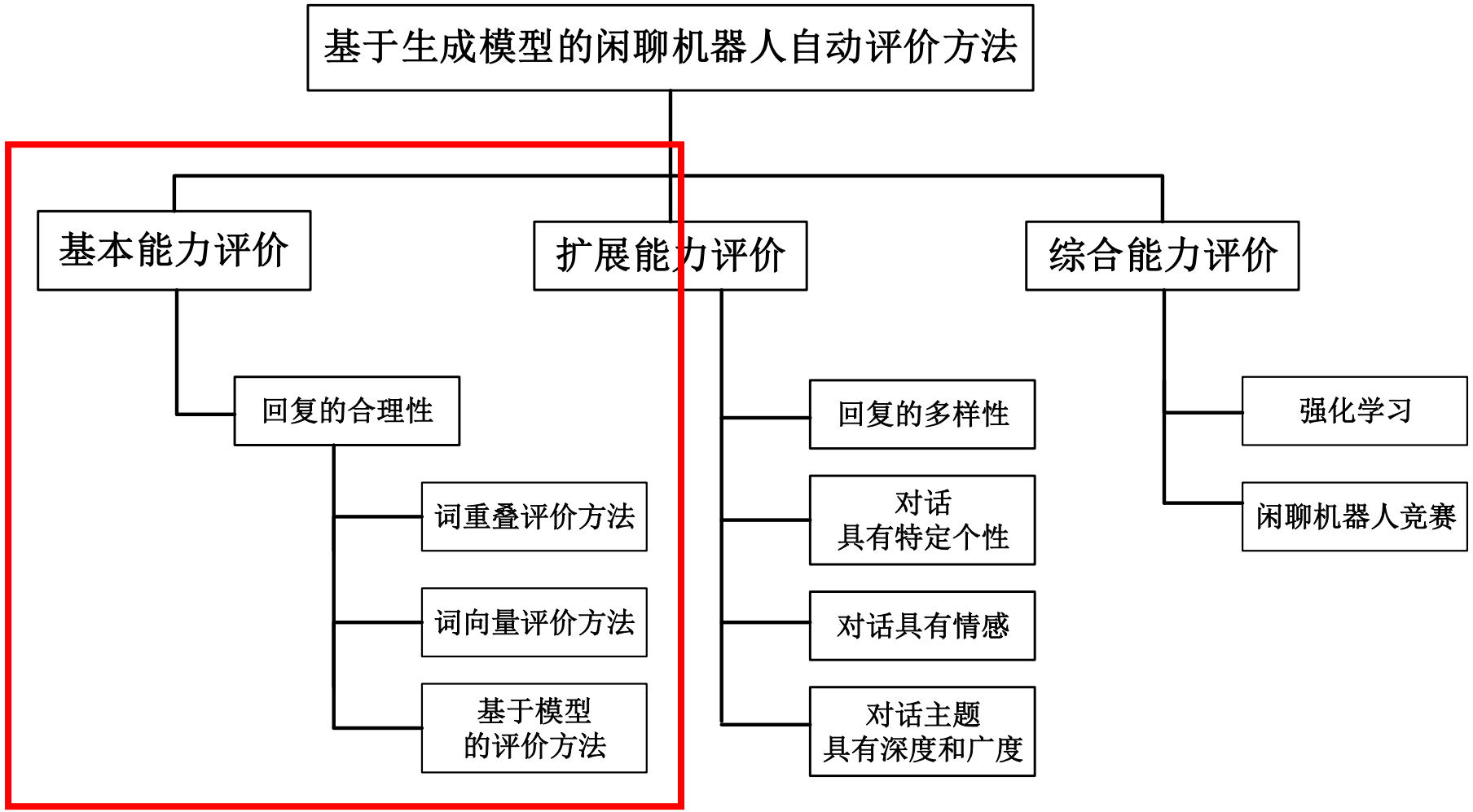
生成式闲聊机器人的自动评价方法

张璐

2019.12.27


基于生成模型的闲聊机器人自动评价方法
基本能力评价
扩展能力评价
综合能力评价
回复的合理性
词重叠评价方法
词向量评价方法
基于模型
的评价方法
回复的多样性
对话
具有特定个性
对话具有情感
对话主题
具有深度和广度
强化学习
闲聊机器人竞赛

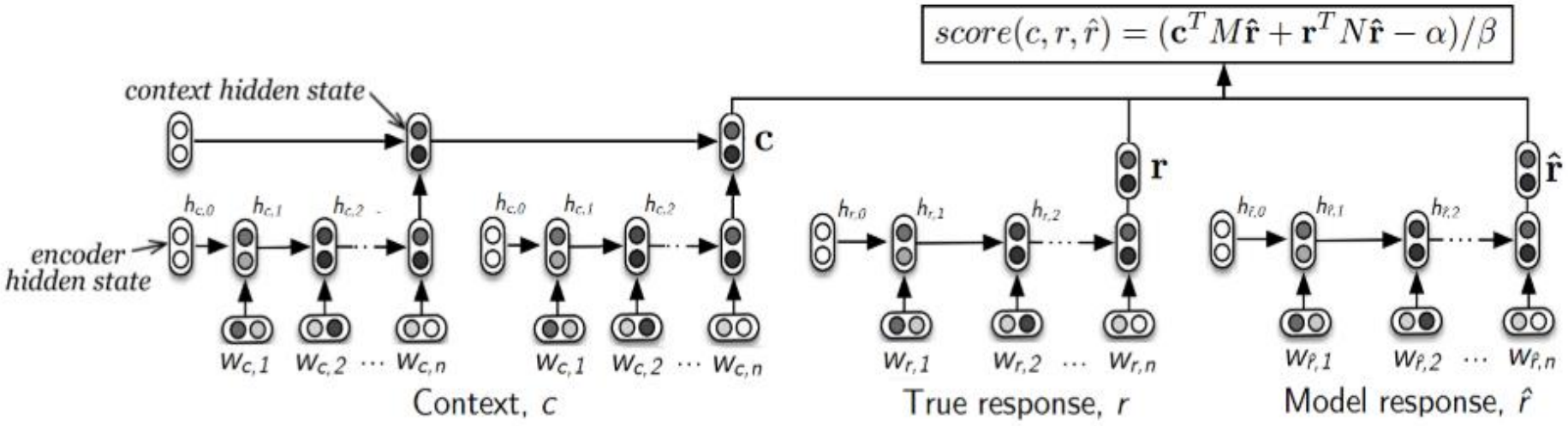
基本能力评价-评价回复的合理性				
类别		代表方法	方法概述	是否需要标注数据
词重叠评价方法		BLEU	计算生成回复与参考回复相同的n-gram数，相同的n-gram数越多，生成回复越合理	不需要
词向量评价方法		向量均值法	计算生成回复与参考回复的语义相似度，语义相似度越高，生成回复越合理	不需要
基于模型的评价方法	模仿打分模型	ADEM	使用具有人工评分的标注数据训练打分模型，使得模型能够模仿人类给生成回复打分	需要
	对抗模型	判别器	借鉴生成对抗网络，训练对抗模型，使得对抗模型能够区分闲聊机器人生成回复和人类产生回复	不需要
	直观打分模型	RUBER	一方面计算生成回复与参考回复的相似性，另一方面训练模型判断生成回复与问题的相关性，然后结合两个评价维度给出最终分数	不需要
其他		PPL	计算生成回复的生成概率，概率越大，PPL值越小，生成回复越合理	不需要

对话历史
对话者A: 你今天晚上想做什么? 对话者B：我们去看电影吧！ ?
参考回复
我不想看电影，我们做点别的吧
闲聊机器人生成回复
1. 别了吧，这段时间没什么好电影 2. 好啊，听说《少年的你》很好看

基本能力评价-评价回复的合理性				
类别		代表方法	方法概述	是否需要标注数据
词重叠评价方法		BLEU	计算生成回复与参考回复相同的n-gram数，相同的n-gram数越多，生成回复越合理	不需要
词向量评价方法		向量均值法	计算生成回复与参考回复的语义相似度，语义相似度越高，生成回复越合理	不需要
基于模型的评价方法	模仿打分模型	ADEM	使用具有人工评分的标注数据训练打分模型，使得模型能够模仿人类给生成回复打分	需要
	对抗模型	判别器	借鉴生成对抗网络，训练对抗模型，使得对抗模型能够区分闲聊机器人生成回复和人类产生回复	不需要
	直观打分模型	RUBER	一方面计算生成回复与参考回复的相似性，另一方面训练模型判断生成回复与问题的相关性，然后结合两个评价维度给出最终分数	不需要
其他		PPL	计算生成回复的生成概率，概率越大，PPL值越小，生成回复越合理	不需要

模仿打分模型-ADEM

- 构建数据集，数据集的形式为{问题-回复-人类评分}，因此首先要收集人类评分。
- 模型设计，层级RNN结构
- 评价方法评价，皮尔逊相关系数和斯皮尔曼相关性系数



对抗模型

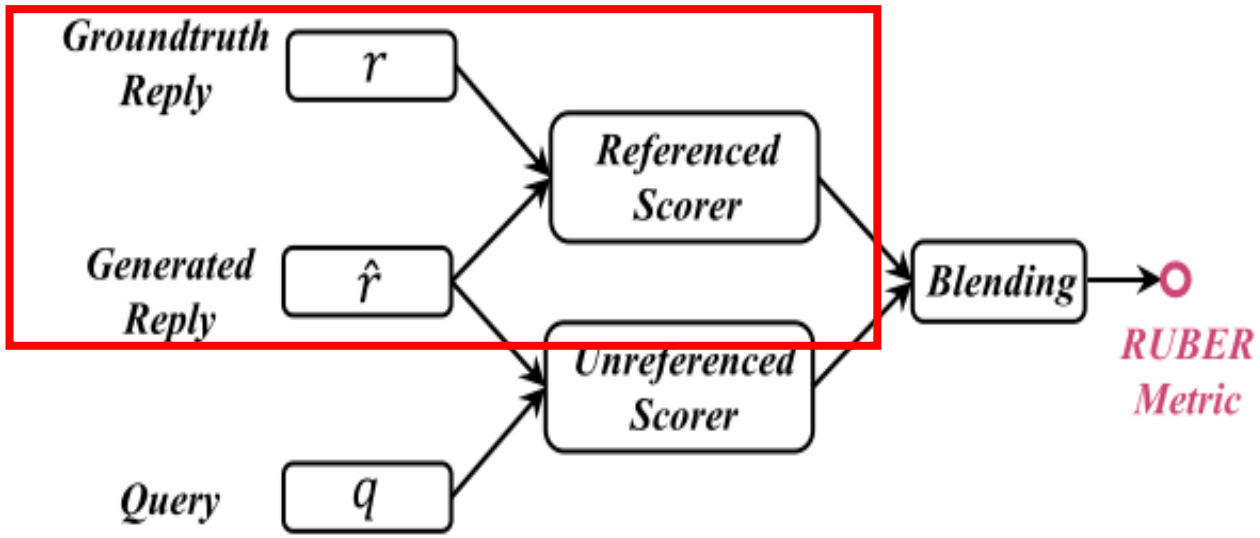
- 实验设计

  实验一：将数据集中一半对话对保留原始回复，一半对话对中的回复替换为随机回复。给定上下文，判断给出的回复是原本的回复还是随机的回复。

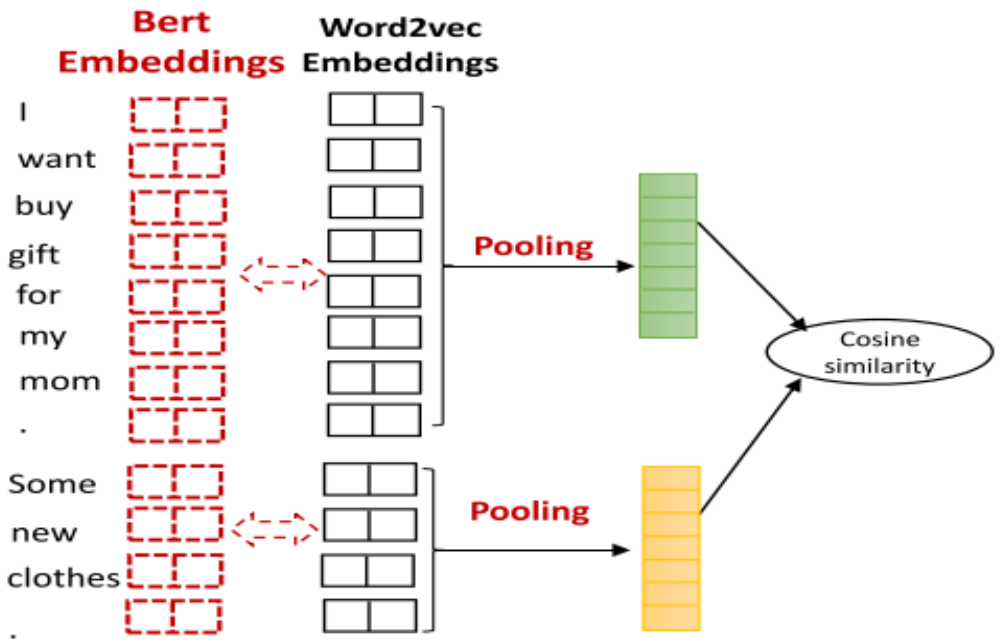
- 实验二：首先训练一个闲聊机器人，闲聊机器人对数据集中的问题生成回复，然后判断回复是闲聊机器人生成的回复还是语料中的参考回复
- 模型设计

  分类器

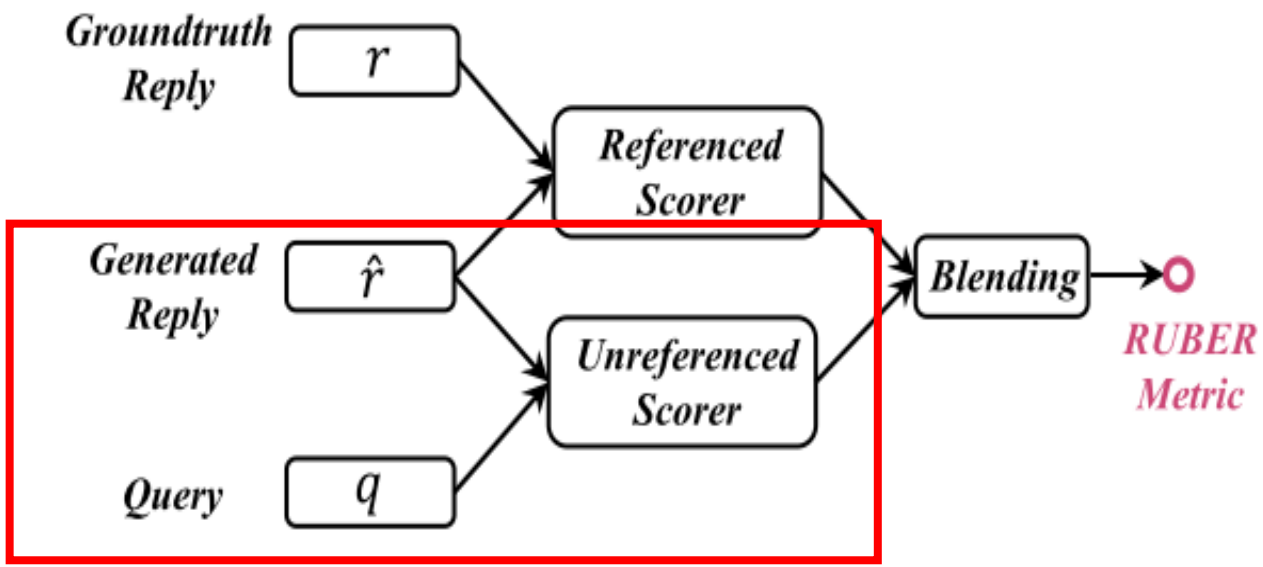
直观打分模型

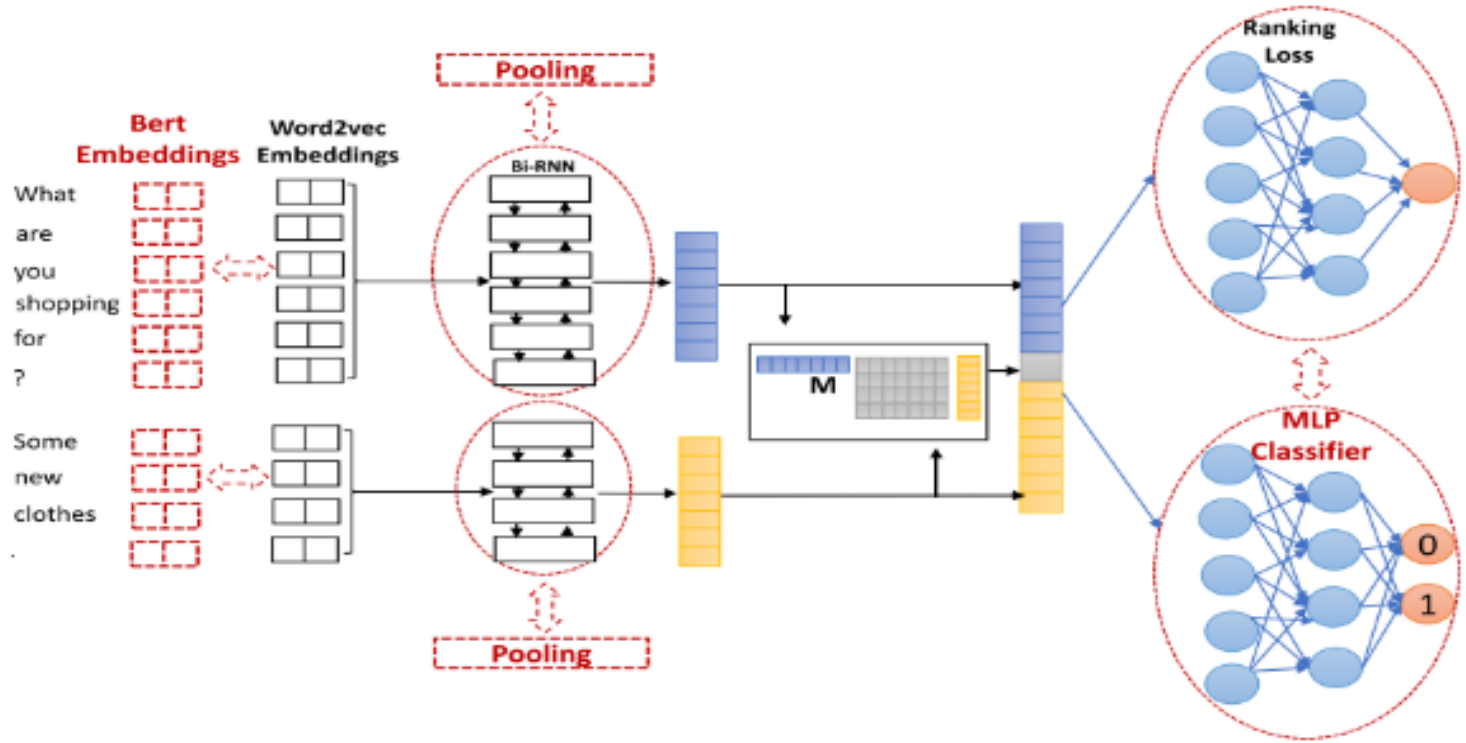

Groundtruth Reply
r
Generated Reply
$\hat{r}$
Query
q
Referenced Scorer
Unreferenced Scorer
Blending
RUBER Metric

直观打分模型

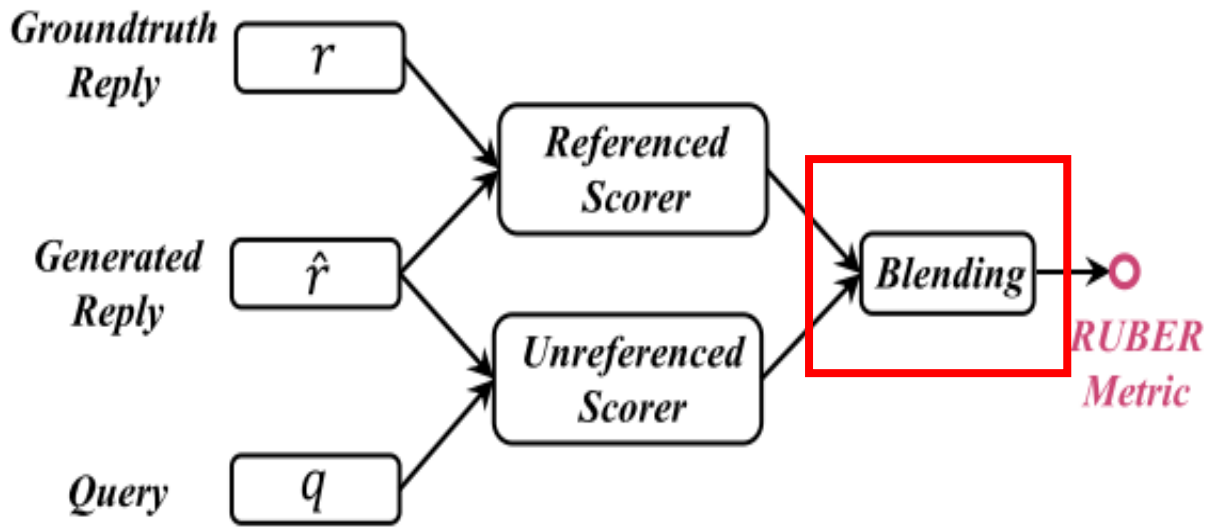

Bert Embeddings
Word2vec Embeddings
I
want
buy
gift
for
my
mom
.
Some
new
clothes
.
Pooling
Pooling
Cosine similarity

直观打分模型

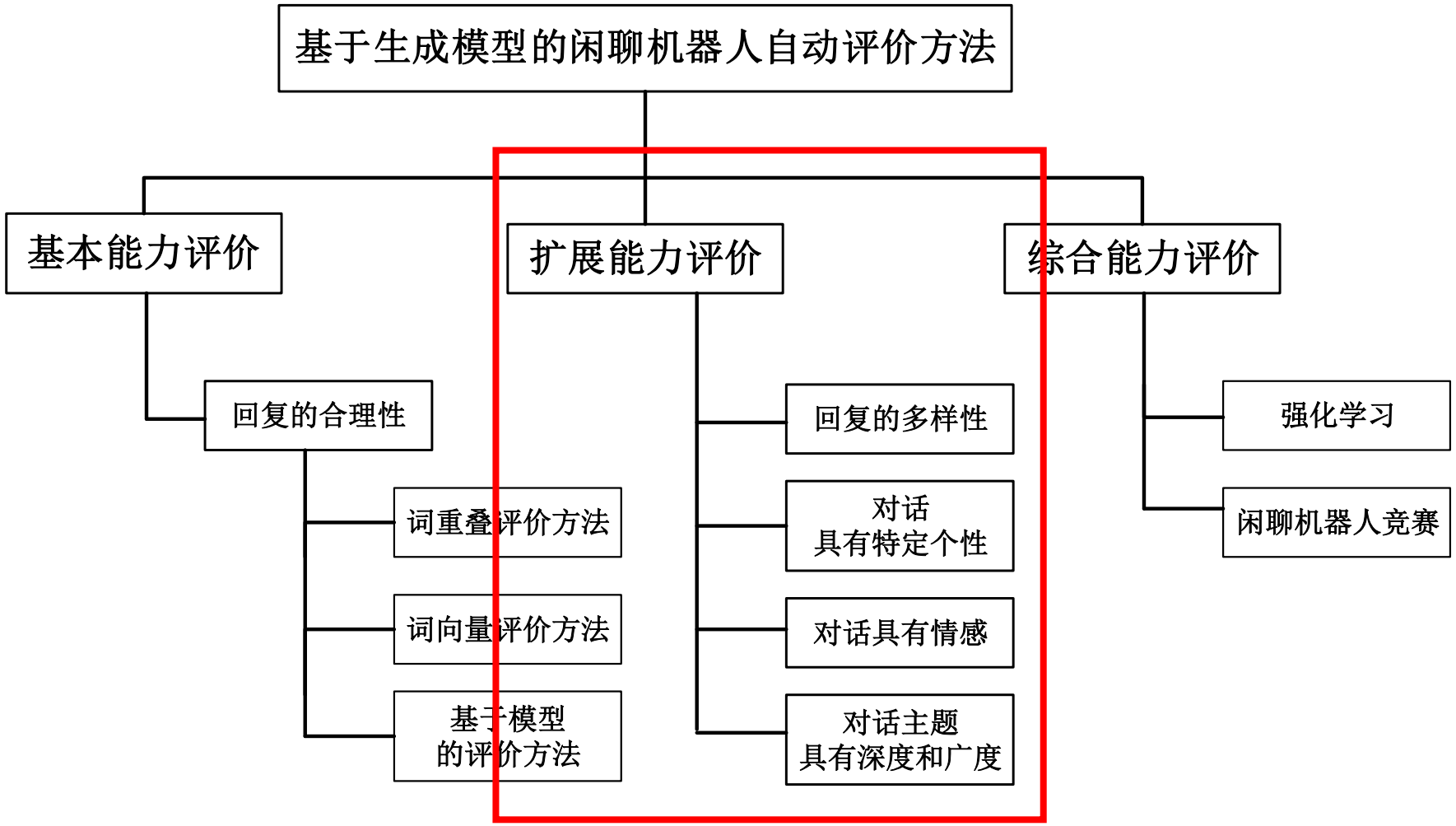

Groundtruth Reply
r
Referenced Scorer
Generated Reply
$\hat{r}$
Blending
RUBER Metric
Unreferenced Scorer
Query
q


Pooling
Bert Embeddings
Word2vec Embeddings
Bi-RNN
What
are
you
shopping
for
?
Some
new
clothes
.
M
Pooling
Ranking Loss
MLP Classifier
0
1

直观打分模型



- 启发式策略:最大，最小，算数平均和几何平均


基于生成模型的闲聊机器人自动评价方法
基本能力评价
扩展能力评价
综合能力评价
回复的合理性
回复的多样性
强化学习
词重叠评价方法
对话
具有特定个性
闲聊机器人竞赛
词向量评价方法
对话具有情感
基于模型
的评价方法
对话主题
具有深度和广度

回复的多样性

- 生成回复的平均长度
- distinct-1，distinct-2
- 熵

闲聊机器人的个性特质

- 构建回复具有个性标注的数据集：{问题-回复-回复体现的个性}

  OCEAN个性识别器：开放性（Openness），尽责性 （Conscientiousness），性格外向（Extraversion），令人愉快（Agreeableness）和神经质 （Neuroticism）

- 评价闲聊机器人的个性特质

  闲聊机器人在测试集上生成回复后，首先使用个性识别器对生成回复体现的个性进行识别，然后再与之前对参考回复识别出的个性进行比较。

闲聊机器人的情感表达

- 构建回复具有情感标注的数据集，{问题-回复-回复体现的情感}

  BiLSTM情感分类器：生气、厌恶、开心、喜欢、伤心和其他

- 评价闲聊机器人的情感表达

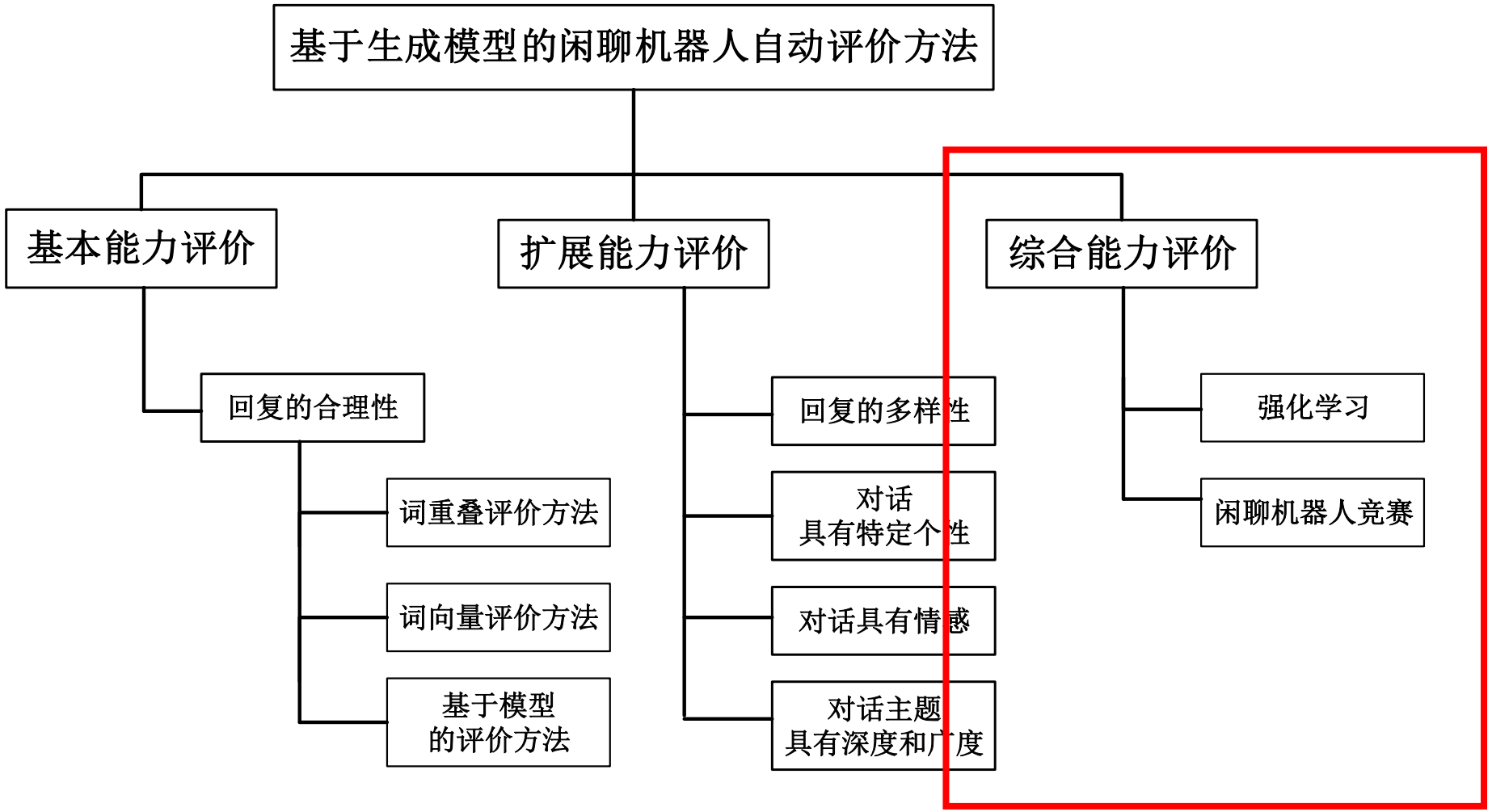
在测试集上，使用情感分类器对闲聊机器人的生成回复的情感进行预测，然后计算参考回复的情感类别与生成回复的预测情感类别之间的一致性，得到情感准确性，情感准确性越高，代表该闲聊机器人的情感表达能力越好。

闲聊机器人对话主题的深度和广度

- 主题广度是衡量闲聊机器人在不重复表达的情况下就各种粗粒度和细粒度主题进行交谈的能力的评价指标。
- 主题深度是衡量闲聊机器人在给定主题上保持长时间连贯对话的能力的评价指标。
- 使用主题分类模型识别每句话语的主题，主题关键字检测器识别每句话语的主题关键字
- 根据定义的概念，进一步定义评价指标

概念	定义
特定主题的轮数T	属于同一主题的{用户话语-机器人回复}对
主题一致的子对话S	连续的同一主题对话，至少包含两轮
子对话的长度l_s	子对话包含的同一主题的轮数
连续对话的长度l_c	讨论不同主题的总轮数

类别		评价指标		定义
主题深度		对话级别	平均主题深度	一次对话中所有子对话的平均长度
		系统级别	平均主题深度	所有对话中所有子对话的平均长度
主题广度	粗粒度主题	对话级别	粗粒度主题广度	一次对话中出现不同主题的数量，其中每个主题至少持续了一个子对话
		系统级别	粗粒度主题广度	对话级别的粗粒度主题广度的平均值
			粗粒度主题数量	所有对话中包含主题的子对话的数量
			粗粒度主题频率	将不同主题的子对话的数量进行归一化
	细粒度主题词	系统级别	主题关键字覆盖率	所有对话中不同关键字的总数
			主题关键字数量	所有对话中的所有关键字的总数
			主题关键字频率	同一关键字的总数比上所有关键字的总数


基于生成模型的闲聊机器人自动评价方法
基本能力评价
扩展能力评价
综合能力评价
回复的合理性
词重叠评价方法
词向量评价方法
基于模型
的评价方法
回复的多样性
对话
具有特定个性
对话具有情感
对话主题
具有深度和广度
强化学习
闲聊机器人竞赛

奖励函数

评价级别	评价指标	评价方法
话语级别	容易回答	通过使用生成无趣回复的负对数似然函数来衡量生成话语是否容易回答
	信息流	提出对闲聊机器人连续产生话语的余弦相似度进行惩罚的方法
	语义连贯性	使用生成回复与上下文的互信息来确保生成回复是连贯的、合理的
系统级别	合理性	将回复的合理性检测看作是分类问题，将回复的合理性分为：合理的，一般的，不合理的三类。使用具有人工评价的标注数据训练分类器
	对话主题深度	讨论同一主题的连续话语数量反映了对话主题深度。同样将对话主题深度检测看作分类问题，对话主题深度可分为“浅”、“中级”和“深”三个等级，使用人工标注数据对分类器进行训练
	信息增益	对话语进行分词后，统计不同单词的数量

闲聊机器人竞赛

评价指标	评价方法	是否需要标注数据
对话 用户体验	平均用户评分	需要人工评分数据
	平均常用用户评分	需要人工评分数据
连贯性	回复错误率	需要回复连贯与否的 标注数据
参与度	参与度评价者评分	需要人工评分数据
	对话时间中位数	不需要
	对话轮数中位数	不需要
域覆盖	反向变化系数R-COV	需要人工评分数据
主题广度	单词表大小	需要主题标注数据
	平均频率	需要主题标注数据
主题深度	平均深度	需要主题标注数据

问题与展望

- 构建标准闲聊评价数据集
- 提高人工评价的可信度
- 研究多轮交互式评价方法

谢谢！