面向非任务型对话系统的人工标注中文数据集

李菁，张海松，宋彦
（腾讯 AI Lab，广东省 深圳市 518052）

摘 要：本文针对非任务导向型对话的回复质量构建了一个大规模的人工标注中文数据集，该数据集包含了从社交媒体收集到的超过 2 万 7 千个对话问题以及超过 8 万 2 千个对话问题的回复[1]。为了产生高质量的标注数据，我们邀请了专业人员根据对话回复的相关性、连贯性、信息性、趣味性、以及是否潜在地具有让对话继续延续的特性进行标注，在标注中我们定义了一个五级评分方法，分别是：极差的、较差的、一般的、较好的、极好的。为了测试标注产生的数据集是否具有有效性和实用性，我们以对话回复选择为任务，在我们的标注数据集上测试了多种无监督和有监督模型。实验结果表明，该数据集对于提升对话回复选择的质量有显著效果。

关键词：对话系统；人工标注；中文数据集

A Chinese Corpus for Non-task-oriented Dialogue Systems with Five-grade Manual Annotations

Jing Li, Haisong Zhang, Yan Song
(Tencent AI Lab, Shenzhen, Guangdong, 518052)

Abstract: This paper presents a large-scale corpus for non-task-oriented dialogue systems, which contains over 27K distinct prompts with more than 82K responses collected from social media. To annotate this corpus, we define a 5-grade rating scheme: bad, mediocre, acceptable, good, and excellent, according to the relevance, coherence, informativeness, interestingness, and the potential to move a conversation forward. To test the validity and usefulness of the produced corpus, we compare various unsupervised and supervised models for response selection. Experimental results confirm that the proposed corpus is helpful in training response selection models.

Keywords: dialogue system; manual annotation; Chinese corpus

1 引言

自从图灵测试构想的诞生[1]以来，构建能够自然地与人类沟通的交互系统便成为了人工智能的使命之一，尤其是人机交互的前端——自然语言对话系统更是被期待着承担连接人与机器的重任。近年来，随着人工智能相关技术的突破以及大量真实对话数据的产生，对话系统的相关研究呈现出巨大发展。很多在真实应用中涌现出的对话系统不但成为了我们日常生活中的必备工具，更吸引了学术界的广泛关注，例如：苹果 Siri[2]、谷歌 smart reply[3]、微软小冰[4]等等。总体来说，在功能层面上，现有的对话系统可以分为两大类，即任务导向

[1] 本文提出的数据集在此处公布：
http://ai.tencent.com/ailab/upload/PapersUploads/A_Manually_Annotated_Chinese_Corpus_for_Non-task-oriented_Dialogue_System

型对话系统[5]和非任务导向型聊天机器人[6]。其中，任务导向型对话系统对对话应用的任务场景做了一定程度的限制，旨在帮助人们完成特定任务，例如订票交互系统帮助用户购买机票的服务[7]、图书馆交互系统回答用户关于图书信息的咨询[8]等。相比之下，非任务导向的聊天机器人更加侧重闲聊功能，这种类型的对话系统不会对对话场景和主题做任何限制，因此聊天主题相对更多样化，话题覆盖程度比任务型对话系统更加广泛。

以往的人机交互系统，例如 Eliza[9]、Parry[10]和 Alice[11]都使用基于规则和模板的方法。这类方法在早期的对话系统中十分流行，然而，规则和模版的设计需要消耗大量人力，很难覆盖多样化的对话主题。在现今的对话系统中，应用最广的是数据驱动型对话系统，这种对话系统不依赖于人总结的规则，完全从数据中学习如何回复用户的问题，可以在很大程度上缓解规则系统所需要的人力和资源[12]。然而，训练一个数据驱动型对话系统往往需要大量的对话数据。为了解决这种数据需求，以往的工作倾向于从社交媒体中收集用户产生的交互文本[2]用于训练对话系统[13]，原因有如下几个方面：其一，社交媒体的数据完全公开，易于收集和获取；其二，社交媒体对文本的长度做了限制，例如新浪微博[3]上单条信息的长度不超过 140 个字，这些文本的长度比较接近对话中的文本长度，因此比较适合被用于学习对话回复；其三，社交媒体的文本往往产生自不同的人，天然地构成了对话形态的文本组织方式；其四，社交媒体的语言风格紧跟潮流，能够比较与时俱进地反应当前的语言使用现状。

然而，通过社交媒体收集的数据也会直接受到社交媒体平台带来的负面影响，包括诸如信息噪声大（包含广告等）、不符合规范、有效信息量小等问题。例如，在新浪微博上，针对用户的对话问题：“我超爱吃苹果!!!”，我们在表 1 中展示了几个用户回复的样例。其中第一个回复是一个针对 RPG 游戏的广告而不是直接回复原始微博的问题，这在社交媒体中广泛存在，属于噪音数据，这样的回复与问题完全无关，属于偏离了主题的极差回复类型。第二个回复虽然包含问题中的关键词“苹果”，但是与问题的配合看来显得并不通顺连贯，属于较差的回复类型。第三个回复虽然通顺自然，却属于在社交媒体上广泛存在的一类“万能回复”[14]，可以应对多种不同类型的问题，因此在内容上针对特定问题并不具备多少信息量[4]。第四个回复被认是极好的回复，因为其不仅主题相关、自然连贯地回复问题，且其中包含的俗语表达提供了“苹果有益身体健康”的丰富信息，还具备一定程度的趣味性。上述实例表明，不同回复的质量很大程度上决定了一个对话进程的持续能力和用户体验。因此，对话系统需要有效区分不同质量的回复。尤其对于数据驱动型的对话系统，回复数据的标准化质量标注显得非常重要，可以有效助益对话系统的回复生成[15]能力和效果评估[16][17]。然而，目前相关研究有限，并且高质量有效标注的语料较为稀缺，在中文对话领域基本没有类似的工作发表，比较明显地阻碍了该领域相应工作的推进。

为了完善当前对话系统研究，并且为学界提供有效的公开标注数据，在本文所述的工作中，我们构建了一个大规模的人工标注对话数据集，其中包含超过 2 万 7 千个中文问题及其对应的 8 万 2 千条回复（每个问题可能对应多个回复）。本文从问题和回复的相关性、连贯性、信息性、趣味性等维度提出五级人工标注评分标准：极差的、较差的、一般的、较好的、极好的。考虑到多数以往工作主要集中使用未标注数据和自动标注数据，据我们了解，本文所述的工作是首次为非任务导向的对话系统构建人工标注中文数据集。同时，为了对比分析，在该数据集的基础上，我们使用不同的对话回复选择模型尝试了多组基础实验。实验结果表明本文提出的人工标注数据集可以有效驱动对话系统选择较高质量的回复。

---

[2] 在社交媒体等场景下，用户对其他公开用户发表的某些状态或者评论等进行相应的回复，以此产生的文本我们称之为交互文本。更一般地，任何用户相互之间进行交流产生的文本都可以被认为是交互文本。

[3] https://weibo.com

[4] 由于社交媒体上通用回复的普遍性，以往通过社交媒体语料训练的聊天机器人，往往倾向于生成类似的“万能回复”，妨碍聊天的正常进行。因此，通用回复与更高质量的回复需要被有效地区分。

表 1 新浪微博上的问题和它的样例回复及其对应标准分析

问题	我超爱吃苹果！！！	
样例回复	**评分等级**	**标准分析**
1、这个 RPG 游戏挺好玩的。URL。	极差的	偏离主题、广告
2、苹果 apple	较差的	包含关键词但不连贯
3、我也喜欢。	一般的	万能回复、信息量少
4、我也喜欢。一天一苹果，医生远离我。	极好的	富有信息量、趣味性

2 相关工作

本文与非任务型对话系统紧密相关。通常，非任务型对话系统可以分为两个不同类别：规则驱动型对话系统和数据驱动型对话系统。规则驱动型对话系统主要出现在对话系统研究的早期，利用人工制定的规则或模版来构建对话系统。核心的方法包括关键词匹配[18]、槽位填充（slot filling）[19]和模版填空[20]等。但是这类方法一方面需要耗费大量人力，另一方面在使用时也存在缺陷，主要原因是非任务型对话系统中，回复的可能性太多，以至于无法被有限的规则总结。

数据驱动型对话系统的蓬勃发展获益于在线数据的大量产生。当前，大规模人人对话数据已经易于获得，这在很大程度上推动了各类对话系统模型和算法的发展。这类对话系统主要利用机器学习的算法，通过引入少量的人工特征[13,21]、或者完全自动的特征学习[22,23]从真实的对话数据中学习类似人人交互方式的对话行为。数据驱动型对话系统不仅极大降低了对人力和资源的需求，而且相比于规则驱动型对话系统更能保证对话回复的多样性。

因此，为了保证数据驱动型对话系统的性能，收集和整理大规模、高质量的对话数据集变得尤为重要。以往的工作主要通过自动[23,24]或半自动[13,25]的方法构建数据集，保证这些方法有效的基本前提是收集的原始数据集已经拥有了比较高的质量。然而，由于社交媒体是当前对话数据集的主要来源[23]，其质量良莠不齐，因此引入人工标注提高数据质量十分重要。据我们了解，本文介绍的工作是第一个中文大规模人工标注对话数据集，有效填补了以往工作在非任务驱动型对话系统数据集方面的空白。

3 数据准备

3.1 数据收集

本文提出的数据集所包含的问题和答案对（简称问答对）收集自社交媒体上真实用户对话中的问题和回复，从包括百度贴吧[5]、百度知道[6]、豆瓣[7]、新浪微博等社交媒体站点通过网络爬虫进行收集。上述网站是中文社区较为流行的社交媒体平台，在这些平台上用户进行交互式讨论的主题具有多样性和高覆盖性等特点。这些数据的收集过程如下：首先，我们从各个平台的索引页面提取主题列表信息，例如：明星、娱乐、军事、体育、游戏等等[8]；接着，我们使用 JSsoup[9]工具抓取各主题页面并且对每个页面进行 HTML 解析，以此提取问题和对应回复的文字。

[5] https://tieba.baidu.com
[6] https://zhidao.baidu.com
[7] https://www.douban.com
[8] 我们从这些不同网站上抽取的主题列表具有比较高地相似性。
[9] https://jsoup.org/

3.2 数据整理

原始数据收集完毕之后，我们采取两个步骤进行接下来的数据预处理，以便于后续的人工标注工作。第一步进行敏感信息过滤，处理如脏话、成人内容、披露私隐等敏感数据。该操作的目的是避免任何使用本文提供的语料进行训练或者评估的聊天机器人产生使人不适的回复或者公开用户的私隐。第二步则聚焦于辨别和过滤带有知识依赖的问题。由于带有知识依赖的问题对应的答案有领域和场景的局限性，通常仅仅针对特定知识，所以很可能在对话过程中产生不适合当前条件的回答，例如：“今天北京天气如何？”、“明天皇马对利物浦的比赛几点开始？”等。因此为了避免在后续对话或评测中出现回复无法匹配场景的问题，我们需要将知识相关的问题和回复进行过滤。为了完成上述两步预处理，我们聘请了四位有经验的标注人员进行人工过滤。

表 2 对话问答对语料标注标准

等级 1（极差的）：回答本身毫无意义（例如[S1]）或者跟问题完全不相关（例如[S2]）。
等级 2（较差的）：回答与问题不连贯或者不一致但是提到了少量关键词（例如[S3]），或者回答与问题高度相似，几乎重复了一遍或者重复部分关键词（例如[S4]）。
等级 3（一般的）：回答是有意义的、相关的以及前后衔接流畅的，但是只能限制在特定时间或者空间上来使用（例如[S5]）或者比较通用的万能回复（例如[S6]）
等级 4（较好的）：回答是前后衔接的，覆盖了相关内容，但是比较简单明了，信息量不是很大以及回复无衍生内容，无更多想象空间（例如[S7]）。
等级 5（极好的）：回答不仅前后衔接以及内容相关，而且是有信息量的、有趣的、富含寓意的、主题相关的并能够产生更多对话（例如[S8]）

(a) 5 个评分等级 8 个类型的相关样例说明

问题：“嘿，北京，我来了！”
[S1]：<**等级 1**>(无意义)ddddd
[S2]：<**等级 1**>(不相关)也许吧？
[S3]：<**等级 2**>(不连贯)北京比香港大一些。
[S4]：<**等级 2**>(学舌)北京。
[S5]：<**等级 3**>(时空限制的)希望暖和一点！现在在下雪。
[S6]：<**等级 3**>(万能回复)很棒~
[S7]：<**等级 4**>(较好的)好好玩！北京还是很漂亮的。
[S8]：<**等级 5**>(极好的)好好玩！北京还是很漂亮的。你到时候住哪家酒店？

(b) “嘿，北京，我来了！”这个问题的真实回复。[Si]是一些样例回复。<等级 i>是样例回复对应的评分等级，类型是根据回复质量进行的解释。

4 数据标注

4.1 标注标准

完成原始数据的准备和整理工作后，我们聘请了四位标注人员对所有回复文本根据表 2 所示的五个等级标准进行等级评定。其中，质量标准从等级 1 到等级 5，分别对应“极差的”、“较差的”、“一般的”、“较好的”、“极好的”回复。对于每个回复，我们保证有两位标注人员分别进行独立评分。详细的评分等级及说明列于表 2(a)。同时为了更好理解标注标准，表 2(b)通过八个类型的数据样例说明各个评分等级的区别。

如表 2 所示，“极差的”回复指那些无意义的（例如[S1]）或者与问题不相关的（例如[S2]）回复。“较差的”回复可能与问题存在一定的相关性，但是在与问题的一致性、连贯性等方面有所欠缺，例如仅仅提到了少量关键词（例如[S3]）或者简单地重复问题中的片段（例如[S4]）等。回复的内容如果处在极差的或较差的评分等级，那么可以认为是低于一般水平的回复类型。

“一般的”评分等级可以认为是达到“及格”水准的回复质量。具有该等级评分的回复必须是内容有意义、前后衔接流畅并且与问题相关的特点。在“一般的”评分等级下，有两种典型的回复类型：其一、回复的内容有时间或者空间上的限制；其二、万能回复。对于第一类回复，回复的适合程度被限定在特定的时间或者空间条件下。例如实例[S5]在冬天看来，可以认为是一个合适的回复；然而，如果正处炎炎夏日，那么该回复就显得不合时宜了。对于万能回复，尽管它们没有时空上的限制，但是由于太过于通用，故而不能为提出的问题提供有效信息，例如[S6]。这类回复可以适用于多种不同类型的问题，前面提到，正因为这个特点，它们在收集的语料中广泛存在。为了有效区分这类回复和高质量回复，我们把万能回复定义为“一般的”而不是“较好的”或者“极好的”回复。

最后，等级 4 所对应的“较好的”回复往往是比较自然的、贴切的，既没有时空上的限制，也没有万能回复的特性，例如[S7]这类回复。而等级 5 所对应的“极好的”回复则可以更进一步，在回复中具备丰富的信息、幽默有趣，并能够有效促进对话过程往后推进，样例[S8]就是一个“极好的”回复，因为回复中提出了“北京的酒店”这一新的话题，因而积极地推动了对话的延续。

4.2 统计分析

由于数据集中的每个问答对都有两位标注人员进行标注，因此我们选择当且仅当他们的评分等级完全一致或差异为 1 时的相应的问答对进入最终的数据集合，最终符合该条件的问答对共有 82,010 对。对于每个入选的问答对，我们选取两位标注人员的平均分作为最终的回复评分。最终数据集包含 27,383 个问题以及 82,010 条回复，每个问题包含不同数量的回复，从 1 到 20 不等。在最终产生的数据集中问题回复数量的分布（百分比）如图 1 所示。从该分布可以观察到，只有少量问题含有 7 个及以上的回复，大多数问题所包含的回复数量在 1 到 6 个之间，包含 2-3 个回复的问题占比超过 60%。

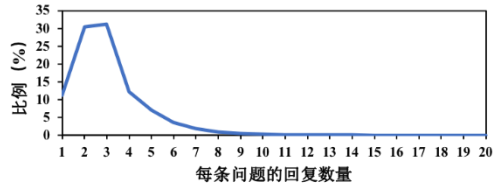

图 1. 每个问题对应回复数量的分布

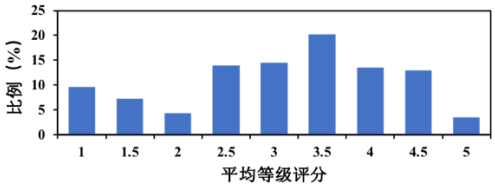

图 2. 两位标注人员的平均评分分布比例

在我们的数据集上，同一问答对不同标注人员之间的标注一致性，我们通过卡帕系数[26]来衡量，在最终数据集中，同一个回复的两个评分之间的卡帕一致性达到 80.3%，这表明了整个数据集上评分的高度一致性，侧面反映了该数据集的评分结果的可靠性。图 2 展示了数据集中总体评分的分布，其中 48.6%的回复评分属于[2.5,3.5]这个区间，反映了数据集中大量存在“一般的”回复。如前文所述，这类回复属于万能回复或者是时空限制的回复，这种类型回复的大量存在显示出从“极好的”和“较好的”实例中分离出“时空限制的”回复以及“万能回复”的重要性，从而能够进一步精确地区分出高质量的回复内容。经过进一步观察，我们发现 23.9%的回复得分在 2.5 分以下，一定程度上说明了社交媒体文本的回复质量良莠不齐，因此当训练和评估聊天机器人的时候，并不能假设所有用户生成的回复都是好

的结果。因此在对话系统中直接使用自动获取的数据具有一定程度的局限性，同时也进一步说明了在对话数据中区分对话回复质量的重要性。基于原始数据和发布数据上的多个维度指标详细信息参考表 3。

表 3 发布数据上的多个维度指标

指标	发布数据
问题个数	27,383
问答对个数	82,010
总字符数	1,386,450
总词数	1,030,629
问题平均含有字符数	6.33
问题平均含有词数	4.63
回答平均含有字符数	10.50
回答平均含有词数	7.88

5 基准实验

为了测试标注数据的合理性以及生成数据集的有效性，我们基于最终标注的数据集比较了不同回复选择模型的性能。这里我们使用回复选择模型作为测试方法的依据是，当前一般非任务型对话系统都是基于检索式的回复选择框架[27]，因此本文的实验设定可以有效反映实际系统的性能。实验设置描述详见 5.1，结果分析阐述于 5.2。

表 4 对比结果（%）。更高的分数表明更好的结果。阈值@N:表示回复评分大于等于 N 被认为是正例，其他就是负例。N 越大表明标准更加严格。

	模型	阈值@3			阈值@4			阈值@5		
		P@1	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR
无监督方法	Cosine sim	84.8	91.1	91.8	67.5	81.3	82.0	40.0	64.9	65.1
	BM25	86.1	**91.4**	**92.4**	70.6	82.7	83.7	53.8	72.3	73.4
有监督方法	SVMRank	**86.2**	91.1	92.2	73.0	**84.0**	85.0	64.3	78.6	79.8
	GBDT	85.9	91.0	92.0	71.4	82.9	83.8	55.7	74.4	74.5
	BiLSTM	85.4	90.7	91.8	**73.8**	**84.0**	**85.3**	68.1	81.0	82.2
	CNN	85.5	90.6	91.8	72.5	83.4	84.6	**70.5**	**81.2**	**83.0**

5.1 实验设置

在预处理中，我们使用结巴分词工具[10]进行中文分词。接着，我们随机选取问答对的 80%作为训练集，10%选入验证集，剩下的 10%作为测试集。在实验中，我们维护一个词典，这个词典包含在训练集中出现的所有词。表 5 中列出了实验数据集的详细统计信息。

在实验中，我们考虑两个非监督的排序（ranking）模型作为基线模型：Cosine Sim[11]和 BM25[28]。Cosine Sim 通过问题和回复的 TF-IDF 来计算余弦相似度，然后将回复根据相似度从高到低排序。BM25 模型根据类似 TF-IDF 的方法对回复排序。Cosine Sim 和 BM25 所使用到的词文档频率（DF）主要基于训练集来计算。同时，我们也测试了基于排序学习

[10] https://github.com/fxsjy/jieba
[11] https://en.wikipedia.org/wiki/Cosine_similarity

（learning-to-rank）的监督模型的结果。我们选择了两个经典模型：SVMRank[29]和梯度提升决策树（GBDT）[30]。这类模型需要依赖人工的特征提取工程，提取的特征与 Wang et. al(2013)[13]提出的方法相似，包括回复的句子长度、回复句子和对应问题的余弦相似度等等。额外地，我们还测试了两种广泛应用的神经网络模型：双向长短时记忆循环神经网络（BiLSTM）[31]和卷积神经网络（CNN）[32]。BiLSTM 和 CNN 可以实现端到端的训练，自动学习特征，无需依赖特征工程，训练方式类似于问答（QA）系统，对话问题和回复分别对应问答系统中的问题与答案。对于所有上面提到的模型，超参数的调节在验证集上进行。其中神经网络模型 BiLSTM 和 CNN 的编码器的隐层大小都设置为 300，使用均方误差（MSE）[33]作为损失函数，并且在训练时使用 early-stop[34]策略来防止过拟合。

表 5. 实验数据集的统计信息，均长表示句子切词之后词的平均个数

	对话问题		对话回复		词典
	数量	均长	数量	均长	大小
训练集	21,964	4.05	65,706	7.01	
开发集	2,669	4.02	8,080	6.98	36,035
测试集	2,750	4.11	8,224	7.03	

5.2 实验结果

我们遵循问答系统的评价方法：即给定一个问题来评价排序过的回复，需要将回复切分为“正样本”和“负样本”两类。因此，我们按照回复的评分等级将二分类的切分阈值 N 分别设为 3、4、5，将标注等级大于等于 N 的回复认为是正样本，其他等级的回复认为是负样本。总体来说，N 越大意味着标准越严格。表 4 展示了在不同的切分情况下，不同模型的实验结果。我们的评测指标是基于测试集得到的：P@1（precision@1）、平均精度均值（MAP）、倒数排名均值（MRR）。特别地，如果某个问题对应的所有回复按照排序阈值进行切分之后只有正负样本其中的一类，对于这类问题及其对应的回复我们会将其移除出我们的测试集，以保证模型评分的公正性。

最后的实验结果可以导出如下观察：1）从整体来看，监督模型比非监督模型结果更好，一定程度反映了我们的标注数据能够帮助监督模型辨别高质量回复。进一步观察监督模型和非监督模型在不同切分阈值上的差距时，我们发现，当标准越严格，监督模型与非监督模型的差距越大。这说明，当标准比较宽松的时候，非监督模型尚能通过一些简单的统计规则区分出真正“差”的回复 。但是当标准愈加严格的时候，非监督模型的性能急剧下降，而监督模型通过学习人工标注，能够很好地区分出更高质量的回复。以上观察说明了我们的标注结果对指导模型学习高质量回复颇有助益。2）对于监督模型而言，在阈值 N=3 和 4 之间的差距比阈值 N=4 和 5 之间的差距要大得多。产生这种现象的原因可能是 “极好的”和“较好的”回复相比于“较好的”和“一般的”回复区分度不大，这一观察也从侧面反映出在“好”的回复中区分出更高质量的回复对于标注人员而言亦是十分困难的任务，从而体现出标注对话数据集工作的挑战性。

6 结论

在本文所述工作中，我们构建了一个大规模人工标注中文对话数据集，其中包含了超过 2 万 7 千个不同的中文问题以及 8 万 2 千个回复。在这个数据集中，每个问题的每个回复根据与问题的相关性、连贯性以及内容的丰富性和趣味性等指标被分为五个评分等级。根据我们的调研，该数据集是第一个由人工标注的专门针对非任务导向的对话系统的中文数据集。相比于自动标注的数据集而言，本文所述数据集的标注质量更为可靠，可以助益于聊天机器

人的训练和评估。通过对话回复选择的实验，在这个数据集上，我们对比了不同模型的性能，实验结果反映了本文提出的数据集的客观性和有效性。

参考文献

[1] Alan M Turing. 1950. Computing Machinery and Intelligence. *Mind* 59(236):433–460.

[2] https://www.apple.com/ios/siri/

[3] njuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, László Lukács, Marina Ganea, Peter Young, Vivek Ramavajjala: Smart Reply: Automated Response Suggestion for Email. KDD 2016: 955-964.

[4] John Markoff and Paul Mozur. 2015. For Sympathetic Ear, More Chinese Turn to Smartphone Program. NY Times.

[5] Steve J. Young, Milica Gasic, Blaise Thomson, Jason D. Williams: POMDP-Based Statistical Spoken Dialog Systems: A Review. Proceedings of the IEEE 101(5): 1160-1179 (2013).

[6] Diana Perez-Marin. 2011. Conversational Agents and Natural Language Interaction: Techniques and Effective Practices. IGI Global.

[7] David S. Pallett, William M. Fisher, Jonathan G. Fiscus, John S. Garofolo: DARPA ATIS Test Results June 1990. HLT 1990.

[8] 李萍, 郑建明. 智慧图书馆中智能交互系统的研究和应用[J]. 图书馆学研究, 2016 (11): 34-38.

[9] Weizenbaum J.. ELIZA–A computer program for the study of natural laguage communication between man and machine[J]. Communications of the ACM, 1966, 9(1):36-45.

[10] Colby K.M., Weber S., Hilf F.D.. Artificial paranoia[J]. Artificial Intelligence, 1971, 2(1): 1-25.

[11] Wallace R.S.. The Anatomy of A.L.I.C.E. [EB/OL], A.L.I.C.E. Artificial Intelligence Foundation Inc., 2004.

[12] I. Serban, R. Lowe, L. Charlin, and J. Pineau. A survey of available corpora for building data-driven dialogue systems. arXiv preprint arXiv:1512.05742, 2015.

[13] Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A dataset for research on short-text conversations. In EMNLP, pages 935–945.

[14] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, Bill Dolan: A Diversity-Promoting Objective Function for Neural Conversation Models. HLT-NAACL 2016: 110-119.

[15] Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, Ray Kurzweil: Generating High-Quality and Informative Conversation Responses with Sequence-to-Sequence Models. EMNLP 2017: 2210-2219.

[16] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, Joelle Pineau: Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. ACL (1) 2017: 1116-1126.

[17] 张伟男, 张杨子, 刘挺. 对话系统评价方法综述. 中国科学: 信息科学, 2017, 47: 953–966.

[18] Tanveer J. Siddiqui, Uma Shanker Tiwary: Integrating Relation and Keyword Matching in Information Retrieval. KES (4) 2005: 64-73.

[19] Glen Pink: Slot Filling. University of Sydney, Australia 2017.

[20] Suket Arora, Kamaljeet Batra, Sarabjit Singh: Dialogue System: A Brief Review. CoRR abs/1306.4134 (2013).

[21] 代六玲 ,黄河燕 ,等. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报 ,2003 ,18(1) :26 - 32.

[22] Coates, A., Ng, A., and Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. pages 215–223.

[23] Lifeng Shang, Zhengdong Lu, Hang Li: Neural Responding Machine for Short-Text Conversation. ACL (1) 2015: 1577-1586.

[24] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, Joelle Pineau: Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. ACL (1) 2017: 1116-1126.

[25] Ryan Lowe, Nissan Pow, Iulian Serban, Joelle Pineau: The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. SIGDIAL Conference 2015: 285-294.

[26] Kilem L Gwet. 2014. Handbook of Inter-rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. Advanced Analytics, LLC.

[27] Ivan Kopecek: Modeling of the Information Retrieval Dialogue Systems. TSD 1999: 302-307.

[28] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to Information Retrieval. Cambridge University Press.

[29] Thorsten Joachims: Optimizing search engines using clickthrough data. KDD 2002: 133-142.

[30] Jerome H Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. Annals of statistics pages 1189–1232.

[31] Ming Tan, Bing Xiang, and Bowen Zhou. 2015. LSTM-based Deep Learning Models for Non-factoid Answer Selection. arXiv preprint arXiv: abs/1511.04108.

[32] Aliaksei Severyn, Alessandro Moschitti: Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks. SIGIR 2015: 373-382.

[33] Mean Squared Error. Encyclopedia of Machine Learning and Data Mining 2017:808.

[34] Rich Caruana, Steve Lawrence, C. Lee Giles: Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. NIPS 2000: 402-408