面向开放的限定领域的交互式问答语料分析*

张耀允，王晓龙，王 轩，徐睿峰，侯永帅，范士喜
哈尔滨工业大学 深圳研究生院 智能计算研究中心，深圳 518055
E-mail: zhangyy@cs.hitsz.edu.cn; wangxl, wangxuan@insun.hit.edu.cn; xuruifeng,houyongshuai@hitsz.edu.cn

摘 要：交互式问答是国际问答技术领域新兴的热门研究方向。它结合自动问答与对话系统技术，可以处理系列相关问题，并能与用户进行对话式交互，但是目前在中文问答领域开展的相关研究还比较少，尤其缺乏对真实环境中大规模交互式问答语料的收集和分析。本文收集了面向开放的限定领域的中文交互式问答语料，首次对其中的语言现象进行了详细统计，并着重在对话结构、话题过渡的方式、上下文相关现象以及它们之间的关系等方面进行了分析和讨论。本文的统计结果和发现对于研究交互式问答的对话模型设计和上下文相关话语的自动理解算法，具有重要的意义。

关键词：交互式问答；对话结构；话题过渡；上下文信息；语料统计；开放的限定领域

Analysis of Interactive Question Answering Corpus in Open-ended Restricted Domain

Zhang Yaoyun, Wang Xiaolong, Wang Xuan, Xu Ruifeng, Hou Yongshuai, Fan Shixi
Intelligent Computing Research Center, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055
E-mail: zhangyy@cs.hitsz.edu.cn; wangxl, wangxuan@insun.hit.edu.cn; xuruifeng,houyongshuai@hitsz.edu.cn

Abstract: Interactive question answering is an emerging hot research topic of international question answering field. It combines technology of automatic question answering and dialogue system, and can handel series of related questions in a dialogue-like way. However, relevant research in this direction is scarce in China. Especially, we are lack of a corpus collected in practical environment and statistical analysis based on it. This paper collect a Chinese interactive question answering corpus in open-ended restricted domain, and analyze the language phenomena in detail, focusing on the discourse structure, topic transition methods, the contextual phenomena and their relations. The results and findings in this paper are significant to research of discourse model design and automatic interpretation of contextual-aware dialogue in interactive question answering.

Keywords: interactive question answering; discourse structure; topic transition; contextual information; corpus statistics; open-ended restricted domain

1 引言

交互式问答(interactive QA) 是国际问答技术领域新兴的热门研究方向。它一方面保持了自动问答系统的开放性，从海量数据中检索并生成答案，同时借鉴了对话系统的交互模式，可以与用户进行对话式的、连续相关的问答[1]，例如：

Q1: 支付宝是什么？
A1: 支付宝是一种“第三方担保交易模式”，由阿里巴巴创办。
Q2: 它和网银一样吗？
A2: 两者是有区别的。
Q3: 充进去的钱可以拿出来吗？
A3: 不可以直接提现……

因此，交互式问答首要需要解决的问题是判断当前问句和前述问句及答案之间的相关性，并建立相应的对话模型，以及对承前省略上下文信息的问句进行恢复和理解。要解决这些问题，就

* 863计划（2006AA01Z197）资助项目，国家自然科学基金（60435020）资助项目。

要对交互式问答语料中的语言现象，尤其是保持对话连贯性的话题过渡(transition)方式和上下文相关现象进行归纳和分析。

在国际评测组织 TREC 和 NTCIR 在 2004-2007 年间[2][3]分别举办的自动问答评测中，所提供的语料均是成组的问句系列，每组问句围绕一个话题展开。参赛队伍首先要对问句中省略的上下文信息进行恢复处理，才能检索答案。但是评测问句都是人工做好的规则问句，并且以事实类的简单问句为主。真实问答环境中存在的用户表述复杂化、口语化现象，具有复杂信息需求的问句，以及当前问句与前述答案之间上下文依赖关系都没有涉及。Chai 和 Jin[4]提出了问句话题过渡的几种类型，但他们的研究仅止于理论模型，并且仍然没有涉及答案在话题过渡中的作用。Bertomeu 等人[5]使用 wizard-of-oz 的方法，即由真人来完成问答系统后台的问句理解和答案检索功能，收集了基于特定任务的和限定领域的数据库的交互式问答语料，并进行了详细的分析，但是其任务和数据源的限制都和传统的对话系统类似，却没有体现出自动问答系统中用户信息需求和检索源的开放性。Schooten 和 Akker[6]也采用了 wizard-of-oz 的方法，他们将连续问答的过程简化为只收集第一个问句，对应的答案以及第二个问句作为语料，并对语料中的话语现象进行了归纳和分类。但是这种收集语料的方式不能覆盖长距离的上下文相关现象，也不能客观地体现真实连续对话中话题过渡的关系。

目前研究中文交互式问答的工作还比较少，伍大勇等人[7]将 TREC 的评测语料翻译成中文使用，并采用了和[6]相同的方法收集语料。

本文同样采用 wizard-of-oz 方法，收集了面向开放的限定领域的中文交互式问答语料。实验对用户进行问答要完成的信息任务没有做事先的限制，答案则是由承担系统角色的真人根据搜索引擎百度或谷歌的检索结果手工生成的。在语料的基础上，本文对交互式问答过程中的语言现象进行了详细统计，并着重在对话结构、话题过渡的方式、上下文相关现象以及它们之间的关系等方面进行了分析和讨论。

2 语料收集环境和标注说明

2.1 语料收集环境配置和要求

本实验对金融和计算机两个限定领域的交互式问答语料进行了收集。我们同样采用了 wizard-of-oz 方法，20 个计算机专业的硕士和博士参与了语料的收集，其中 2 个人为一组，交替充当交互式问答的前端用户和系统的角色。实验为每个领域提供了 50 个话题，每个用户从中选择 8~12 个话题，并对每个话题提问不超过 15 句。用户在前端使用的问答界面和真实的系统完全相同。对于选择的话题，用户可以从任意的角度进行提问。如果在对话过程中出现了感兴趣的新话题，用户也可以自由转换话题进行后续提问。系统首先将出现信息缺省的问句补充完整，然后提交给搜索引擎百度或谷歌，根据检索结果生成答案返回给前端。

除正常的提问外，用户如果对当前给出的答案不满意，或者还需要补充，可以直接提出要求。例如：

Q4: 卡巴斯基的软件分哪些类型，服务器和客户端是怎么回事

A4: 卡巴斯基实验室创新性地将反病毒软件所面临的软件分为了三类：恶意程序，正常程序和未知程序。服务器负责客户端软件的更新、维护，尤其是病毒库的更新。

Q5: 我想问的是，它的软件是如何分发的

另一方面，如果用户提问具有歧义性，或者内容过于宽泛，则系统可以对用户进行询问。例如：

Q6: Photoshop 软件是哪一年推出的?

A6: 你是指第一个版本吗

Q7: 是的

因此，交互式问答过程中的对话不仅包括正常的问句和答案，还包括用户和系统之间澄清性的问题和应答(clarification question&response)。

2.2 语料收集和标注情况

我们最终收集得到交互式问答语料 218 组[1]，共 2080 句，1040 轮对话。平均每个用户选择了约 11 个话题。共涉及 81 个话题，其中金融领域涉及 43 个话题，共出现 125/人次；计算机领域涉及 38 个话题，共出现 93/人次。

除了对语料中的对话平均长度、问句类型等基本信息的统计和分析外，实验还对以下几个方面进行了手工的标注：

2.2.1 连续问答过程中的话题过渡关系

Chai 和 Jin[4]定义了连续问句之间话题过渡的三种类型，即话题扩展(Extension)，话题探索(Exploration)和话题转换(Shift)，但是当前问句话题和前述答案包含信息的过渡关系却没有定义。本文在上述三种类型的基础上进行扩展，增加了答案在话题过渡中的功能，详细的类别描述如下：

（1）话题扩展：当前问句和前述某一问句的话题相似，但是话题的参与者或者限制条件发生改变，例如在 Q9 中，Q8 里话题的事件没有改变，但是参与者却改变了：

Q8: 怎样快速提高 QQ 等级?

Q9: QQ 农场呢?

（2）话题探索：当前问句和前述某一问句的话题相同，但询问的是该话题不同方面的信息，即问句的焦点改变了，例如 Q11：

Q10: 和讯网的网址是什么？

Q11: 它的盈利模式是什么？

（3）话题转换：即当前问句和前述某一问句询问的是两个不同的话题，例如 Q12 到 Q13：

Q12: windows7 售价是多少？

Q13: 为什么自带的播放器打不开 mkv 视频?

（4）答案做话题：当前问句的话题或者话题的参与者是前述某一个答案中的信息，可以分为三种子类型：1. 答案本身做话题，例如：A14 到 Q15；2. 答案中出现的事件做话题，例如 A16 到 Q17；3. 答案中出现的实体做话题，例如：A18 到 Q19。

Q14: 那么人民币升值会对别国的经济有什么影响呢?

A14: 在中国同世界各国的经贸联系日益紧密的今天，人民币升值不仅不利于我国经济的发展，也会……

Q15: 对你提到的情况我国政府该如何应对?

Q16: 雅虎中国和阿里巴巴是什么关系？

A16: 雅虎中国被阿里巴巴收购了。

Q17: 当时收购花了多少钱啊？

Q18: 我是在校大学生，适合用哪种卡呢？

A18: 在校大学生可以申请交通银行 Y-POWER 信用卡……

Q19: 这个对大学生有什么优惠吗？

[1] 本文语料的下载网址为 http://qa.haitianyuan.com/IQACorpus.html/。

2.2.2 对话语句的功能

如 2.1 节所述，交互式问答过程中的对话不仅包括正常的问句和答案，还包括用户和系统之间澄清性的问题和应答，因此本文对这四种对话语句的功能分别进行了标注。

2.2.3 上下文相关现象

上下文相关现象表示当前问句与上下文信息，即前述问句和答案中的信息，在语义上的关联和依赖关系。本文将上下文相关现象分为以下几类：

（1）问句语义完整：1. 完整和上下文无关，一般是对话中的第一个问句；2. 完整且和前述问句相关；3. 完整且和前述答案相关。

（2）问句中有指代现象：1. 指称性代词，例如他，他们，它，其、者等；2. 指示性代词，例如这、这些、那个等；3. 指示性名词/名词词组，例如 Q21 中的“这个表”。

Q20:买股票时要关注资产负债表里的什么内容

Q21:公司会对这个表的数据作假吗

（3）问句中有省略现象：1. 名词性省略，即省略一个名词或名词词组，例如 Q3；2. 动词性省略，即省略一个动作或事件，例如 Q17。

（4）问句是一个片段：问句仅有限定成分或者名词，有些时候结构和前述的问句相似，补全时可以把前述问句的内容直接连接上，或者做一些替换，例如 Q9。

另外，问句中的指代现象按照先行词(antecedents)的位置可以分为句间指代和句内指代。句间指代和省略的先行词又可能出现在前述问句或者答案里。

同时，某些连词、语气词或者表述方式可以起到衔接上下文的作用，例如“那么”，“哦”，“嗯”，“这样啊”，“你刚刚提到”等等。我们也对这些现象进行了标注和统计。

3 语料统计和分析

3.1 语料基本信息统计

表 1 交互式问答语料基本信息表

	子句个数		问句长度			答案长度			每组对话长度			事实类问句	复杂问句
	1	2	AVG	MIN	MAX	AVG	MIN	MAX	AVG	MIN	MAX		
金融	596	8	8.0	1	28	42.7	1	307	4.8	3	12	178	417
计算机	424	12	8.2	1	33	35.6	1	308	4.7	2	12	129	295
合集	1020	20	8.1	1	33	39.5	1	308	4.8	2	12	307	712

表 1 第二列中的子句是指用户提交的一个问句中有单句。单句之间以问号、句号和叹号为分隔符。语料中有 20 个问句包含 2 个单句，例如：

Q22: 美元指数到底有什么用？知道美元指数我们就能知道什么东西呢？

问句长度最短只有 1 个词，这种情况是用户对系统提出的澄清性问句的应答。答案的平均长度在 30 个词以上，说明真实环境的交互式问答中答案以段落为主。对话长度以问答的轮数计算，最短的只有 2 轮，最长则达到了 12 轮。通过对以命名实体为信息需求的事实类问句和询问原因、过程、对比等复杂信息需求的问句进行统计，可以发现面向开放的限定领域的交互式问答中复杂的信息需求比例占 2/3 以上，这个发现说明仅仅研究由事实类的简单问句组成的对话是有局限性的。

除此之外，观察发现语料中存在大量口语化的表述，比如“哦”、“嗯”等语气词，以及错字、

漏字等拼写错误，其中最突出的就是指称代词“他”和“它”的混淆，例如：

Q23: 那么新浪和网易比较，他有什么优势吗？

3.2 话题过渡类型分析

表2 话题过渡类型统计

	问句—问句			答案—问句		
	话题扩展	话题探索	话题转换	答案做话题	事件做话题	实体做话题
金融	11	54	368	3	10	33
（465）	2.30%	11.27%	76.83%	0.63%	2.09%	6.89%
计算机	14	63	262	1	6	20
（357）	3.83%	17.21%	71.58%	0.27%	1.64%	5.46%
合集	25	117	630	4	16	53
（822）	3.04%	14.23%	76.64%	0.49%	1.95%	6.45%

如表2所示，话题转换是话题过渡类型中最常见的，占后续问句总数的76.64%。其次是话题探索和答案中的实体做话题。对于话题转换，最常见的形式是前述问句的实体型话题做当前问句的事件型话题的定语或参与者，例如Q25:

Q24: 什么是新股？

Q25:一般都会涨吗？

值得一提的是，有23个问句同时被标注为话题转换和答案事件/实体做话题。也就是说，这些问句同时与前述问句和答案相关，比较普遍的现象是答案中的事件或实体做当前问句的话题，但问句中缺省的定语或者状语则需要前述问句中的信息来补全。因此，它同样是和前述问句被约束在同一个宏观的话题的上下文语境里的。例如在Q27中，A26中的实体作为话题，但是缺省了Q26中的信息作为对话题的限定。

Q26: Thinkpad哪个型号好用？

A26: 给你推荐X2系列……

Q27: X2系列最高配置在什么价位？

3.3 对话语句功能分布

如2.1节的描述，语料集中共包括29对澄清性的问题和应答，它们的分布如表3所示。尽管在功能上和正常的问答做了区分，但是对于可以放在上下文的语境中进行分析的部分澄清性问题和应答，本文都对它们的话题过渡方式和上下文相关现象进行了标注。

表3 交互式问答中对话语句的功能分布

	问句&答案对	澄清性问题&应答对
金融	586	18
计算机	425	11
合集	1011	29

3.4 上下文相关现象统计

如表4所示，上下文信息缺省的现象占用户对话总数的45.43%的，如此高的比例进一步印证了上下文相关信息识别的重要性。最频繁的缺省现象是指称性代词和名词性省略。表格中列出的指代现象都是句间指代，句内的指代一共有14个,有12个是指称性代词，2个指示性名词。

表 4　交互式问答的上下文相关现象分布

	完整			指代			省略		片段	衔接性表述
	上下文无关	问句相关	答案相关	指称性代词	指示性代词	指示性名词	名词性省略	动词性省略		
金融	125 22.98%	192 35.29%	10 1.84%	83 15.26%	4 0.74%	25 4.60%	81 14.89%	9 1.65%	15 2.76%	112 20.59%
计算机	93 19.38%	134 27.92%	17 3.54%	95 19.79%	1 0.21%	12 2.50%	103 21.46%	10 2.08%	15 3.13%	67 13.96%
合集	218 20.96%	326 31.35%	27 2.60%	**178** **17.40%**	5 0.48%	37 3.56%	**184** **17.98%**	19 2.02%	30 3.08%	**179** **17.21%**

由连词、语气词或者其他表述方式组成的衔接性表述现象比例高达 17.21%，可以作为判断前后问句之间是否相关的有效特征。

表 5　当前问句与上下文相关的前述问句或答案之间的距离分布

距离	1	2	3	4	5
问句个数	796	14	6	5	1

表 5 中的距离是以当前问句和与其相关的问句或答案之间相隔对话轮数来定义的。长距离的上下文相关共有 26 个，这种现象的出现主要是因为用户会对答案中出现的新实体或事件进行提问，即出现了话题嵌套的现象。例如 A28 和 Q31 相隔 3 轮对话，这是由于用户对答案中出现的多个的实体逐一进行提问。

Q28: 公认的蓝筹股有哪些?

A28: 长江电力 600900、中国石化 600028、中国联通(600050)，宝钢股份 600019 等

Q29: 长江电力怎么样

Q30 近期会不会涨?

Q31: 中国石化现价如何?

3.5　话题过渡方式与上下文缺省现象的关系

表 6　话题过渡方式和上下文缺省现象的共现频率

	问句—问句			答案—问句		
	话题扩展	话题探索	话题转换	答案做话题	事件做话题	实体做话题
指称性代词	1	**20**	**149**	0	2	**6**
指示性代词	0	0	2	1	0	2
指示性名词	0	4	26	**3**	1	3
名词性省略	**5**	**22**	**131**	0	**8**	**17**
动词性省略	3	2	11	0	**3**	0
片段	**4**	2	18	0	2	3

从表 6 中我们可以看出，不同的话题过渡方式对应的上下文缺省现象具有不同的特点。话题探索、话题转换和答案实体做话题和指称性代词、以及名词性省略共现的频率高；话题扩展和名词性省略，以及片段的共现频率高；答案本身做话题和指示性代词共现的频率高；而答案事件做话题则和省略现象共现的频率高。

4 结论与讨论

本文收集了面向开放的限定领域的中文交互式问答语料，首次对其中的语言现象进行了详细统计，并着重在对话结构、话题过渡的方式、上下文相关现象以及它们之间的关系等方面进行了分析。基于前文的分析和统计，我们可以得出以下几点结论：

1. 尽管在话题过渡方式中，话题转换和答案信息做话题的比例比较高，达到82.39%。但在这两种现象中，当前问句和前述问句仍然处于由一个宏观的话题约束的上下文环境中，这个宏观的话题主要是由整组对话的第一个问句引入的。因此，我们计算了后续问句和第一个问句相关的个数。其中，相关的后续问句个数为604，占后续问句总数的73.48%；存在信息缺省的问句个数为261，占信息缺省问句总数的55.29%。因此，在判断前后问句相关性和对问句中的缺省信息进行识别时，可以将第一个问句中的话题作为基准线来参考和对比[8]。

2. 在上下文相关现象中，最普遍的信息缺省现象是指称性代词和名词性省略，这两种现象的先行词都是命名实体，领域专有名词，或名词性的词组，因此这三种实体的识别是生成先行词候选集合，进而对问句补全的关键。另一个方面，领域知识，尤其是专有名词和属性，以及专有名词和事件之间关系的收集，则是从多个候选实体中识别出正确的先行词的关键。

3. 后续问句理解中的两个难点是前述答案到当前问句话题过渡的识别和远距离上下文相关现象的识别。一方面，真实语料中的答案比较长，里面包含多个实体和事件，在很大程度上增加了候选先行词的范围。另一方面，统计显示，远距离上下文相关现象的比例仅占后续问句总数的3.16%，但是由于其后续问句的理解要基于当前问句对上下文信息的正确解析，不加处理有可能会产生错误的积累和传递。因此，研究高效而经济的对话模型的表示方法和研究解决这种现象的算法和策略是有必要的。

参 考 文 献

[1] N. Webb, B. Webber. Special issue on interactive question answering: introduction. Natural Language Engineering, 2008, 15(1): 1-8.

[2] Hoa Trang Dang, Diane K, and Jimmy L. Overview of the TREC 2007 question answering track. In: Proceedings of the Sixteenth Text REtrieval Conference. Gaithersburg, USA: NIST Special Publication, 2008.

[3] Tsuneaki K, Jun'ichi F, Fumito M, and Noriko K. Handling information access dialogue through QA technologies—A novel challenge for open-domain question answering. In: Proceedings of the HLTNAACL Workshop on Pragmatics of Question Answering. Boston, USA: ACL, 2004. 70-77.

[4] Joyce Y. Chai and Rong Jin. 2004. Discourse structure for context question answering. In Proceedings of the HLT-NAACL 2004 Workshop on Pragmatics in Question Answering, Boston, MA.

[5] N´uria Bertomeu, Hans Uszkoreit, Anette Frank, Hans-Ulrich Krieger and Brigitte J¨ org. Contextual phenomena and thematic relations in database QA dialogues:results from a Wizard-of-Oz Experiment. In: Proceedings of the HLT-NAACL Workshop on Interactive Question Answering. New York, USA: ACL, 2006. 1-8.

[6] Boris van Schooten and Rieks op den Akker. Follow-up utterances in QA dialogue. Traitement Automatique des Langues, 2005, 46(3): 181-206.

[7] 伍大勇, 张宇, 刘挺. 交互式问答用户问题相关检测研究. 中文信息学报. 2010, 24(3): 11-18.

[8] B. w. Van schooten, R. Op den akker, S. Rosset, O. Galibert, A. Max, G. Illouz. Follow-up question handling in the imix and ritel systems: A comparative study. Special issue on Interactive Question answering, Natural Language Engineering, 2008, 15(1): 119-141.