类表示式。如果句类分析已经成功,根据 HNC 的基本观点——语义块是句类的函数,那么语义块的个数、概念类别以及次序也就确定了。各小句的表示式分别为:

小句 1:XJ=A+X+B

小句 2:XJ=A+X+B

小句 3:R511=RB1+R+RB2

小句 4:jDJ=DB+jD+DC

根据句类表示式,写出语义块的概念类别。

小句 1:{wj2/各少数民族聚居的地方}+{va02}+{ga108}

小句 2:{ }+{va01+v311}+{px+(v44e61,l12,pw008)}

小句 3:{ }+{vc451}+{rc44e61+(ga11;ga01)}

小句 4:{wj2/各民族自治地方}+{jl11}+{jgu40-0}

此时第 2、3 小句的{ }内出现空白,表示在语言空间出现省略或语义块共享,需要进行回指消解,按照步骤 3~5 顺序执行。小句 1 和小句 4 中的 JK 概念类别相同或接近,则假设其为同形回指,直接进入步骤 5"语境单元验证"。

步骤 3 激活预期知识。根据语义块的概念关联知识和同行优先性,在空缺处激活预期概念类别知识,显然,这是在概念空间的操作。

小句 2、{wj2}+{va01+v311}+{px+(v44e61,l12,pw008)} //设立

小句 3、{pe;wj2}+{vc451}+{rc44e61+(ga11;ga01)} //行使

步骤 4 获取候选指代语。从当前位置分别向前和向后紧挨的句子搜索,判断各个块是否符合步骤 3 中的预期知识。按照符合程度进行排队,优先级高的在前,优先级低的在后,完全无关的则跳过,以便减少回溯(backtrack),提高性能。此时,句 2 中的 JK1 只有一个候选指代语,说明匹配成功,可以直接进入步骤 5"语境单元验证";句 3 因为有两项可能的匹配,需要依靠语境知识进一步验证。可见,候选指代语是通过向前和前后搜索获得的,是对语言空间里的符号匹配操作。

小句 2、{wj2/各少数民族聚居的地方}+{va01+v311}+{px+(v44e61,l12,pw008)}

小句 3、{pe/自治机关;wj2/各少数民族聚居的地方}+{vc451}+{rc44e61+(ga11;ga01)} //行使

步骤 5 语境单元验证。通过语境单元信息,确认是否消解成功。

通过句群中的领域词汇"自治机关"、"设立"、"行使",以及标题"中华人民共和国宪法"、发布机关"全国人民代表大会"等背景知识,可以判定句群领域为法律活动 a5,进而激活法治活动的领域句类表示式:

SCD=SCD(a51e21)=Cn-1|ReC(RtC)D01-42XY*211J

该表示式对应的世界知识是"政府机关按照法律行使权利",先验地赋予计算机。

SGUD=(8y:|DOM;SIT;BACE;BACA) (HNC3-2)

SIT=SCD(A,B,C) (HNC3a)

领域 DOM:法治政府侧面 a51e21,政府依法行事。

SIT=SCD(A,B,C);A(政府机关)、C(行使权利)

BAC——背景(宪法强流式关联于政治制度;宪法可以对民族关系作出规定)

根据这一预期知识,可知小句 2 中的候选指代语可以共享小句 1 的 JK1,构成回指;小句 4 的 JK1 可以作为指代语,指向小句 1 的 JK1,回指关系成立。

小句 1:{wj2/各少数民族聚居的地方}+{va02}+{ga108}

小句 2:{wj2/各少数民族聚居的地方}+{va01+v311}+{px+(v44e61,l12,pw008)}

小句 4:{wj2/各民族自治地方}+{jl11}+{jgu40-0}

小句 3 有两个候选指代语,即{pe/自治机关;wj2/各少数民族聚居的地方},在语境单元验证阶段,根据"政府机关行使权利"这一法治活动的领域句类知识,小句 3 中的 A 语义块只能是"e/自治机关","wj2/各少数民族聚居的地方"不符合这项领域世界知识,应予排除。消解成功。

5 小结

句间回指消解是一个非常重要而又富有挑战性的研究课题,在 HNC 理论的指导下,对法律文本的句间回指消解进行了探讨,描述了实现句间回指消解的基本思路和方法,具有较强的可行性。直接从语义和语用入手、采取演绎和归纳的研究方法,对回指进行形式化描述和消解,这是主要思路,从已经标注的语料来看,这种模型对回指消解具有较好的处理能力,能够涵盖绝大多数回指现象。今后的工作重点是加强概念基元(词语级)、句类(句子级)、语境单元知识(句群级)等知识库建设,增加语料类型和数量,并进行上机实验,提高回指消解的准确率。

参考文献:

[1] 侯敏,孙建军.汉语中的零形回指及其在汉英机器翻译中的处理对策[J].中文信息学报,2005(1):14-20.

(下转 122 页)