中国科学: 信息科学 2017 年 第 47 卷 第 8 期: 953–966
SCIENTIA SINICA Informationis
从大数据到大知识工程专刊 · 评述

对话系统评价方法综述

张伟男, 张杨子, 刘挺*

哈尔滨工业大学社会计算与信息检索研究中心, 哈尔滨 150001
* 通信作者. E-mail: tliu@ir.hit.edu.cn

收稿日期: 2017–04–06; 接受日期: 2017–06–21; 网络出版日期: 2017–07–24
国家重点基础研究发展计划 (批准号: 2014CB340503) 和国家自然科学基金 (批准号: 61502120, 61472105) 资助项目

摘要 本文介绍了对话系统的发展历史以及随着对话系统发展而衍生出的多种对话系统评价方法, 从任务型对话系统与开放域对话系统两个方向进行了调研和总结, 分析了不同评价方法的利弊, 每种评价方法的侧重点和不同方向上最新的研究进展. 在任务型对话系统方面, 介绍了 Steve Young 等人的近期研究成果, 总结了几种被广泛使用的评价思路. 在开放域对话系统方面, 从客观指标评价和模拟人工评分两个角度探索了开放域聊天系统的评价方法, 对于不同的指标和不同的研究思路做了分析及介绍. 最后, 本文通过总结及分析对话系统的经典评价方法和当前最新的基于神经网络模型的对话评价方法, 对对话系统评价方法的发展趋势进行了展望.

关键词 对话系统评价方法, 开放域对话系统, 任务型对话系统, 自然语言处理, 人工智能

1 对话系统发展概述

1950 年, 图灵在哲学刊物《思维》上发表 “计算机器与智能” 的文章 [1], 提出了后来被奉为经典的图灵测试 —— 交谈能检验智能, 即如果一台计算机能像人一样对话, 它就能像人一样思考. 图灵也由此获得 “人工智能之父” 的称号. 从此之后, 人类就一直在研究能让机器与人进行无障碍对话的方法. 1966 年的 ELIZA 是历史上出现的第一个人机对话系统 (简称对话系统), 1964~1966 年间由 Joel Weizenbaum 在 Massachusetts Institute of Technology (MIT) 编码完成, 主要用于在临床治疗中模仿心理医生对患者提供咨询服务. 1988 年出现的 UC 机器人用于帮助用户学习使用 UNIX 操作系统. 1999 年出现的 YAP 是用于查询英国电话黄页的机器人. 2004 年的 CSIEC, Sofia 等都是以完成特定任务为目的的对话系统, CSIEC 是一个外语学习伴侣, Sofia 帮助教师进行数学教学, 与用户互动的同时连接 Mathematica 软件实时为用户解决代数问题. 早在 1994 年就出现了第一个以娱乐闲聊为目的的开放域对话系统 (也称聊天机器人) A.L.I.C.E, 科学家华莱士将这个聊天程序安装到网络服务器, 然后观察网民与它的聊天行为. 近年来, 对话系统由于其应用的广泛性受到了越来越多的关注, 百度度秘、

引用格式: 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, doi: 10.1360/N112017-00125
Zhang Y Z, Zhang W-N, Liu T. Survey of evaluation methods for dialogue systems (in Chinese). Sci Sin Inform, 2017, 47: 953–966, doi: 10.1360/N112017-00125

表 1 目前几种主流对话系统功能比较

Table 1 The functional comparison of several conversational systems

System	Language	Platform	Function				
			QA	Speech	Recommendation	Instruction execution	Chit-Chat
Microsoft Cortana	Multilingual	Windows	√	√	√	√	√
Apple Siri	Multilingual	IOS	√	√	√	√	√
Google Allo	Multilingual	Android	√	√	–	√	√
Facebook Messenger	English	IOS & Andriod & Web page	√	√	–	√	–
Baidu Duer	Chinese	IOS & Andriod	√	√	√	√	√
Microsoft Xiaoice	Chinese	WeChat & Microblog	√	√	–	–	√

苹果 Siri、微软小冰等对话系统的诞生更是将人机对话的研究推向了前所未有的高度. 表 1 给出了目前比较热门的 6 个对话系统样例, 并根据支持的语言、搭载的平台以及具有的功能做了简单的比较.

随着神经网络和深度学习的快速发展, 循环神经网络、深度强化学习等多种学习方法让对话系统在对话生成这个方向取得了很大的进展, 但如何评价对话系统, 一直都是对话系统相关研究的重点和难点, 本文将从任务型对话系统和开放域对话系统这两个大方向上介绍目前对话系统评价的主流方法和思路.

2 任务型对话系统评价方法

随着任务型对话系统的诞生, 与其对应的评价方法也逐渐成为了一个活跃的研究方向. 1997 年, Walker [2] 提出了一个将对话持续时间及其他许多特征融入线性方程的系统 PARADISE, 用于推测用户的满意度, 该系统主要采用的方法是用一个已标注用户满意度的对话数据集和一个客观评价的数据集, 通过线性回归的方法对已标注的数据集求出一个可以用来表示用户满意度的权重指标, 指标的决定因素是对话成功率和对话成本消耗 (如对话时长、系统给出确认性质回复的次数等), 再通过强化学习将这个指标变成一个损失函数作为网络的奖励 (Reward), 由于这个方法很好地考虑了对话系统的多个不同因素, 这种方法后来也被用于对话策略学习 [3]. 由于实际操作中发现对话系统的成功率和对话的长度基本可以被认为是最重要的两个指标, 后来的研究也往往将最大化成功率与最小化对话长度作为任务型对话系统评测的指标.

然而当系统真正与人进行交互的时候, 任务完成的程度是很难界定的, 不仅如此, 生成模型理论上的有效性等一系列问题使得 PARADISE 的评价效果不尽如人意 [4]. 因此基于标注语料的数据驱动型对话评价模型成为了一个被广泛讨论的方向: 2012 年有研究者提出用协同过滤的方法来实现对用户反馈的表示 [5]; 利用重塑反馈函数也可以起到加速对话策略学习的目的 [6]; Ultes 与 Minker 等人 [7] 的研究发现专家满意度对于对话系统的回复成功率有很大的影响. 所有的这些方法和尝试都表明, 优质的训练数据对于对话系统的生成结果是至关重要的. 但是得到优质的标注数据是非常困难的, 耗费大量的专家资源来对数据进行系统完整的标注是非常不现实的, 所以后续有研究者提出了用机器模拟人类标注数据的过程来标注数据, 这样既可以减少人工消耗, 也可以产生更多的可用数据. 基于这个想法, 有研究者提出了动态学习 (active learning) 的方法, 为了减少标注误差而采用多种方式相结合的办法来对数据进行自动标注 [8].

Steve Young 等人 [9] 在 2012 年总结了任务型对话系统评价的基本情况. 数据驱动的自然语言处理 (natural language processing, NLP) 任务有很多种评价方法, 但由于多轮交互性的影响, 对对话系统的评价难度要更大一些. 尽管目前已经有很多针对不同指标的评价矩阵 (即将客观指标用特征矩阵的方式进行表示, 从而达到可运算的目的), 但如何将他们很好地结合起来用于评价对话系统仍然是个难于解决的问题. 事实上, 对话系统评价的最终目标是测评用户的满意度, 但总有许多因素造成我们无法将评价结果与用户的体验感受完全吻合, 即使通过给出的人工制定的评价指标来对一个系统进行测试, 也会造成不同程度的偏差, 而且也很难将所有的特征罗列出来并加以对比达到全面的评价效果, 因此现有的评价过程大都无法准确的满足用户的要求. 针对这些问题, 文章提出了 3 个针对任务型对话系统评价的对策: 1) 通过构造某种特定形式的用户模拟系统进行评价; 2) 人工评价; 3) 在动态部署的系统中进行评价 [10].

2.1 用户模拟

用户模拟是一种有效且简单的评价策略, 并且是最有可能覆盖最大对话空间的方法, 这是由于通过模拟不同情境下的对话, 可以有效地在大范围内进行测试和评价 [11~13]. 然而这种方法的缺点也很明显, 就是真实用户的反映与模拟器的反应之间潜在的差距, 这个差距的影响大小某种程度上取决于用户模拟器的好坏. 然而即使这个问题无法解决, 用户模拟仍然是任务型对话系统评价中最常用的评价方法, 曾被用于评价多种不同的基于部分可观察 Markov 决策过程 (partially observable Markov decision process, POMDP) 的对话策略 [14, 15].

实验对比了在不同噪声等级下, 同一个用户模拟系统针对 3 个不同的对话管理器的评价结果 (实验结果参看文献 [9] 的 Fig.6.), 其中 HDC 表示系统将人工选取出最像模拟器生成的结果作为输入, BUDS 表示系统输入是非人工选择的, BUDS-HDC 表示系统的输入是一个人工编码的对话策略产生的, BUDS-TRA 表示系统输入是通过 NAC 这一策略进行训练之后得到的模型生成结果 (NAC 策略是一种强化学习策略, 在文献 [9] 前文中提及, 此处不做详解). 系统模拟的情境是一个自助旅游信息问询系统, 用户可能会问到涉及一个虚拟小镇中的旅馆、餐馆、酒吧等多种娱乐信息的问题. 评价方程定义为对话轮数的最大容忍值 (即若在这个轮数内系统没有给出答案则认为此次对话失败, 此处预先设定取值为 20) 与实际对话轮数的差值, 图中的纵坐标 Mean reward 表示每一次对话之后评价方程的取值, 横坐标 Confusion rate 被定义为对话管理器产生的语义结果与假设中给出的结果不相同的概率. 图中可以清晰的看到不同输入会产生不同的结果特征, 起初在低错误率的范围内每个对话系统的表现都很相近, 而在高错误率的范围内, 可以明显看出 BUDS 的系统鲁棒性要优于 HDC, 也就是说 BUDS 对于噪声的抵抗性要优于 HDC. 后续跟踪评价结果也发现, 在多轮对话内容间产生如前后内容不一致、对话突然终止等冲突时, BUDS 的处理能力也优于 HDC. 除此之外, 将经过策略学习产生对话结果的 BUDS-TRA 与人工编码策略的 BUDS-HDC 相比较, 可以发现 BUDS-TRA 后期的效果有较明显的提升, 这表明了通过强化学习策略 NAC 进行的优化是有效的.

2.2 人工评价

第 2 个任务型对话系统评价方法是通过雇佣测试人员对对话系统生成的结果进行人工评价, 这样做的好处是能够产生更多真实的评价数据. 到目前来看, 这种评价方法更多地出现在实验室等研究资源雄厚的环境中, 测试人员在预定任务领域内对系统进行评测, 通过一些预设的询问方式与系统进行对话, 根据对话结果对系统的表现进行评分. 通过 POMDP 进行对话策略学习的对话系统已经采用了

人工评价的方法, 并将结果应用于人工编码策略与 MDP 策略的结果提升任务中 [15~17].

然而这种评价方法最大的问题在于如何雇佣足够多的测评人员, 很明显这需要大量的开销. 后期出现的外包模式以及借助网络媒介延迟较小的特点在网络上进行实时评价等方法都可以用于尝试解决这个问题. 例如使用 AMT (the Amazon Mechanical Turk) 服务 [18]: 给出预定义的任务以及基础的培训指令, 雇佣评测员对指定的任务进行评测, 评测员可以通过免费电话对对话系统进行评价, 每次对话之后给出反馈信息. 这个方法可以有效地产生大量对话系统与人的真实对话数据, 从而产生大量的数据统计结果 [19]. 除开销巨大之外, 这种评价方法还存在外包选择的评测人员是否真的能够代表所有用户的问题. 事实上如果对实验集合没有很好的监控, 人工评价的动机和目的就会成为最后评价结果的重要影响因素, 也有事实证明人工评价并没有非常完整地表现出对话的效果特点 [20].

2.3 部署动态系统的评价

任务型对话系统评价的理想状态就是在真实用户群中检测用户的满意度, 这种评价方法通常是在商业广告中植入对话系统或构建一个能够让公众主动去使用对话系统的服务设施. 很明显这两种方法都是较难实现的. Carnegie Melon University (CMU) 的研究者曾提出过一个评价架构, 称为对话系统挑战 (spoken dialogue system challenge). CMU 首先开发了一个对宾夕法尼亚州匹兹堡的用户提供在线公交信息查询的对话系统, 用户可以通过给这个对话系统打电话来查询公交信息 (在此之前这项服务是在工作时间由人工完成的). 由于产生了非常真实的用户需求, 通过提供这样一个全时间段自动化的查询服务, CMU 开发的对话系统成为了这项服务的一个标准. 后期出现参与挑战的系统, 如果通过测试证明效果更好, 可以替换这个系统 (例如在 2010 年的挑战中, 就有学生给出了比原有系统鲁棒性更强的新系统, 并参与了挑战 [21]).

3 开放域对话系统评价方法

目前对开放域环境下的聊天机器人评价的方法主流有两种思路, 客观指标评价与模拟人工评分. 本文对客观指标部分的介绍主要包括两方面: 一是以 BLEU [22], METEOR [23] 和 ROUGE [24] 为代表的词重叠评价矩阵; 二是以 Greedy Matching [25], Embedding Average [26], Vector Extrema [27] 为代表的基于词向量的评价矩阵. 模拟人工评分部分本文主要介绍了前沿的 3 种用神经网络模拟人工评分的方法: Google Brain 的 Anjuli Kanan 和 Google Deepmind 的 Oriol Vinyals 等人 2017 年 1 月提出的一种类 GAN 结构的对抗评价模型; McGill 大学 Ryan 等人提出的基于 RNN 的 automatic dialogue evaluation model (ADEM) 对话评价系统和基于 ANN 结构的对话评价系统.

3.1 基于客观指标的评价方法

3.1.1 基于词重叠率的评价矩阵

根据已有的 NLP 任务经验, 当提到如何评价模型生成结果的质量时, 首先想到的就是根据生成的回复与标准答案之间的词重叠率来进行评价: BLEU 和 METEOR 是在机器翻译任务中取得很好效果的两种评价方法; ROUGE 也在文本的自动摘要任务中取得了不错的评价效果. 在非对话系统领域内这些指标被普遍认为可以准确反映生成结果的部分特征, 虽然还没有彻底的适应对话系统类型的任务, 但也有了不少值得探讨的尝试.

下文将介绍几种不同的客观指标, 若没有特殊说明, 下文出现的公式中均用 r 表示语料中的真实回复 (这里假设每一句输入只有一个确定的真实回复, 下文都将写为 “参考答案”), $\hat{r}$ 来表示模型生成的输出, 对于第 j 个序列, 产生真实回复 r 与产生 $\hat{r}$ 的序列的分别记作 w_j 和 $w_{j'}$.

BLEU. BLEU 是一种对模型输出和参考答案的 n-gram 进行比较并计算匹配片段个数的方法. 这些匹配片段与它们在上下文 (Context) 中存在的位置无关, 这里仅认为匹配片段数越多, 模型输出的质量越好. BLEU 首先会对语料库中所有语料进行 n-gram 的精度 (Precision) 计算 (这里假设对于每一条文本, 都有且只有一条候选回复):

$$P_n(r,\hat{r}) = \frac{\sum_k \min(h(k,r), h(k,r_i))}{\sum_k h(k,r_i)}, \tag{1}$$

式中 $P_n(r,\hat{r})$ 代表语料库中所有语料 n-gram 的精度, k 表示长度为 n 的序列的 n-gram 值, $h(k,r)$ 为 r 中 n-gram k 的个数. 为了避免使用精度值带来的问题 (精度选择更倾向于短回复) 演变出了 BLEU-N, N 表示了经过选择后 n-gram 中 n 的最大值:

$$\text{BLEU-}N = b(r,\hat{r}) \exp\left(\sum_{n=1}^{N} \beta_n \log P_n(r,\hat{r})\right), \tag{2}$$

β_n 一般是一个恒定的权重, $b(r,\hat{r})$ 表示惩罚因子. BLEU-4 在许多任务中都被广泛使用. 由于如果回复中不存在 4 元重叠组则几何平均值会为 0, 这样会导致计算错误, 所以现在许多不同的 BLEU-4 版本都需要顺滑操作来保证重要信息不会受损. BLEU 通常是在数据集级别进行计算, 常被用于在多个候选答案中选择更好的那一个.

METEOR. METEOR 矩阵会在候选答案与目标回复之间产生一个明确的分界线 (这个分界线是基于一定优先级顺序确定的, 优先级从高到低依次是: 特定的序列匹配、同义词、词根和词缀、释义). 有了分界线之后, METEOR 可以把参考答案与模型输出的精度 (Precision) 与召回率 (Recall) 的调和平均值作为结果进行评价. 具体的作法是: 对于一个模型输出 c 与其对应的参考答案 r 的 (c,r) 序列 m, METEOR 矩阵值是其精度 P_m 与召回率 R_m 的调和平均值, Pen 是根据已有的正确答案预先计算出的一个惩罚因子, 公式中的 α,β,γ 都是具有默认值的超参数常量.

$$\begin{aligned} F_{\text{mean}} &= \frac{P_m R_m}{\alpha P_m + (1-\alpha) R_m}, \\ \text{Pen} &= \gamma(\text{frag})^{\theta}, \\ \text{METEOR} &= (1-\text{Pen}) F_{\text{mean}}. \end{aligned} \tag{3}$$

ROUGE. ROUGE 是一系列用于自动生成文本摘要的评价矩阵, 记为 ROUGE-L, 它是通过对候选句与目标句之间的最长相同子序列 (longest common subsequence, LCS) 计算 F 值 (F-measure) 得到的. LCS 是在两句话中都按相同次序出现的一组词序列, 与 n-gram 不同的是, LCS 不需要保持连续 (即在 LCS 中间可以出现其他的词). 公式中 s_{ij} 表示与候选回复 c_i 对应的第 j 个模型输出, $l(c_i,s_{ij})$ 表示两者间 LCS 的长度, β 是超参数常量.

$$\begin{aligned} R &= \max_j \frac{l(c_i,s_{ij})}{|s_{ij}|}, \\ P &= \max_j \frac{l(c_i,s_{ij})}{|c_{ij}|}, \end{aligned} \tag{4}$$

$$\mathrm{ROUGE_L}(c_i, s_i) = \frac{(1+\beta^2)RP}{R+\beta^2 P}.$$

3.1.2 基于词向量的评价矩阵

除了词重叠率外, 另一种考量回复效果的思路是通过了解每一个词的意思来判断回复的相关性, 词向量是实现这种评价方法的基础. 依据语义分布, 采用 Word2Vec [28] 等方法, 给每一个词分配一个向量用于表示这个词, 这种表示方法通过计算这个词在语料库中出现的频率来近似地表示这个词所表达的含义. 所有的词向量矩阵通过向量连接就可以近似为句子级的句向量. 通过这种方法可以分别得到候选回复句与目标回复句的句向量, 再通过余弦距离进行比较, 就可以得到二者的相似度.

Greedy matching. 贪婪匹配方法 (greedy matching) 是基于词级别的一种矩阵匹配方法. 在给出的两个句子 r 和 $\hat{r}$, 每一个词 $w \in r$ 都会经过词向量转换后变为词向量 e_w, 同时与 $\hat{r}$ 中的每一个词序列 $\hat{w} \in \hat{r}$ 的词向量 $e_{\hat{w}}$ 最大程度进行余弦相似度匹配, 最后得出的结果是所有词匹配之后的均值:

$$G(r, \hat{r}) = \frac{\sum_{w \in r} \max_{\hat{w} \in \hat{r}} \cos_{\mathrm{sim}}(e_w, e_{\hat{w}})}{|r|},$$

$$GM(r, \hat{r}) = \frac{G(r, \hat{r}) + G(\hat{r}, r)}{2}. \tag{5}$$

由于上面的计算公式是不对称的, 所以需要在各个方向上都对 G 求均值以保证结果的准确性. 贪婪匹配最早在智能导航系统中提出, 后续研究发现这种方法选择的最优解往往偏向于中心词与参考答案在语义上有较大相似度的结果.

Embedding average. 向量均值法 (embedding average) 是通过句子中的词向量计算一个句子特征向量的方法, 通过对句子中每一个词的向量求均值来计算句子的向量. 这种方法在除对话系统之外的很多 NLP 领域内都应用过 (例如计算文本相似度的任务), 公式中 $\bar{e}$ 表示句子 r 中所有词组的词向量均值:

$$\bar{e} = \frac{\sum_{w \in r} e_w}{|\sum_{w' \in r} e'_w|}. \tag{6}$$

对比 r 与 $\hat{r}$ 时可以将两个句子分别按照上述方法计算向量均值, 再把二者的余弦相似度作为相似性的指标进行评价, 即 $EA := \cos(\bar{e}_r, \bar{e}_{\hat{r}})$.

Vector extrema. 另一种在句子级向量上计算相似度的方法是向量极值法 (vector extrema). 通过筛选词向量的每一维来选择整句话中极值最大的一维作为这个句子的向量表示:

$$e_{rd} = \begin{cases} \max\limits_{w \in r} e_{wd}, & \text{if } e_{wd} > |\min_{w' \in r} e_{w'd}|, \\ \min\limits_{w \in r} e_{wd}, & \text{otherwise.} \end{cases} \tag{7}$$

公式中 d 代表词向量中的维度; e_{wd} 是 w 的词向量中第 d 维. 当然想要更准确的表达两个回复的相似度, 仅计算向量极值是不够的, 还需要计算回复之间的余弦距离才能更好地表示它们之间的相似程度. 直观上看, 在某个文本中具有特殊意义的词应当具有比常用表达更高的优先级, 但由于常用表达往往会出现在更多的文本中, 这种计算方法会使得在向量空间中它们离得更近, 计算相似度之后常用表达就会占据输出向量排序更靠前的位置, 这会使得具有重要语义信息的词被 “挤” 到靠后的位置. 由于这个原因, 在采用向量极值法的时候需要有意识地忽略常用表达 [29].

3.2 基于评分模拟的评价方法

3.2.1 GAN 结构

生成式对抗网络 (generative adversarial networks, GAN) 自 2014 年由 Ian Goodfellow 提出至今, 不仅催生了很多理论论文, 也带来了层出不穷的实际应用, 已经成为人工智能学界一个热门的研究方向. Google Brain 的 Anjuli Kanan 和 Google Deepmind 的 Oriol Vinyals 等人 2017 年 1 月提出了一种类 GAN 结构的对抗评价模型, 并设计了一种类 GAN 的网络结构, 用于直观评价生成器 (Generator) 产生的回复结果与人类回复的相似程度.

受到 GAN 在图像生成任务上的成果启发, Anjuli 等人的工作也是采用了 GAN 的基本生成器 - 分类器 (Generator-Discriminator) 结构, 通过训练得到的 Generator 用于生成回复, Discriminator 用于区分人的回复与 Generator 生成的结果.

与传统 GAN 的基本结构不同, 模型的 Generator 是一个序列对序列 (sequence to sequence, Seq2-Seq) 模型, 包含一个完整的循环神经网络 (recurrent neural network, RNN) Encoder + Decoder 结构, 而 Discriminator 虽然也是一个 RNN, 但采用的是 Encoder 加一个二元分类器的结构.

训练时将数据集中的信息对 (o, r) 输入 Generator, 其中 o 代表原始数据, 由序列 $o_1, o_2, ..., o_n$ 构成, r 代表回复信息, 由序列 $r_1, r_2, ..., r_m$ 构成. 训练的目标是在给定原始信息 o 的情况下, 使生成的回复结果 r 出现的概率最大, 从而根据评价方程 P_g 选择最合适的回复结果:

$$P_g = \sum_{(o,r)} \log P(r_1, r_2, ..., r_m | o_1, o_2, ..., o_n). \tag{8}$$

对于 Discriminator, 训练时输入的信息为 (o, r, y), 其中 o 和 r 分别代表了原始信息与生成回复的信息, y 则用于区分二者, $y = 1$ 代表数据为训练数据, $y = 0$ 则相反. 训练阶段的目标是让 r 与 o 相近程度高的信息对分数接近 0, 即分数越高, Discriminator 越容易识别出生成这个结果的是 Generator 而不是人类, 认为 Generator 生成的结果越不好. 评价方程为 P_d:

$$P_d = \sum_{(o,r,y)} \log P(y | o_1, o_2, ..., o_n, r_1, r_2, ..., r_m). \tag{9}$$

实验的数据集 C 由邮件的回复对 (o, r) 构成, Generator 的训练集由这个数据集和 Gmail 提供的一些智能回复共同组成; Discriminator 的训练数据集则更加复杂, 一部分是抽取 C 中的部分数据 C' 将对应的分值 y 标注为 1, 另一部分是将 C' 中信息对 (o, r) 里的 y 用 Generator 生成的结果 r' 替代变为 (o, r'), 并标注为 0.

通过实验结果可以清晰地看出 Discriminator 的区分效果与回复的长短有很大的关系 (参看文献 [30] 的 Figure.1), 图像很清晰地展示了实验的效果, 左图表示的是分数与回复长度的关系, 横坐标表示回复的长度, 纵坐标表示 Discriminator 给出的分数, 随着回复长度的增加, 分数在逐渐变高, 即回复的长度越长, Discriminator 越容易区分. 右图也展示了召回率与精度的关系, 随着训练次数的迭代, 精度越来越高的同时召回率也在减小, 即 Generator 产生的结果与原始数据越来越相近, 意味着 Discriminator 就越来越难区分回复来自于 Generator 还是来自训练数据. 表 2 也列举了一些回复结果的实例, 根据例子可以发现长度较短的一些生成结果要更贴近真实的回复 [30].

表 2 Discriminator 与 Generator 对于同一问题的相同长度回复结果对比 [30]

Table 2 The comparison of responses ranking of the same length by discriminator score and generator score [30]

Original	Length	Top responses (Discriminator)	Top responses (Generator)
Actually, yes, let's move it to Monday	3	Oh alright.	OK thank you
		That's fine.	That's fine
		Good....	Thank you!
Are you going to Matt's party on Sunday?	1	Ya	Yes
		Maybe	No
		Yeah	Yes

3.2.2 RNN 结构

McGill 大学的 Ryan 等人 2016 年 6 月的研究发现, 传统的客观评价指标都具有一定的局限性, 无法很完整地表示评分与人类评价的相关度, 于是尝试使用 RNN 的方法进行自动评分模型的训练, 并提出了一个对话系统自动评价模型 (automatic dialogue evaluation model, ADEM) 用于预测回复的人工评价结果, 同时也将 ADEM 的评分结果与传统指标 BLEU, ROUGE 进行了对比, 证明了自动评价系统的可行性.

ADEM 是一个通过半监督性学习方法训练得到的多层 RNN 结构的评价模型, 使用了多层编码器 (Encoder) 来将训练语料中文本转化为向量, 训练阶段的输入为对话文本 c, 生成回复 $\hat{r}$, 参考回复 r. 首先由 ADEM 中的 Encoder 将这些语料分别转化为向量, 然后通过对这些向量进行线性变换得到一个分数:

$$\text{score}(c,r,\hat{r}) = \frac{c^{\mathrm{T}} M\hat{r} + r^{\mathrm{T}} N\hat{r} - \alpha}{\beta}, \tag{10}$$

式中的 $M, N \in R^n$ 是经过学习后得到的参数, 用于初始化等式; α, β 是标量常数, 用于将得到的分值控制在 $[0,5]$ 这个范围内. M 和 N 可以被看作将 $\hat{r}$ 分别映射到 c 和 r 中的一种线性变换方法, 基于这一点不难理解得分高的回复一般会跟 c 与 r 的向量表示比较相似. 由于上面的公式是可微的, 所以所有的参数都可以通过反向传播的方法进行学习, 在这个实验中, 通过 L1 正则化方法最小化 score 与人工评分的方差来确定参数 $\theta = M, N$:

$$L = \sum_{i=1:k} [\text{score}(c_i, r_i, \hat{r}_i) - \text{human\_score}_i]^2 + \gamma||\theta||_1. \tag{11}$$

图 1 出自文献 [31], 通过 ADEM 结构可以看出, ADEM 采用的复合 RNN Encoder 结构包括两层, 底层用于将文本中的词变为词向量, 上层则是根据词向量产生整个上下文 (Context) 的向量表示. ADEM 采用的这种 RNN 的结构与对话系统所使用的 RNN 结构最大的区别在于 ADEM 将参考回复作为一个变量参与训练, 从而在训练的过程中, 可以随时将模型生成的结果与参考回复做对比, 大大降低了对比两种答案的难度.

由于人工标注数据费时费力, 实验希望在训练过程能够用更少的标注数据达到更准确的预测效果, 所以采用了预训练的方法学习 Encoder 的参数, 实验采用的预训练方法是将原模型中的 Encoder 产生的结果当作输入送入一个独立的 RNN, 然后经过对这个 RNN 的训练产生特定条件下对特定 Context 的回复, 并把这些数据当作原 RNN 的训练数据. 这样一来, 同一句 Context 就可以产生许多句不同的回复, 从而可以得到更多的训练数据.

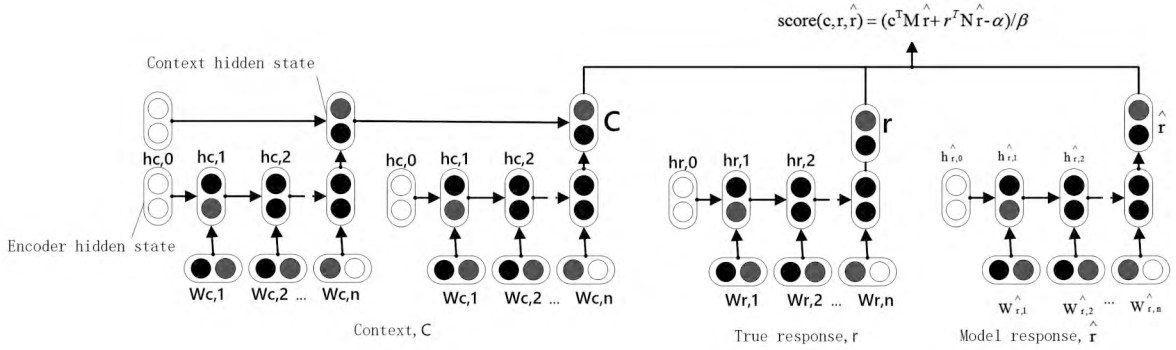

图 1 ADEM 的基本结构 [31]

Figure 1 The basic structure of ADEM [31]

实验使用了 Twitter 数据集, 在开始实验之前, 作者先通过 Amazon Mechanical Turk 的志愿者对数据集中给出的不同问题的不同回复进行评分, 并且对人工评价的分数做了分析, 根据人工评价得到的分数特点结合现有的 Context 与回复, 再通过上文提到的预训练方法生成大量数据加入到实验数据集中.

实验将 ADEM 的评价结果与 METEOR, BLEU-N (N=1,2,3,4) 的结果在 Spearman 与 Pearson 两个指标上进行比较, 见表 3. 表格中的 T2V 代表 ADEM 仅使用 Twitter 数据集作为训练数据集, VHRED 代表的是训练数据集为 Twitter 数据集加上预训练得到的数据, C-ADEM 代表模型训练时生成回复结果仅与 Context 进行对比, R-ADEM 代表结果仅与参考回复进行对比; Full dataset 表示全部训练数据集, 包含 4104 个回复, Validation set 表示验证数据集 (可以理解为小数据量的训练数据集), 包含 616 个回复, Test set 表示测试数据集, 包含 616 个回复. 同时也通过拟合图像的方式展示了实验的结果 (参看文献 [31] 的 Figure 3). 直观上看, 与人工评分的结果分布相比, ADEM 的分数要比简单的 BLEU 与 ROUGE 拟合程度更好一些, 即认为 ADEM 的效果要好于 BLEU 和 ROUGE [31].

3.2.3 基于神经网络结构的其他评分模拟

训练数据的匮乏一直是开放域对话系统领域面临一个重大难题, 基于数据处理的难度和数据标注消耗巨大的现状, Ryan 的团队认为应该从大量未标记数据中寻找训练对话系统的方法, 由此提出了用下文回复分类 (next utterance classification, NUC) 作为从文本数据中训练对话系统的一个开端 [32].

实验为了比较人工评价的结果与人工神经网络 (artificial neural network, ANN) 产生结果的差别, 在训练阶段采用了 Dual Encoder (DE) 来训练数据, DE 由带有长短期记忆模型 (long short-term memory, LSTM) 的 RNN 组成, 用于将 Context c, 候选回复 r 处理为向量形式. 在 t 时刻, 一个词 x_t 被送入 LSTM, 隐层 h_t 根据 $h_t = f(W_h h_{t-1} + W_x x_t)$ 进行更新, W 是权重矩阵, f 是线性激活函数. 当一句话中所有 T 个词全部进入 LSTM 之后, 最后一个隐藏层状态 h_T 就可以被看作整句话的向量表示.

为了确定一个回复 r 是一段 Context 的真实回复的概率, 模型会将 c 与 r 分别做一个加权点积:

$$P(r \text{ is correct response}) = \sigma(c^{\mathrm{T}} M r), \tag{12}$$

其中 M 是要学习的参数, σ 是 sigmoid 函数, 用经典的交叉熵函数作为损失函数, 模型训练过程要将 (c, r) 对的交叉熵最小化, 从而得到最准确的评价模型.

表 3 自动评价指标与人工判断结果的相关性 [31]a)

Table 3 Correlation between metrics and human judgements [31]

Metric	Full dataset		Test set	
	Spearman	Pearson	Spearman	Pearson
BLEU-1	0.026 (0.102)	0.055 (<0.001)	0.036 (0.413)	0.074 (0.097)
BLEU-2	0.039 (0.013)	0.081 (<0.001)	0.051 (0.254)	0.120 (<0.001)
BLEU-3	0.045 (0.004)	0.043 (0.005)	0.051 (0.248)	0.073 (0.104)
BLEU-4	0.051 (0.001)	0.025 (0.113)	0.063 (0.156)	0.073 (0.103)
ROUGE	0.062 (<0.001)	0.114 (<0.001)	0.096 (0.031)	0.147 (<0.001)
METEOR	0.021 (0.189)	0.022 (0.165)	0.013 (0.745)	0.021 (0.601)
T2V	0.140 (<0.001)	0.141 (<0.001)	0.140 (<0.001)	0.141 (<0.001)
VHRED	– 0.035 (0.062)	– 0.030 (0.106)	– 0.091 (0.023)	– 0.010 (0.805)
Metric	Validation set		Test set	
	Spearman	Pearson	Spearman	Pearson
C-ADEM	0.272 (<0.001)	0.238 (<0.001)	0.293 (<0.001)	0.303 (<0.001)
R-ADEM	0.428 (<0.001)	0.383 (<0.001)	0.409(<0.001)	0.392 (<0.001)
ADEM (T2V)	0.395 (<0.001)	0.392 (<0.001)	0.408 (<0.001)	0.411 (<0.001)
ADEM	0.436 (<0.001)	0.389 (<0.001)	0.414 (<0.001)	0.395 (<0.001)

a) 表中括号内为 p- 值, ADEM (T2V) 表示 ADEM 方法采用 tweet2vec 词嵌入结果, VHRED 表示采取点乘 VHRED 词嵌入矩阵的方法, C- 和 R-ADEM 表示 ADEM 模型训练过程中分别只对比模型产生的回复与上下文 (Context) 或参考答案之间的相似性.

实验的数据集有 3 部分组成. 1) Ubuntu 数据集: 从 Ubuntu 上的开放聊天平台 IRC 中提取出的对话数据 [33]; 2) Twitter 数据集: Twitter 提供的大量 Twitter 用户的对话数据 [34]; 3) The SubTle 数据集中电影部分 [35]: 一些已标记对话结束位置的电影对白. 人工评分参考的数据是 145 名在 AMT 上找到志愿者 (志愿者包括不同性别、年龄、不同学历、不同工作等多种类型) 进行评分的结果, 具体志愿者的信息可以参看文献 [32] 中提供的表 4, 可以看出这些志愿者对不同数据集中的数据都进行了评分, 用于最终的 ANN 模型生成结果的参考.

表 5 出自文献 [32], 是实验的最终结果. 实验结果将召回率作为评价指标, R@k 表示在排序生成的前 k 个结果中出现正确答案的比例, 把实验提出的 ANN 结构与人工评分结果进行对比. 不难发现在 3 个数据集上, ANN 的表现都低于人类评分, 在 Twitter 与 Ubuntu 数据集中, R@2 与人工评价结果较为接近.

通过表 5 我们可以发现, 人类在对回复进行分类的时候有明显的规律, 随机情况只占极少数, 对于不同领域不同知识层次的人, 分类也有不同的结果, 这是有规律可循的. 此外, 这项研究还发现, 在对话自动生成的任务中, 机器的实验结果与人类幼儿的回复结果非常相近, 但与专家级别的回复结果差距较大, 所以可以认为对话系统仍有很大的进步空间.

4 未来趋势

基于对话系统现有的两种不同分类方法, 研究者们有针对性地提出了各种不同的评价方法, 对话系统评价这个领域也在近两年有了飞速的发展, 各种各样的评价方法层出不穷. 对于任务型对话系统

表 4 145 名 AMT 评测人员信息统计 [32]

Table 4 Statistical items of 145 AMT participants [32]

Question	Item	Value
What is your gender?	Male	56.5%
	Female	44.5%
What is your age?	18–20	3.4%
	21–30	38.1%
	31–40	33.3%
	41–55	14.3%
	55+	10.2%
How would you rate your fluency in English?	Beginner	0%
	Intermediate	8.2%
	Advanced	6.8%
	Fluent	84.4%
What is your current level of education?	High school or less	21.1%
	Bachelor's	60.5%
	Master's	13.6%
	Doctorate or higher	3.4%
How would you rate your knowledge of Ubuntu?	I've never used it	70.7%
	Basic	21.8%
	Intermediate	5.4%
	Expert	2.7%

表 5 每个数据集中的平均得分 [32]

Table 5 Average results on each corpus [32]

	Number	Movie corpus		Twitter corpus		Ubuntu corpus	
	of users	R@1	R@2	R@1	R@2	R@1	R@2
AMT Non-experts	135	65.9 ± 2.4%	79.8 ± 2.1%	74.1 ± 2.3%	82.3 ± 2.0%	52.9 ± 2.7%	69.4 ± 2.5%
AMT experts	10	–	–	–	–	52.0 ± 9.8%	63.0 ± 9.5%
Lab experts	8	69.7 ± 10%	94.0 ± 5.2%	88.4 ± 7.0%	98.4 ± 2.7%	83.8 ± 8.1%	87.8 ± 7.2%
ANN model (Lowe et al., 2015a)	(Machine)	50.6%	74.9%	66.9%	89.6%	66.2%	83.7%

可以利用任务完成程度当作硬性的评价指标进行评价, 对于开放域的对话系统, 通过对客观指标进行评分, 可以很好地根据每个指标对对话系统模型进行修改和提高. 通过模拟人工评分的结果, 可以对评价这件事有一个整体的宏观的认知, 这些都是对话系统评价领域内非常好的研究思路.

根据上文所述, 对于任务型对话系统的评价仍然需要继续提高与人工评价结果的拟合程度; 对于开放域对话系统, 客观指标和模拟评分这两种方法都有各自的优缺点, 如何利用二者不同的特点来提高评价的效率和准确率是未来研究的重点. 如何能够减轻模型比较和选择的负担, 增强评价系统的可扩展性, 实现一种可迁移到不同数据集中进行评价任务的评价方法, 是未来要重点关注的研究方向.

5 总结

本文通过介绍目前两种对话领域的主流评价思路, 总结了目前在对话评价领域内的研究成果. 对于任务型人机对话系统, 主流评价思路仍然是以人工评价结果为目标不断进行探索. 对于开放域的人机对话系统, 目前的研究从切入点来区分主要有两个大方向: 客观指标与模拟评分. 客观指标主要包括词重叠评价矩阵和基于词向量的评价矩阵. 模拟人工评分主要是通过神经网络的训练方法, 用机器来模拟人的打分过程, 从而实现对对话系统的评价.

虽然每种评价方法都有自己的优势, 如对于开放域对话系统的评价, 客观指标可以更有针对性地发现对话系统的问题, 根据不同的指标可以有的放矢地提高对话系统的性能; 模拟评分则采用了更加宏观的角度, 抛开细枝末节直接对整体进行模拟, 从而契合人工评价的结果, 而且通过模拟还可以产生更多的标注语料, 减少人工功耗. 但是它们也都有自己的不足, 如评价矩阵更多的是用于非对话系统的其他 NLP 任务中, 还没有针对人机对话任务进行很好的整合迁移; 而直接抛开细节的评分模拟, 很可能也忽略了重要的因素, 所以效果不是非常出众. 基于这样的现状, 对于对话系统评价的研究还有很大的空间可以供研究者探索.

参考文献

1 Turing A M. Computing machinery and intelligence. Mind, 1950, 59: 433–460
2 Walker M A, Litman D J, Kamm C A, et al. PARADISE: a framework for evaluating spoken dialogue agents. In: Proceeding of the 8th Conference on European Chapter of the Association for Computational Linguistics, Madrid, 1997. 271–280
3 Rieser V, Lemon O. Learning and evaluation of dialogue strategies for new applications: empirical methods for optimization from small data sets. Computat Linguist, 2011, 37: 153–196
4 Larsen L B. Issues in the evaluation of spoken dialogue systems using objective and subjective measures. In: Proceedings of 2003 IEEE Workshop on Automatic Speech Recognition and Understanding, St Thomas, 2003. 209–214
5 Yang Z J, Levow G, Meng H. Predicting user satisfaction in spoken dialog system evaluation with collaborative filtering. IEEE J Select Topics Signal Proc, 2012, 6: 971–981
6 Asri L E, Laroche R, Pietquin O. Task completion transfer learning for reward inference. In: Proceeding of AAAI Workshop on Machine Learning for Interactive Systems, Quebec, 2014. 38–43
7 Ultes S, Minker W. Quality-adaptive spoken dialogue initiative selection and implications on reward modelling. In: Proceeding of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Prague, 2015. 374–383
8 Su P H, Gašić M, Mrkšić N, et al. On-line active reward learning for policy optimisation in spoken dialogue systems. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, 2016. 2431–2441
9 Young S, Gašić M, Thomson, B. POMDP-based statistical spoken dialogue systems: a review. Proc IEEE, 2012, 101: 1160–1179
10 Hirschman L, Thompson H S. Overview of evaluation in speech and natural language processing. In: Survey of the State of the Art in Human Language Technology. New York: Cambridge University Press, 1997. 409–414
11 Watambe T, Araki M, Doshita S. Evaluating dialogue strategies under communication errors using computer-to-computer simulation. IEICE Trans Inform Syst, 1998, E81-D: 1025–1033
12 Ai H, Weng F. User simulation as testing for spoken dialog systems. In: Proceedings of Annual Meeting of the Special Interest Group on Discourse and Dialogue, Columbus, 2008. 164–171
13 Schatzmann J. Statistical user and error modelling for spoken dialogue systems. Dissertation for Ph.D. Degree. Cambridge: University of Cambridge, 2008
14 Williams J. Applying POMDPs to dialog systems in the troubleshooting domain. In: Proceedings of the HLT/NAACL Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technology, New York, 2007. 1–8
15 Thomson B, Young S. Bayesian update of dialogue state: a POMDP framework for spoken dialogue systems. Comput Speech Language, 2010, 24: 562–588

16 Henderson J, Lemon O, Georgila K. Hybrid reinforcement supervised learning for dialogue policies from communicator data. In: Proceedings of the IJCAI Workshop on Knowledge and Reasoning in Practical Dialog Systems, Edinburgh, 2005. 68–75
17 Gašić M, Lefevre F, Jurčiček F, et al. Back-off action selection in summary space-based POMDP dialogue systems. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, Merano, 2009. 456–461
18 Jurčiček F, Keizer S, Gašić M, et al. Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk. In: Proceedings of Interspeech Conference, Florence, 2011. 3061–3064
19 McGraw I, Lee C, Hetherington L, et al. Collecting voices from the cloud. In: Proceedings International Conference on Language Resources and Evaluation, Malta, 2010. 1576–1583
20 Gašić M, Jurčiček F, Thomson B, et al. On-line policy optimisation of spoken dialogue systems via live interaction with human subjects. In: Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding, Hawaii, 2011. 312–317
21 Black A, Burger S, Conkie A, et al. Spoken dialog challenge 2010: comparison of live and control test results. In: Proceedings of Annual Meeting of the Special Interest Group on Discourse and Dialogue, Portland, 2011. 2–7
22 Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, 2002. 311–318
23 Banerjee S, Lavie A. METEOR: an automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, 2005
24 Lin C Y. Rouge: a package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, 2004. 25–26
25 Rus V, Lintean M. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In: Proceedings of the 7th Workshop on Building Educational Applications Using NLP, Stroudsburg, 2012. 157–162
26 Wieting J, Bansal M, Gimpel K, et al. Towards universal paraphrastic sentence embeddings. arXiv: 1511.08198
27 Forgues G, Pineau J, Larcheveque J M, et al. Bootstrapping dialog systems with word embeddings. In: Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, Cambridge, 2004
28 Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: Proceedings of International Conference on Neural Information Processing Systems, Lake Tahoe, 2013. 3111–3119
29 Charlin L, Pineau J. How NOT to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. arXiv: 1603.08023v2
30 Anjuli Kannan, Oriol Vinyals. Adversarial Evaluation of Dialogue Models. arXiv: 1701.08198v1
31 Lowe R, Noseworthy M, Serban I V, et al. Towards an automatic turing test: learning to evaluate dialogue responses. 2017. In press
32 Lowe R, Serban I V, Noseworthy M, et al. On the evaluation of dialogue systems with next utterance classification. In: Proceedings of Annual Meeting of the Special Interest Group on Discourse and Dialogue, Los Angeles, 2016
33 Lowe N, Pow N, Serban J I, et al. The ubuntu dialogue corpus: a large dataset for research in unstructured multi-turn dialogue systems. In: Proceedings of Annual Meeting of the Special Interest Group on Discourse and Dialogue, Prague, 2015
34 Ritter A, Cherry C, Dolan B. Unsupervised modeling of twitter conversations. In: Proceedings of Annual Conference on North American Chapter of the Association for Computational Linguistics (NAACL), Los Angeles, 2010. 172–180
35 Banchs R E. Movie-dic: a movie dialogue corpus for research and development. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Cincinnati, 2012

Survey of evaluation methods for dialogue systems

Wei-Nan ZHANG, Yangzi ZHANG & Ting LIU*

Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin 150001*, China*

* Corresponding author. E-mail: tliu@ir.hit.edu.cn

Abstract This paper introduces the history of dialogue systems and their evaluation methods. The evaluation methods are categorized as either task-oriented dialogue systems or open domain dialogue systems. This paper investigates and summarizes the different methods of evaluating dialogue systems, analyzes the pros and cons of the different methods, discusses the emphasis of each method, and presents the progress of recent research for the two categories. For task-oriented dialogue systems, this paper introduces the recent research results of Steve Young. In addition, this paper sums up several widely used evaluation approaches. The evaluation methods for open domain chatting systems are explored from two angles: objective index evaluation and simulated artificial scoring. The various indices and different research ideas are analyzed and introduced as well. Finally, through summarizing and analyzing classical evaluation methods of dialogue systems as well as the newer evaluation methods based on neural network models, this study aims to predict developmental trends in evaluation methods for dialogue systems.

Keywords evaluation methods for dialogue systems, open domain dialogue systems, task-oriented dialogue systems, natural language processing, artificial intelligence

Wei-Nan ZHANG is a lecturer at the Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. His research interests include human-computer dialogue systems, natural language processing, and information retrieval.

Yangzi ZHANG is a graduate student at the Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. Her research interests include the evaluation of human-computer dialogue systems and dialogue system user simulation.

Ting LIU is a professor at the Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. His primary research interests are in natural language processing, information retrieval, and social computing.