交互式问答中基于话语结构的指代消解研究

张超，孔芳，周国栋

苏州大学 自然语言处理实验室，江苏 苏州，215006

E-mail: zcxyxywy@gmail.com

摘要：交互式问答系统能够与用户进行对话式交互进而处理用户提出的一系列问题，但指代现象阻碍了系统准确地理解用户的问题。本文深入研究了交互式问答中的指代消解任务，在交互式问答文本语料中标注指代链信息和话语结构信息，并在基准平台的基础上提出了基于话语结构的特征集。实验研究了新闻文本上提出的基准特征集在交互式问答文本中的性能，并在此基础上引入话语结构特征。实验结果表明，与基准系统相比，改进系统的 F 值提高了 2.6%，指代消解平台的性能有较大的提升。

关键词：交互式问答；话语结构；指代消解

Coreference Resolution in IQA based on Discourse Structure

Zhang Chao, Kong Fang, Zhou Guodong

Nature Language Processing lab, Soochow University, Suzhou Jiangsu 215006

E-mail: zcxyxywy@gmail.com

Abstract: Interactive Question Answering (IQA), a hot research topic in the area of QA, allows system to interact with users to process a series of questions just with a dialogue. This paper systematically explores the methodology of coreference resolution in IQA. We labeled coreference information and discourse structures in IQA text and experiment on the IQA text to study the flexibility of the baseline system. Then we proposed several features about discourse structures in IQA. Experimental results show that our features are effective.

Keywords: interactive question answering; discourse structure; coreference resolution

1 引言

问答系统（Question Answering，QA），作为一种特殊的信息检索方式，拥有比传统的信息检索技术更加智能更加便捷的特点。问答系统允许用户用自然语言表达自己的搜索需求，省去了对自己的需求进行关键词提取的步骤。问答系统还能够以自然语言的形式返回答案，而非只返回相关文档，然后还需让用户进一步提取信息。一直以来，问答系统都是自然语言处理领域的重要研究内容之一[1]。早期对于问答系统的研究主要集中在一问一答的形式，这种形式存在一个缺陷：单个问题提供的信息量太少，不能满足用户大信息量的需求。因此，近几年来越来越多的学者开始研究另一种形式的问答——交互式问答（Interactive Question Answering，IQA）。交互式问答系统允许用户提出一系列与主题相关的问题，以对话的形式逐个回答用户关于同一实体或事件多方面的问题[1]。为促进交互式问答的研究，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[2]，如例 1 所示：

Q1: What film introduced Jar Jar Binks?

Q2: What actor is used as his voice?

Q3: To what alien race does he belong?。

时至今日，还没有出现能够通过图灵测试的智能机器人，其中一个重要的原因是自然语言中存在着各种各样的机器难以理解的语言现象，这些语言现象阻碍了计算机充分理解用户话语表达中的所有含义。其中，指代现象是较为常见的一种语言现象。为了使表达简洁高效，用户会用指代词去指代文中出现过的语言单位，这就是指代现象。如例 1 中 *Q2* 中的 *his* 和 *Q3* 中的 *he* 是指代 *Q1* 中的 *Jar Jar Binks*。指代现象对计算机理解问句造成了阻碍，因此，需要为篇章中的指代词找回指代的语言单位，这个过程即为指代消解。

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新闻文本语料中的指代消解方法研究，交互式问答中的指代消解方法研究还较少。因此，本文主要研究了交互式问答文本的特点，在前人研究的基础上，提出了基于话语结构的交互式问答文本指代消解方法。实验结果表明，相比与基准系统，本文的方法在交互式问答文本中有更好的性能。

本文将在接下来的第 2 节中介绍交互式问答以及指代消解的相关研究;在第 3 节介绍本文使用的基准系统；第 4 节介绍交互式问答中的话语结构；第 5 节给出了实验设置及其结果和分析；最后一节是本文的内容总结以及工作展望。

2 相关工作

交互式问答技术的重要性日益凸显,自TREC2004的QA任务中加入系列问题任务以来,国内外许多学者相继开展了交互式问答技术的相关研究。Chai[3]等人较早地发现仅仅一个问题往往满足不了用户的需求,用户往往是想询问关于一个特定主题的信息或者是如何解决一个特定的任务，因此他指出以对话的形式获取的信息比一问一答的形式更加全面和准确。Dongsheng Wang[4]提出使用本体与模板的方法来利用交互式问答中的上下文信息，Joyce Y . Chai[5]根据交互式问答的特点提出了对话话语结构来使用上下文信息，但两者都并未真正在指代词上进行消解工作。交互式问答技术主要是通过对话解答用户的一系列问题,Carbonell[6]和 Nils[7]等都指出了对话领域中指代现象出现频繁，是计算机理解人机对话的一大障碍。

早期的指代消解主要是基于领域和语法知识构建复杂逻辑规则的方法，具有代表性的有：Hobbs 算法、中心理论、基于句法知识的方法等。随着标注语料库的不断出现，基于语料库的指代消解方法研究越来越多并且取得了比较好的性能。Dagan 等[8]提出优先考虑出现频率较高的先行语候选作为代词的先行语，对代词“it”的消解进行了研究。Ge 等[9]提出了一种基于贝叶斯概率统计模型的方法，并将它应用于单数第三人称代词的消解中。Cardie 等[10]提出了通过聚类方法进行名词短语的指代消解，主要是通过收集篇章中的基本名词短语，根据短语的特征对名词短语聚类，判断两个名词是否属于同一个类。McCarthy 等[11]把判断先行语的问题转换成分类问题,通过分类器判断指代语与每个先行语候选之间是否存在指代关系，这一思想为日后指代消解的主要框架。Soon 等[12]则首次给出了详尽完整的实现步骤，并开发出了实用的指代消解平台。之后，许多学者在此基础上作了许多不同程度的延伸。Ng 等[13]对 Soon 等的研究进行了扩充，抽取了 53 个不同的词法、语法和语义特征。目前，大多数指代消解系统都采用局部优化方法，即对于每个指代语，依据不同算法，选择最佳的先行语。

但至今大多数的指代消解都在新闻文本语料上进行。本文进行了在交互式问答文本上的指代消解研究，首先将基准系统运用到交互式问答文本中，观察在新闻文本上的指代消解方法的适应性。

3 基准系统

本文使用 Soon 等[12]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。它由三个基本流程构成：

1）训练生成分类器模型;

2）使用分类器模型进行分类;

3）对分类结果进行评测。

基准系统所使用的特征集如表 1 所示：

表 1 指代消解基准系统所使用的特征集

ANPronoun	照应语是代词取 1，否则 0
ANDefiniteNP	照应语是有定名词短语取 1，否则 0
ANDemonstrativeNP	照应语是指示性名词短语取 1，否则 0
CAPronoun	先行语是代词取 1，否则 0
ANCAGenderAgreement	照应语和先行语满足词性一致取 1，不一致取 0
ANCANumberAgreement	照应语和先行语满足单复数一致取 1，不一致取 0
ANCAAppositive	照应语和先行语是同位语 1，否则取 0
ANCAHeadStringMatch	如果照应语和先行语满足中心词匹配取 1，否则取 0
ANCASentDistance	照应语和先行语在一句内取 1，两句 0.9，….大于 10 句 0
ANCAWORDSENSE	从 Word Net 中获得的语义信息类有相同的为 1，不同为 0
ANCABothProperName	照应语和先行语候选词均为专有名词取 1，否则取 0
ANCANameAlias	照应语和先行语候选词存在别名关系取 1，否则取 0

在基准系统的基础上，根据交互式问答文本的特点，本文又提出了基于话语结构的指代消解特征集。

4 话语结构

Harabagiu 等[14]提出在问题和答案层面分别应用富知识型的自然语言处理技术能对问答系统的性能有显著的提升。Joyce Y.Chai 等[5]研究了交互问答中某一问题的上下文信息能否对其他问题的理解和答案的抽象有帮助的问题，在这基础上提出了富语义的话语模型，包括问题中的话语角色和问题间的话语转换。

4.1 话语角色

在交互式问答中，每一个问题都在一个有上下文的情境中。除了句子中实体带有的语义信息，每个问题中还都含有与整个问答相关的话语角色信息。

在一个完整的交互式问答中，用户不仅仅只问问题，也可能会去重复确认系统提出的问题或者是简单的确认。因此，Grosz 等[15]提出获取用户问题的意图（intention）是十分重要的。话语信息包括话语的话题和话语的中心等一些语义信息。除此之外，承载用户答案的媒体也是十分重要的，比如用户可能会要求一个人或者一个实体的照片或者视频资料。因此，Joyce Y.Chai 等提出了三种类型的话语角色：意图（Intent），内容（Content）和媒质（Media）。

1. 意图信息

意图信息可以进一步分为行为（Act）和动机（Motivator）。行为是指用户是从系统获取信息或者回复系统。动机表明了用户获取信息的方式，是对信息的检索还是对信息的确认。

2. 内容信息

根据各自的特点，内容信息可以分为目标（target）、主题（topic）和中心（focus）。目标是指问题所指向的答案的类型，包括实体（比如时间、位置、姓名等）和观点（比如原因、步骤、看法等）。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，是指当前主题中问题最关心的一方面。主题和中心与一个句子的语义信息有关联，主题可以有实体（Entity）和活动（Activity）。活动可以再细分为活动类型（ActType）、参与者（Participant）和边缘信息（Peripheral）。参与者是活动中具有不同语义类型的实体。实体又可以细分为语义角色、语义类型等语义信息。边缘信息是指活动的地点、时间、原因等。

3. 媒质信息

媒质信息是问题需求信息的媒介，可分为格式（Format）和类型（Genre）。格式包括有图片、表、文本等等。类型是指答案是需要总结还是需要一个列表，比如：列举出 10 个最大的城市，就需要列出符合要求的规定数量的答案。

以问句 *Q4:What is the name of the volcano that destroyed the ancient city of Pompeii?* 为例，图 1 给出了 *Q4* 的话语角色信息：意图说明了 *Q4* 的用户是向系统获取答案的；内容说明了 *Q4* 的主题是 *destroy* 这个活动并且有两个参与者，一个是作为 Agent 的 *volcano*，另一个为作为 Theme 的 *Pompeii*；Q1 的中心为 *destroy* 这个活动的参与者 1 这个实体的名字；最后问题以文本的形式返回答案。

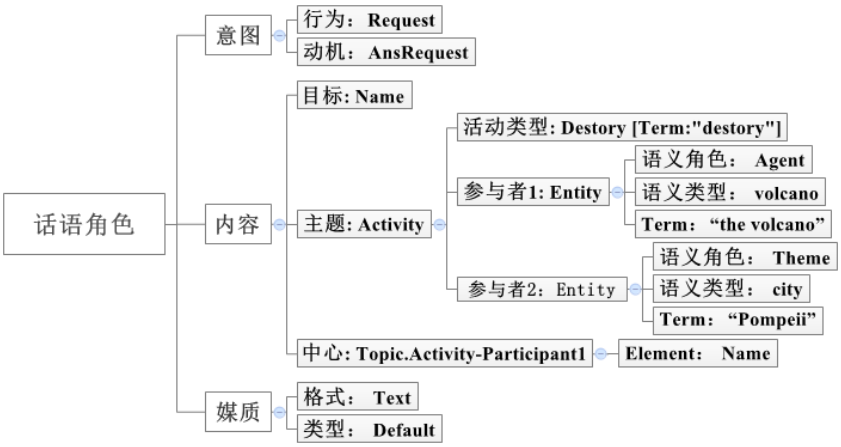

图 1　Q1 的话语角色信息图

划分话语角色的粒度可以是多样性的。划分的越好越细致，上下文信息的利用率就越高。但是，越好的划分需要越深入的语义识别处理。这种富语义的表达形式能够更好地帮助理解交互式问答中的上下文信息。

4.2 话语转换

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，所以，很大程度上各个问题间是如何联系在一起的与问题的主题是如何演化的非常相关。因此，围绕问题的主题，可以把话语的转换分为以下三类：

1. 话题的延伸

下一个问题的主题与上一个问题的主题相同，但是参与者或者一些其他方面有所不同，这种情况又可以细分为：

1）约束变化

下一个问题与上一个问题有着相同的主题，但有着不同或修改过的约束条件。例如：

Q5: What's the crime rate in Maryland and Virginia ?

Q6: What is it ten years ago?

Q5，*Q6* 有相同的主题 crime rate，约束条件不同，*Q5* 是 *Maryland and Virginia* 的 *crime rate*，而 *Q6* 是 *ten years ago* 的 *crime rate*，作为 Q5 约束的一个修改。再比如：

Q7: What's the crime rate in Maryland and Virginia?

Q8: What was it in Alabama and Florida?

Q7 和 *Q8* 同样有着相同的主题，但是两者的约束条件不同，*Q7* 与 *Q8* 询问的两处不同地方的犯罪率。

2）参与者转移

下一个问题与上一个问题有着相同的主题，但两个问题的参与者不相同，例如：

Q9: In what country did the game of croquet originate?

Q10: What about soccer?

在这个例子中，两个问题的主题都是 *originate*，但是两者的参与者不同，*Q9* 为 *croquet*，*Q10* 为 *soccer*。

2. 话题的扩展

两个相连的问题有着相同的主题，但是两个问题的中心不同。比如：

Q11: What is the name of the volcano that destroyed the ancient city of Pompeii?

Q12: When did this happen?

在这个例子中，*Q11* 和 *Q12* 具有相同的主题，但两者关注主题的方面不同。*Q11* 是询问主题 destroy 的参与者之一的名字，而 *Q12* 是询问主题的发生时间。

3. 话题的转移

两个相连的问题也可能是关于两个不同的主题。根据两个问题间不同的语义关系，话题的转移可以分为两类：

1）活动主题转移为另一个活动主题

例如在下面这个例子中：

Q13: What is the name of the volcano that destroyed the ancient city of Pompeii?

Q14: How many people were killed?

两个问题的主题都是活动，但是是不相同的活动，*Q13* 为 *destroy* 而 *Q14* 为 *kill*。

2）活动主题转移为实体主题

例如：

Q15: What is the name of the volcano that destroyed the ancient city of Pompeii?

Q16: How tall is this volcano?

上述例子中，*Q15* 的主题为 *destroy*，而 *Q16* 为实体 *volcano*。这样的转移中包含了可以作为指代消解的依据，因为 *Q16* 中的实体是 *Q15* 活动中的参与者之一。

5 话语结构特征

话语结构能够很好地表示交互式问答中的上下文信息以及单个问题中主要部分间的关

系。因此本文根据交互式问答中的结构特征提出了描述部分话语结构的特征，如表 2 所示。统计语料发现，先行语和照应语作为话语结构中参与者实体的概率很高，同时两者作为问题中的主题的情况也十分常见。因此，本文在消解平台中加入了标注好的实体与主题特征。在交互式问答中，不同位置的问题所含的未知信息量不同，因此我们也加入了问题在这个问答中的位置信息特征。将这些新特征加入到交互式问答基准平台中，进行消解实验，测试改进后的新平台的性能。

表 2 话语结构特征集

ANSentenceid	问题总数减去照应语所在问题的 ID
CASentenceid	问题总数减去先行语所在问题的 ID
ANEntity	照应语是问题中的实体取 1，否则取 0
CAEntity	先行语是问题中的实体取 1，否则取 0
ANTarget	照应语是问题的主题取 1，否则取 0
CATarget	先行语是问题的主题取 1，否则取 0
ANCAEntity	照应语和先行语同为实体取 1，否则取 0

6 实验结果与分析

6.1 实验设置

本文采用 TREC2004 至 TREC2007 的 QA 评测任务的 286 个问题集 1962 个问题作为实验语料[2]，在语料上标注了指代关系链，以及话语结构信息。

在使用机器学习方法的步骤中，本文采用 SVM-Light 工具中径向基核函数（RBF）来进行训练与测试指代消解平台的性能，选取 60%的语料进行训练，20%的语料作为开发集，其余 20%作为测试语料。整个实验流程如下：

首先，去除语料中标注的标签，生成生语料；对语料进行词性标注、命名实体识别、句法分析等预处理工作。在预处理的结果上提取出名词列表，并两两组成训练实例对，根据单复数、性别、语义类别等规则进行实例对过滤，去除不可能具有指代关系的实例对；随后，根据设定的特征空间，结合从标注样本中提取的话语结构信息，获取实例对的各特征向量值，并结合指代链标注信息确定实例对间的正负关系，形成训练文件；将训练文件提交给相应的分类器算法，训练生成分类器模型。

生成分类器模型之后，再去除测试语料中的所有标注信息，形成生语料；与训练时类似，对生语料进行预处理工作；将所有名词性短语两两组对，根据设定的规则进行实例对过滤，去除不可能具有指代关系的实例对；根据设定好的特征空间获取各特征的值；将实例对的特征向量提交给分类器，分类器使用训练时生成的分类器模型进行分类，并返回分类结果。

最后将分类结果与标注好的指代关系进行比对，利用相应的评测算法计算得出准确率、召回率和 F 值。

6.2 实验结果及其分析

表 3 给出了基准系统在 ACE2003 NWIRE 语料上的实验结果，表 4 给出了基准系统与改进后的系统在交互式问答文本中指代消解的实验结果。

表 3 基准系统在 ACE2003NWIRE 上的实验结果

	准确率	召回率	F 值
总体性能	64.5%	51.9%	57.5%

表 4 基准系统和改进系统在交互式问答文本中的实验结果

词类别	基准系统			改进系统		
	准确率	召回率	F 值	准确率	召回率	F 值
总体性能	50.12%	66.54%	57.17%	52.94%	68.50%	59.72%
代词	91.24%	23.81%	37.76%	91.24%	23.81%	37.76%
有定名词	52.60%	55.31%	53.92%	56.67%	52.72%	54.63%
专有名词	52.73%	58.86%	55.63%	52.86%	62.99%	57.48%

从表中可以看出：

1）对比表 3 与表 4，基准系统在交互式问答文本上的总体性能良好，F 值和新闻文本相比相差不大，说明基准系统的特征集在交互式问答文本中的适应性良好。但基准系统在交互式问答文本中的代词消解性能不佳，召回率很低，导致 F 值很低，说明交式问答文本中代词的情况和新闻文本中代词的情况有较大的差异性。

2）加入了话语结构特征以后，系统总体性能的各项指标都有所上升，准确率上升了 2.8%，召回率上升了 2%，F 值上升了 2.6%，由此说明话语结构特征的加入对于提高交互式问答中的指代消解性能作用明显。

3）进一步分析改进系统对不同类别词的指代消解性能结果显示：对于代词，各项性能指标均无显著的变化，说明引入了话语结构信息对于代词的消解没有太大帮助。改进系统对有定名词和专有名词的消解性能均有提高，两者的 F 值均有提高。有定名词上的准确率上升明显，但召回率略有下降；专有名词上的准确率和召回率均有不同程度地提高。

7 小结与展望

本文将现有的新闻文本的指代消解平台运用到交互式问答中，观察方法的适应性，并在此基础上提出了基于话语结构信息的特征集。通过实验后发现，原基准系统在交互式问答中的性能与在新闻文本上的总体性能差异不大，在加入了本文提出的话语结构信息特征后，改进系统的指代消解性能有了显著的提高，F 值上升了 2.6%。但基准系统和改进系统对于交互式问答文本中的代词的指代消解性能表现都不好，因此未来我们将进一步改进现有系统，对交互式问答中代词进行更加深入细致的研究。

参 考 文 献

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

[2] Ellen M. Voorhees. Overview of the TREC 2004 Question Answering Track. http:/ /trec.nist.gov/, 2004.

[3] Chai J, Jin R. Discourse structure for context question answering[C]. HLT-NAACL 2004 Workshop on Pragmatics of Question Answering, 2004:23–30.

[4] Dongsheng Wang. Answering contextual questions based on ontologies and question templates[J]. Front. Comput. Sci. China, 2011, 5(4): 405–418.

[5] Joyce Y. Chai, Rong Jing. Discourse structure for context question answering[C]. Proc of of the Workshop on Pragmatics of Question Answering at HLT-NAACL, 2004:23–30.

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

[7] Nils D, Jonsson A. Empirical studies of discourse representations for natural language interfaces[C]. Proc of 4th Conference of the European Chapter of the ACL, 1989:291–298.

[8] Dagan I. and Itai A. 1990. Automatic Processing of Large Corpora for the Resolution of Anaphora References. ACL’1990:330-332.

[9] Ge N.Y., Hale J. and Charniak B. 1998. A statistical approach to anaphora resolution. VLC’1998: 161-170.

[10] Cardie C. and Wagstaff K. 1999. Noun phrase coreference as clustering. EMNLP’1999: 82- 89.

[11] McCarthy and Lehnert W. 1995. Using Decision Trees for Coreference Resolution. In Proceedings of the Sixth Message Understanding Conference (MUC-6).

[12] Soon W M, Ng H T, Lim D. A machine learning approach of coreference resolution of noun phrase[J]. Computational Linguistics, 2001, 27(4): 521-544.

[13] Ng V. and Cardie C. 2002. Improving machine learning approaches to coreference resolution. ACL’2002: 104-111.

[14] Harabagiu, S., Pasca, M., and Maiorano, S. Experiments with Open-domain Textual Question Answering. In

Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000), 2000.

[15] Grosz, B. J. and Sidner, C. 1986. Attention, intention, and the structure of discourse. Computational Linguistics , 12(3):175-204. 1986.