

文章编号: 1003-0077(2016)04-0090-08

# 基于中心理论和话语结构的交互式问答文本指代消解

李映,孔芳

(苏州大学 计算机科学与技术学院,江苏 苏州 215006)

**摘要:** 与传统新闻文本相比,交互式问答中蕴含着更为丰富的语言现象。在传统的针对新闻文本的指代消解方案的基础上,融入了交互式问答特有的特征集,给出了一个适于交互式问答文本的指代消解方案。具体而言,基于浅层语义角色分析的结果进行话语结构的识别,根据识别出的话语结构进行话语中心及中心跳转的识别。将获取到的话语中心及跳转信息组织成交互式文本特有的特征集,使用交互式问答领域广泛使用的 TREC2004 和 TREC2007 的评测语料进行指代消解的实验,结果表明给出的方案能大大提高交互式问答文本中指代消解的性能,系统 F 值提高了 3.2%。

**关键词:** 指代消解;交互式问答;中心理论;话语结构

**中图分类号:** TP391      **文献标识码:** A

## Centering Theory and Discourse Structure Based Approach to Coreference Resolution for Interactive Question Answering Text

LI Ying, KONG Fang

(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

**Abstract:** The interactive question answering texts is rich in linguistic phenomena. Taking this a advantage, a novel coreference resolution approach for interactive question answering text is proposed. On basis of shallow sematic role analysis, the discourse structure is identified, upon which the preferred center and types center shift are further identified. These form a new feature set related to centering theory and discourse structure. Experiments on TREC2004 to TREC2007 corpora show that the proposed approach can significantly improve the performance of coreference resolution about 3.2% in F-measure for interactive question answering texts.

**Key words:** coreference resolution; interactive question answering; centering theory; discourse structure

自动问答(Question Answering)是指在对用户以自然语言形式提出的问题进行分析、理解的基础上,借助各类信息检索技术给出回答。自动问答给出的答案并非只是相关文档,而是以自然语言形式重新组织的文本片段。传统的问答技术局限于对单一问题进行回答,因此对问题的理解相对片面,提供的答案的信息量有限,不能满足用户了解同一实体或事件多方面复杂信息的需求。近年来新提出的交互式问答(Interactive Question Answering, IQA)很好地解决了这一问题。用户提出一系列主题相关的问题,在综合分析这些问题的基础上,交互式问答系统能以对话的形式逐个回答用户关于同一实体或事件不同方面的问题<sup>[1]</sup>。交互式问答作为自动问答的

一项重要技术受到了越来越多的关注。文本信息检索会议(TREC)在 2004 年的 QA 任务中开始以系列问题的方式给出问题评测集<sup>[2]</sup>,每个系列针对一个 Target(对象),系列问题中的每个问题都与此 Target 相关,如图 1 所示的例子。

交互式问答相关的应用也逐渐进入我们的生活,例如,苹果公司的 siri 语音助手、许多领域相关的自动咨询机器人等。但至今还没有能够通过“图灵测试”<sup>[3]</sup>的智能机器人出现,其中一个重要的原因是自然语言中存在众多机器难以理解的语言现象,这些语言现象使得用户提出的问题很难被机器充分理解,从而给出正确的反馈。在这些语言现象中,指代是最为常见的语言现象之一。对于前文中提及过

的语言单位,或是当前阐述的中心对象,为了表达的简洁高效,都会选择指代词表示。例如,图 1 中,Q2 中的“his”和 Q3 中的“he”都指代 Q1 中的“Jar Jar Binks”。指代词可以是代词,也可以是指示性名词短语,或是别名形式的其他表述方法。指代现象的存在成为机器流畅理解问题的阻碍,因此需要为篇章中的指代词找回指代的语言单位,以便机器能更准确、全面地理解用户的提问,而为指代词找回指代的语言单位的过程称为指代消解。

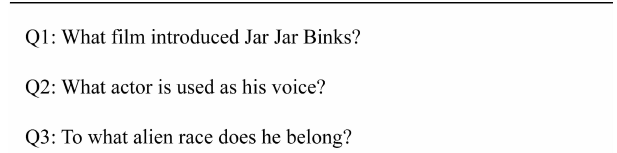


图 1 系列问题示例

近年来,国内外许多学者对指代消解进行了大量细致的研究,但大部分研究都集中在对新闻文本中指代消解的研究上,对交互式问答中指代消解的研究相对较少。本文考虑交互式问答文本的特点,结合传统的针对新闻文本的指代消解解决方案,给出了一个融合中心理论和话语结构的针对交互式问答文本的指代消解解决方案。在交互式问答文本语料(TREC2004 和 TREC2007 的评测语料)上的实验表明,该方案能大大提高交互式问答文本中指代消解的性能,系统 F 值提高了 3.2%。

本文结构安排如下:第一部分介绍与本文相关的工作;第二部分基于传统的针对新闻文本的指代消解方案给出了一个基准系统;第三部分介绍融入了中心理论和话语结构的指代消解方案;第四部分给出了交互式问答语料上指代消解实验的结果及分析;最后总结全文。

1 相关工作

本节分别介绍交互式问答和指代消解领域的相关研究。

1.1 交互式问答

Chai 等人<sup>[4]</sup>以及 Tsuneaki 等人<sup>[5]</sup>发现仅仅一个问题往往满足不了用户的需求,通常用户的问题总是围绕一个特定的主题,或者是希望解决一个特定的任务,因此他们指出以对话形式获取的信息比一问一答的形式更加全面和准确。近年来交互式问答的重要性日益凸显,相关的研究也日益丰富。

交互式问答主要是以对话形式解答用户的一系列问题,而 Carbonell 等人<sup>[6]</sup>和 Nils 等人<sup>[7]</sup>很早就提出,对话领域中频繁出现的指代和省略是人机对话处理的一大障碍。交互式问答的相关研究目前主要集中在问题理解方面,主要思想是提取与交互式问答相关联的上下文信息来辅助对问题自身的理解。代表性工作包括:Chai 等人<sup>[4]</sup>就问题的上下文信息能否对其他相关问题的理解以及答案的抽取有帮助这一问题展开了研究,提出了适于交互式问答的富语义话语模型,探讨了问题中的话语角色和问题间的话语转换等相关话题。Kirschner 等人<sup>[8]</sup>则针对问题分类展开,他们给出了一个利用浅层语义进行相似度计算,从而有效确定下一问题所属类型的方法。Bernardi 等人<sup>[9]</sup>在 Kirschner 等人研究的基础上进一步探讨了深层次的对话和话语结构特征对确定下一问题所属类型的作用。Wang 等人<sup>[10]</sup>提出了使用本体与模板方法提取交互式问答中的上下文信息的方案,并在传统问答系统 OTQAs<sup>[11]</sup>中应用这一方案动态提取上下文信息,构造了一个完整的中文交互式问答系统 OTCQAs。

本文主要解决影响机器理解问题的指代现象,将在 Chai 等人<sup>[4]</sup>提出的富语义话语模型的基础上,融合中心理论与话语结构,给出一个适于交互式文本中的指代消解的方案。

1.2 指代消解

作为信息抽取的核心组成部分之一,指代消解一直都是自然语言处理领域的研究热点。早期指代消解方法均采用启发式规则方法,从二十世纪九十年代开始,随着各类指代消解标注语料的不断发布,以及一些有影响力的自然语言处理会议和公开评测的召开,例如,MUC(Message Understanding Conference)、ACE(Automatic Content Extraction)、CoNLL shared task 等,指代消解的研究重点也转向了数据驱动的基于机器学习的指代消解方法研究。

目前,基于机器学习的指代消解研究可以分成四类:一是基于表述对(mention-mention)的消解方案,基本思想是从文本中提取各种表述,根据前后关系两两配对,针对配对的两个表述分别提取各类词法、语法和语义特征借助机器学习方法生成训练模型,测试时仍然针对配对好的两个表述进行判断,确认它们间是否存在指代关系,最后根据指代关系的传递性构建生成指代链。这种方法的最大优势在于

简单,但仅考虑配对的两个表述的相关特征(局部特征)有时是不够的。代表性工作包括: Soon 等人<sup>[12]</sup>, Ng 等人<sup>[13]</sup>, Yang 等人<sup>[14]</sup>, Kong 等人<sup>[15]</sup>等;二是实体与表述对(entity-mention)的消解方案,基本思想是将已经形成的部分指代链(实体)与当前表述配对,再判别当前表述是否属于这一指代链,代表性工作包括: Yang 等人<sup>[16]</sup>, Rahman 和 Ng<sup>[17]</sup>等;三是实体与实体对(entity-entity)的消解方案,基本思想是借助聚类形成局部的指代链片段,再将指代链借助合并的方式扩充成完整的指代链,代表性工作包括: Stamborg 等人<sup>[18]</sup>;四是采用联合模型,将指代消解、Entity linking 和 NER 等任务进行联合学习,代表性工作包括 Hajishirzi 等人<sup>[19]</sup>。

目前大部分指代消解研究都集中在新闻语料上,针对交互式问答文本的指代消解研究极少,代表性工作包括张超等人<sup>[20]</sup>。本文主要针对交互式文

本中的指代现象,以中心理论为基础,结合话语结构,利用话语结构获取问题句中的优选中心,再结合话语结构中话语转换信息确定句子间的跳转关系,最终提出并实现了一个针对交互式问答文本的指代消解方法。

## 2 基准系统

与目前大多数指代消解研究类似,我们使用 Soon 等人<sup>[21]</sup>提出的基于机器学习的指代消解平台作为实验的基准系统,它的基本构成如图 2 所示。我们选取 Soon 等人提出的 12 个基本特征来构建指代消解基准平台,而待消解项识别模块则采用了孔芳等人<sup>[21]</sup>给出的平面特征集构建。限于篇幅,特征集请参见相关论文。

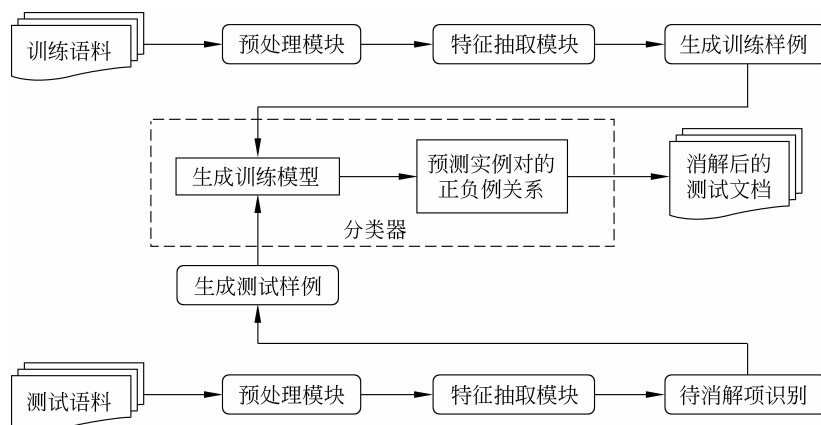


图2 基于机器学习的指代消解平台

## 3 应用中心理论和话语结构的交互式问答文本指代消解

相关研究工作表明合理应用中心理论可以进一步提升指代消解的性能。另一方面,话语结构与交互式问答关系密切,交互过程中话语的焦点(即中心)及其跳转、延续都将影响交互式问答中指代消解的性能。本节将针对交互式问答文本给出一个基于中心理论和话语结构的指代消解方案。

### 3.1 背景知识

#### 3.1.1 中心理论

中心理论是计算语言学中的一个理论模型,它主要分析了代词在语篇中的分布规律,以及影响代词实现的各种条件。中心理论详细、全面地论述了代词是如何促进语篇连贯的<sup>[22]</sup>,它指出语段中出现

的所有话语实体都是语篇的中心,这些中心在前后语段中的突显程度以及它们的语言实现形式都会影响到语篇的连贯性。语篇是以中心为基础连接前后语段的,每一语段都有三种中心。

1) 潜在中心(forward-looking center,  $C_f$ ): 是指一个语段可能存在的会话焦点,它提供了与后继语段联系的纽带,包括一系列的对象,这些对象按照在注意状态中突显度的不同形成一定的等级排列。

2) 现实中心(back-looking center,  $C_b$ ): 是指一个语段的当前会话焦点,它只包含一个对象,负责与先前语段建立联系,即前一语段的若干  $C_f$  中,突显度最高的一个对象就是本句的  $C_b$ 。

3) 优选中心(preferred center,  $C_p$ ): 是指潜在中心中突显度最高的那个对象。

在语段中区分出潜在中心、现实中心和优选中心的目的是为了说明各语段之间的连贯性。中心理

论根据前后两个语段(分别设为  $U_{n-1}$  和  $U_n$ )的三种中心的变化情况定义了四种前后语段的跳转类型,具体如表 1 所示。

表 1 中心理论的跳转类型

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

这四种跳转类型代表了前后两个语段间的连贯程度,显然语段连贯性由高到低的顺序为: Continue>Retain>Smooth Shift>Rough Shift。

在上述语段中心和跳转类型的基础上,中心理论进一步给出了一些规则和约束来描述篇章中代词的分布规律,表 2 给出了这些规则和约束。总结上述中心理论的规则和约束,孔芳等人<sup>[15]</sup>提出了如下结论。

表 2 中心理论的约束和规则

约束:
1) 一个篇章片段中的每一语段都可以有一组潜在中心,但现实中心 $C_b$ 最多只能有一个,或者没有。一般说来,篇章的第一个语段没有现实中心。
2) 每一个语段的潜在中心 $C_f(U_n)$ 是当前语段中的所有实体组成的集合。这些潜在中心在下一语段中起码应该部分实现。
3) 每一个语段的现实中心 $C_b(U_n)$ 是 $C_f(U_{n-1})$ 中的一个,而且是 $U_n$ 所实现的 $C_f(U_{n-1})$ 中阶(rank)最高的那个 $C_f$ 。
规则:
1) 前一语段的潜在中心 $C_f(U_{n-1})$ 中,存在某一元素,它在本语段中是以代词的形式实现的,则当前语段的现实中心 $C_b(U_n)$ 也必须以代词的形式表示。
2) 语段连贯性由高到低的顺序为: Continue>Retain>Smooth Shift>Rough Shift

1) 中心理论是与上下文相关的,中心是上下文的构成要素,是一个语义层的概念。

2) 前一语段的  $C_f$  中突显度最高的一个对象就是本语段的  $C_b$ 。Grosz 等人<sup>[23]</sup>进一步研究发现,语段中  $C_b$  是不受出现的先后次序、语法角色以及实施/受施等语义角色影响的。但位置以及语法角色等要素会影响  $C_f$  中各对象的突显度,其中语法角色是确定  $C_f$  中各对象突显度的关键要素之一。英语中  $C_f$  中各对象的突显度按照语法角色排列顺序为: 主语>宾语>其他。

3) 当本语段包含其它代词时, $C_b$  必须以代词的形式来表示,否则会造成这一语段阅读时间的增加。

4) 各语段间频繁发生硬转(Rough Shift)将会影响文章内容的连贯性。为了保持文章内容的连贯性,文章的作者必然会制定写作计划,减少焦点的切换次数。

中心理论是一个通用的语言学模型,并没有针对对话的特点展开讨论,因此各种中心的识别是非常困难的。本文在交互式对话这一背景下展开讨论,与对话中的话语结构相结合,在中心的识别上更容易实现。

3.1.2 话语结构

现实生活中,人们通常会借助一系列连续、相关的提问来获取某一感兴趣事物的多方面的详细信息。Chai 等人<sup>[4]</sup>指出要将所有相关问题联系起来,每一个问题和答案在整个问题集中的话语状态是一个重要信息。他指出话语状态包括两部分: 话语角色和话语转换。

在交互式问答中,每一个问题都有一个上下文情境。句子中的实体除了带有一定的语义信息,还常常带有句子的话语角色信息。实体能够表示两种不同的话语角色: 主题(Topic)和中心(Focus)。主题代表着一个问题所谈论的事物,而中心则是所谈论主题的一个方面。

话语转换预示着话语角色如何从一个问题到另一个问题间进行转变以及这种转变是如何反映了用户需求的信息。在交互式问答里,一个问题向另一个问题的转换过程中包含了许多上下文信息,这些信息可以用来处理问题和获取答案。问题的内容主要是围绕着问题的主题进行,主题如何演化与问题间如何关联关系密切。因此围绕问题的主题,可以把话语的转换分为以下三类。

(1) 主题的延伸: 下一问题的主题与前一问题相同,但参与者或某些方面略有不同;

(2) 主题的扩展: 前后问题的主题相同,但两个问题的中心不同;

(3) 主题的转移: 前后问题讨论了不同的主题。

3.2 中心理论与话语结构的结合

本节首先给出话语结构的识别方法,在此基础上介绍了基于话语结构的语段中心及其跳转类型的识别方法,并基于此进一步提出与中心和话语结构相关的新特征集,并将其应用于交互式问答文本的

指代消解。

### 3.2.1 话语结构的识别

获取话语结构信息是一项比较困难的任务,由于话语结构语料的匮乏,无法采用直接的有监督的机器学习方法。近年来浅层语义分析取得了长足的发展,语义角色标注(SRL, Semantic Role Labeling)性能不断提升,特别是英文的 SRL 系统已进入了可应用阶段。本文采用 Gildea 和 Surdeanu<sup>[24]</sup>等人提出的问句中语义角色识别方法进行交互式问答文本中问题句的语义角色标注,并基于话语角色和语义角色间的映射关系获取话语角色信息,再根据前后问题中话语角色信息自动标注话语转换的具体类别。

### 3.2.2 基于话语结构的中心识别

就英文而言,中心理论中指出语段中各对象突显性按照语法角色排序为:主语>宾语>其他。但语法与语义间仍然存在一定的分歧,不同的时态,不同的表述方法使得当前的话语中心并不一定充当主语。基于话语结构的中心识别方法的基本思想是,从语义视角获取中心,无法确定再由传统的语法角色排序决定。

在话语结构中主题是指当前问题的讨论范围,而中心是主题中特定的一部分,是指主题中当前问题最关心的一方面。主题和中心与句子的语义信息密切相关。其中,主题可以是实体(Entity)或活动(Activity)。当主题为实体时,该实体即为当前问题讨论的中心,就可以认为该实体就是当前语段的  $C_p$ ;而当主题为活动时,活动参与者的突显性应该高于语段中的其他实体,因此接着考察话语中心,如果中心是某一个主题(活动)的参与者时,则认为此参与者实体是当前语段中突显度最高的,即为本语段的  $C_p$ ;当主题是一个活动,并且话语中心也不是主题参与者实体时,即无法从语义视角确定  $C_p$ ,则按照传统的语法角色排序决定。 $C_p$  确定后,根据中心理论的定义可以很方便的得到下一语段的  $C_b$ 。结合话语结构,我们得到了图 4 所示的语段中心的识别方法。

### 3.2.3 基于话语结构的中心跳转识别

在识别出问题句中  $C_p$  和  $C_b$  后便可进一步识别出语段跳转关系。另一方面,话语结构中的话语转换从语义的视角描述了话语角色中主题和中心的转换关系。前一小节我们讨论了基于话语结构的中心识别方法,本节我们将进一步考虑在话语结构的基础上如何更有效地进行语段中心跳转类型的识别。

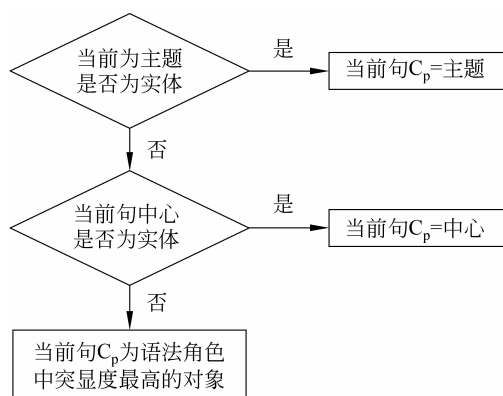


图 4 基于话语结构的中心识别方法

针对 3.1.2 节介绍的三种话语转换,我们分别给出了相应的语段跳转类型识别方法。

1) 话题的延伸: 表示下一个问题的主题与上一问题的主题相同,但参与者或某些约束不同,可将其细分为约束变化和参与者转移。如果主题为实体,由 3.2.1 节给出的中心识别方法可知它是当前语段的  $C_p$ ,且与上一语段的  $C_p$  相同,即  $C_b(U_n) = C_p(U_n)$ ,由此可以确定语段的跳转关系为延续或顺转;如果主题是活动,则进一步区分:(1)约束变化:指对主题的约束条件有变化,并不带来主题的参与者的变化,因此  $C_b(U_n) = C_p(U_n)$ ;(2)参与者转移:若当前话语中心是参与者,那么  $C_b(U_n) \neq C_p(U_n)$ 。此外其他情况我们均认为在话语转换过程中,语段各实体的突显度变化很小,语段的中心未发生变化,即  $C_b(U_n) = C_p(U_n)$ 。

2) 话题的扩展: 表示两个相连的问题有着相同的主题,但两者的中心不同。与话题延伸关系类似,因为前后问题有相同的主题,如果主题是一个实体,便可得到  $C_b(U_n) = C_p(U_n)$ ,由此可以确定跳转关系为延续或顺转。如果两个语段的  $C_p$  都是话语中心,或者两个语段的  $C_p$  只有一个是话语中心,因为中心发生了变化,可以得到  $C_b(U_n) \neq C_p(U_n)$ 。此外我们都认为语段中心没有发生变化。

3) 话题的转移: 表示两个相连问题的主题不同,问句关注的内容有了较大的变化。可以将话题转移细分成三类:(1)一个活动主题转移为另一个活动主题:话语中心都是这两个活动的参与者,那么,如果话语中心没有变化,为同一个参与者,则可认为  $C_b(U_n) = C_p(U_n)$ ;如果话语中心变化,则认为  $C_b(U_n) \neq C_p(U_n)$ ;(2)活动主题与实体主题之间的转移:此时考虑活动主题中的参与者中心是否为另一实体主题,如果是则认为  $C_b(U_n) = C_p(U_n)$ ,

否则  $C_b(U_n) \neq C_p(U_n)$ 。

根据  $C_b(U_n)$  和  $C_p(U_n)$  的关系,以及  $C_b(U_n)$  和  $C_b(U_{n-1})$  的关系,我们就能进一步识别中语段中心的跳转情况。

3.2.4 与中心和话语结构相关的新特征集

结合中心理论和话语结构确定语段的中心及其跳转,话语的主题和中心,以及话语的跳转后,我们提取了如表 3 所示的与中心和话语结构相关的新特征集,并将它们应用于基准平台。从表 3 给出的新特征集我们可以看到,新特征主要描述了四个方面

的信息。

- 1) 照应语和先行语候选所在问题在整个系列问题集中所处的物理位置(特征 1),不同位置的问题所含的已知/未知的信息量是不一样的;
- 2) 照应语和先行语候选的话语角色信息(特征 2~4)和语段中心信息(特征 5);
- 3) 照应语和先行语候选所在语段间的跳转关系(特征 6)和话语转换关系(特征 7);
- 4) 照应语和先行语候选间包含的其他语段间的语段跳转和话语转换关系(特征 8~9)。

表 3 中心及话语结构相关的特征集

	序号	特征	说明
话语角色相关	1	AN/CASenId	问题总数减去照应语/先行语候选所在问题的 ID
	2	AN/CAEntity	照应语/先行语候选是问题的中心取 1,否则取 0
	3	ANCAEntity	照应语和先行语候选均为问题中心取 1,否则取 0
	4	AN/CATarget	照应语/先行语候选是问题的主题取 1,否则取 0
	5	AN/CACp	照应语/先行语候选是本句中的 $C_p$ 取 1,否则取 0
话语转换相关	6	NTCenter	先行语候选与照应语在相邻问句且跳转关系为 Continue 取 1,retain 取 2,SmoothShift 取 3,RoughShift 取 4
	7	NTDiscourse	先行语候选与照应语在相邻问句且话语转换关系为话题延伸取 1,扩展取 2,转移取 3
	8	ALTShift	先行语候选与照应语间隔一问句且中间句与后一句的跳转关系为 Continue、retain、SmoothShif、RoughShift 分别取值 1-4
	9	CNTBoth	先行语候选所在句到照应语之间所有相邻句的跳转关系均为 Continue 时取 1,否则取 0

4 实验及分析

4.1 实验设置

本文采用 TREC2004 至 TREC2007 的 QA 评测任务的问题集作为实验语料,共计 286 个问题集 1 962 个问题,其中每个系列问题都标注了问题的主题(target)。参考传统的新闻类指代消解语料的标注方案,我们在此语料上增加了指代信息的标注,最终的标注结果如表 4 所示,从中可以看到,代词所占比重最大,约为 53.48%,有定名词也占据相当比重,约为 31.97%,专有名词约占 13.43%。由于语料规模较小,后续实验均采用五倍交叉验证,机器学习算法使用 SVM-Light 工具中的径向基(RBF)核

函数进行,所有参数均使用默认值。实验结果采用 Precision、Recall 和 F1-score 进行评测。

表 5 给出了基准系统在新闻类语料(ACE2003 NWIR)和交互式问答语料上的性能,我们可以看到,就系统总体性能的 F1 值而言,基准系统在新闻语料和交互式语料上的结果相差不大。但在交互式文本上占比重最大的代词的消解性能偏低,特别是召回率,仅为 23.8%,说明交互式文本中的代词与新闻文本中的代词使用上有较大差异。

表 5 基准系统在新闻语料和交互式问答语料上的性能

类别	ACE 2003 NWIRE 新闻语料			TREC QA 交互式 问答语料		
	P/%	R/%	F1	P/%	R/%	F1
代词	66.1	58.1	61.9	91.2	23.8	37.8
有定名词	79.2	29.1	42.6	52.6	55.3	53.9
专有名词	82.1	80.21	81.1	52.7	58.9	55.6
总体性能	64.5	51.9	57.5	50.1	66.5	57.2

表 4 语料中指代关系的各类别分布

类别	代词	专有名词	有定名词	其他	总数
数量	430	108	257	9	804

需要特别说明的是,中心理论和话语结构并不仅仅针对交互式文本,新闻类文本中也存在中心、话语角色以及话语转换等概念,但与交互式文本不同,新闻类文本中不存在话语的分割(即 utterance 的概念),而新闻类文本中话语自动分割的规范不统一,其性能也未达到服务于其他应用的程度。因此本文后续的实验仅针对交互式文本展开。

表 6 给出了应用中心理论和话语结构改进的指代消解系统在交互式问答语料上的性能,从中我们可以看到,提出的中心理论及话语结构相关的新特征能改进指代消解在交互式语料上的性能。

表 6 基于中心理论和话语结构的指代消解系统在交互式问答语料上的性能

	类别	P/%	R/%	F1
+ 话语角色 相关特征	代词	79.6	40.6	53.7
	有定名词	55.3	54.3	54.8
	专有名词	53.1	60.6	56.6
	总体性能	51.2	67.9	58.4
+ 话语转换 相关特征	代词	79.6	40.6	53.7
	有定名词	57.5	53.7	55.6
	专有名词	54.2	63.3	58.4
	总体性能	53.5	69.3	60.4

1) 相比基准系统,话语角色相关特征的引入提高了对所有类别的消解项的消解性能,其中对代词的贡献度最为明显。

2) 引入话语角色相关特征的基础上,再加入话语转换相关特征,从表 6 可以看到,代词消解的性能维持不变,有定名词、专有名词消解的性能有了一定的提升,最终系统的总体 F1 值提升了约 2%。

3) 相比基准系统,加入所有与中心理论和话语结构相关的新特征后,系统对各类别的消解性能均有不同程度的提高,总体性能的各项指标都有所上升,准确率上升了 3.4%,召回率上升了 2.8%,F1 值提高了 3.2%。

4) 相比基准系统,新加入的特征对交互式文本中比重最大的代词的消解,贡献度最大。召回率提高了近 17%,准确率下降了约 10.6%,F1 值提高了 16%。但 40.6%的召回率仍然不太理想,还有待进一步的提高。

5 结论

本文给出了一个结合中心理论和话语结构的针对交互式问答文本的指代消解方法。首先介绍了基于语义角色和话语角色映射关系的话语结构识别方法,接着给出了基于话语结构的语段中心及语段跳转关系识别方法,并基于此提出了与中心理论和话语结构相关的新特征集。交互式问答文本上指代消解的实验结果表明,新提出的特征集能够提升指代消解的性能,特别是代词消解的性能。

不过中心理论和话语结构都更关注文本的局部信息,如何更好的结合交互式文本的全局信息将是进一步的工作。此外,交互式问答文本中的答案对后续问题的主题、中心的转换也起着很重要的作用,如何在指代消解系统中考虑问题的回答也是未来的工作方向。

参考文献

[1] Nick Webb. Introduction of Interactive Question Answering Workshop[C]//Proceedings of the Interactive Question Answering Workshop at HLT-NAACL, 2006.

[2] Ellen M Voorhees. Overview of the TREC 2007 Question Answering Track[C]//Proceedings of the TREC, 2007.

[3] Alan M Turing. Computing Machinery and Intelligence[J]. Mind, 1950: 433-460.

[4] Joyce Y Chai, Rong Jin. Discourse structure for context question answering [C]//Proceedings of the Workshop on Pragmatics of Question Answering, HLT-NAACL, 2004: 23-30.

[5] Tsuneaki Kato, Fumito Masui Jun'ichi Fukumoto, Noriko Kando. Handling information access dialogue through QA technologies—A novel challenge for open-domain question answering [C]//Proceedings of the HLT-NAACL 2004 Workshop on Pragmatics of Question Answering, 2004: 70-77.

[6] Jaime G Carbonell. Discourse pragmatics and ellipsis resolution in task-oriented natural language interfaces [C]//Proceedings of the 21st Annual Meeting on Association for Computational Linguistics, 1983: 164-168.

[7] Nils Dahlback, Arne Jonsson. Empirical studies of discourse representations for natural language interfaces[C]//Proceedings of the fourth conference on European chapter of the Association for Computational

- Linguistics, 1989; 291-298.
- [8] Manuel Kirschner, Raffaella Bernardi, Marco Baroni, et al. Analyzing interactive qa dialogues using logistic regression models[C]//Proceedings of the AI \* IA : Emergent Perspectives in Artificial Intelligence, 2009; 334-344.
- [9] Raffaella Bernardi, Manuel Kirschner, Zorana Ratkovic. Context Fusion; The Role of Discourse Structure and Centering Theory[C]//Proceedings of the LREC, 2010.
- [10] D S Wang. Answering contextual questions based on ontologies and question templates[J]. Frontiers of Computer Science in China, 2011, 5(4): 405-418.
- [11] D S Wang. A domain-specific question answering system based on ontology and question templates[C]//Proceedings of the Software Engineering Artificial Intelligence Networking and Parallel/Distributed Computing (SNPD), 2010 11th ACIS International Conference on, 2010; 151-156.
- [12] Soon W M, Ng H T, Lim D C Y. A machine learning approach to coreference resolution of noun phrases[J]. Computational Linguistics, 2001, 27(4): 521-544.
- [13] Ng V, Cardie C. Improving machine learning approaches to coreference resolution[C]//Proceedings of the ACL 2002. 104-111.
- [14] Yang X F, Zhou G D, Su J, et al. Coreference resolution using competition learning approach[C]//Proceedings of the ACL 2003. 177-184.
- [15] Kong F, Zhou G D, Zhu Q M. Employing the centering theory in pronoun resolution from the semantic perspective[C]//Proceedings of the EMNLP 2009. 2009. 987-996.
- [16] Yang X F, Su J, Tan C L. A twin-candidate model for learning-based anaphora resolution[J]. Computational Linguistics, 2008, 34(3): 327-356.
- [17] Altaf Rahman, Vincent Ng. Coreference Resolution with World Knowledge[C]//Proceedings of the ACL 2011; 814-824.
- [18] Stamborg M, Medved D, Exner P, et al. Using syntactic dependencies to solve coreferences[C]//Proceedings of the Joint Conference on EMNLP and CoNLL-Shared Task, 2012.
- [19] Hannaneh Hajishirzi, Leila Zilles, Daniel S. Weld, et al. Joint Coreference Resolution and Named-Entity Linking with Multi-Pass Sieves[C]//Proceedings of the EMNLP 2013; 289-299.
- [20] 张超, 孔芳, 周国栋. 交互式问答系统中待消解项的识别方法研究[J]. 中文信息学报. 2014; 4(28): 111-116.
- [21] 孔芳, 朱巧明, 周国栋. 中英文指代消解中待消解项识别的研究[J]. 计算机研究与发展. 2012. 49(5): 1072-1085.
- [22] Peter C Gordon, Barbara J Grosz, Laura AGilliom. Pronouns, names, and the centering of attention in discourse[J]. Cognitive science, 1993, 17(3): 311-347.
- [23] Barbara J Grosz, Scott Weinstein, Aravind K Joshi. Centering; A framework for modeling the local coherence of discourse[J]. Computational linguistics, 1995, 21(2): 203-225.
- [24] Daniel Gildea, Daniel Jurafsky. Automatic labeling of semantic roles[J]. Computational linguistics, 2002, 28(3): 245-288.
- [25] Hobbs J R. Resolving pronoun references[J]. Lingua. 1978, 44(4): 311-38.
- [26] 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; 1987; Association for Computational Linguistics; 1987; 155-62.
- [27] Strube M. Never look back; An alternative to centering[C]//Proceedings of the 17th international conference on Computational linguistics-Volume 2; 1998.
- [28] Tetreault J R. Analysis of syntax-based pronoun resolution methods[C]//Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics; 1999.
- [29] Ng V. Semantic class induction and coreference resolution[C]//Proceedings of the ACL 2007; 536-543.
- [30] 孔芳, 周国栋. 基于树核函数的中英文代词消解研究. 软件学报[J], 2012. 23(5): 1085-1099.



李映(1976—), 硕士, 实验师, 主要研究领域为中文信息处理。

E-mail: liy@suda.edu.cn



孔芳(1977—), 博士, 副教授, 主要研究领域为机器学习, 自然语言处理, 篇章分析。

E-mail: kongfang@suda.edu.cn