



迈向创造性语言生成: 汉语幽默自动生成的探索

谭红叶^{1,2*}, 闫真¹, 李茹^{1,2}, 敬毅民¹

1. 山西大学计算机与信息技术学院, 太原 030006

2. 计算智能与中文信息处理教育部重点实验室, 太原 030006

* 通信作者. E-mail: hytan.2006@126.com

收稿日期: 2018-06-20; 接受日期: 2018-09-27; 网络出版日期: 2018-11-14

国家自然科学基金 (批准号: 61673248, 61772324) 和山西省“1331 工程”重点学科建设计划资助项目

摘要 幽默生成是计算创造性任务之一, 能够赋予计算机一定的个性化与创造性, 而且可以提升用户体验. 本文以笑话的生成进行汉语幽默生成的探索性研究. 首先提出一个符合当前自然语言生成技术的笑话生成任务: 给定笑话的主体部分, 生成相应的笑点句. 然后, 尝试了基于经典编码器-解码器框架的方法与基于生成对抗网络的方法来完成该任务. 为了克服编码器-解码器框架中对幽默特点没有建模的局限, 本文在生成对抗网络方法中融入了歧义性、不一致性、语音相似性、普遍性等笑话属性特征来评价、指导笑话的生成. 实验结果表明: 在生成对抗网络方法中融入笑话属性特征后, 系统输出构成笑话的比例提升 6 个百分点. 尽管从总体来看系统自动生成的笑点句构成笑话的比例还偏低, 但本文通过对幽默生成问题的研究探索, 带动了对创造性语言生成问题的洞察与理解, 标志着我们向创造性语言生成的探索迈进了一步.

关键词 幽默生成, 笑话生成, 深度学习, 编码器-解码器框架, 生成对抗网络

1 引言

随着人工智能与计算创造 (computational creativity, CC) 技术的发展, 人们不仅期望计算机能以人类智能的方式完成任务, 还期望计算机能像人一样进行写作、绘画等方面的创造性活动. 目前计算创造性任务主要集中在音乐、视觉、语言文字等方面的创作. 其中, 语言文字创作包括: 隐喻生成、故事创作、菜谱生成、诗歌创作及幽默生成等.

幽默是人类特有的复杂认知行为, 不仅体现了人类的创造性思维, 而且在人际交往中扮演着重要角色, 如: 化解冲突、使问题更容易解决、增加个人魅力与独特性等. 如果计算机具备了人类的幽默感, 就具有了一定的创造性与个性化特点, 同时也使得计算机在完成复杂任务时能够提升用户体验. 如已

引用格式: 谭红叶, 闫真, 李茹, 等. 迈向创造性语言生成: 汉语幽默自动生成的探索. 中国科学: 信息科学, 2018, 48: 1497–1509, doi: 10.1360/N112018-00158
Tan H Y, Yan Z, Li R, et al. Towards creative language generation: exploring Chinese humor generation (in Chinese). Sci Sin Inform, 2018, 48: 1497–1509, doi: 10.1360/N112018-00158

有研究表明, 幽默在虚拟招待员这样的人机交互系统中引入了笑话, 使得总体对话量增加了 50% [1]. 因此幽默的生成值得我们关注.

幽默包括语言型幽默 (textual or language humor) 与非语言型幽默 (non-language humor) [1]. 前者指通过语言文字表现的幽默 (如: 笑话、双关语等), 是日常生活中最常见的幽默; 后者指通过语言之外的方式 (如: 动作、表情等) 表现的幽默. 本文主要进行语言型幽默的生成研究 (以下简称为幽默生成).

幽默生成的研究相对较少, 而且大多关注特定形式、以词语选择为中心任务的幽默生成 [2~5]. 有少数研究者探索开放形式的幽默生成, 如: Ren 等 [6] 尝试使用神经网络方法, 生成笑话中的每一个词, 但高质量笑话生成的比例较低, 只有 12.1%. 这在一定程度上反映出完整的幽默生成是一个比较困难的问题.

笑话是幽默的一种主要形式, 一般由主体 (set-up) 与笑点 (punchline, 也称妙语) 两部分构成 [7]. 其中, 主体部分是笑话的基础, 包含读者的结果预期; 笑点部分是笑话中使人发笑的句子或词语, 一般是笑话的结尾.

本文提出一个汉语笑话的笑点句生成任务: 给定一个笑话的主体部分, 要求系统生成对应的笑点句来达成幽默. 本文将笑点句的生成看作一个序列到序列的自然语言生成问题, 并尝试了两种基于深度学习的方法. 第 1 种方法基于经典的编码器 – 解码器框架, 编码器与解码器均采用 LSTM (long short-term memory) 模型. 该方法能够依据给定的上下文, 实现笑点句的生成. 但这种方法使用最大似然估计作为目标函数, 偏向于生成通用或者训练集中频繁出现的句子作为回复, 而且没有对幽默自身的特点进行建模刻画. 为了缓解这些局限, 本文又尝试了第 2 种方法: 基于生成对抗网络 (generative adversarial networks, GANs) 的方法构建生成器. 该模型包含生成器 G 与判别器 D 两个构件, 生成器 G 用来生成笑点句, 判别器 D 用来评价 G 的生成结果, 其输出被看作是生成器 G 的期望奖励函数. 与此同时, 本文还将笑话的属性特点融入到期望奖励函数, 显式地对幽默质量进行评价, 弥补了原有模型对幽默特点刻画不足的缺陷.

相关实验结果表明: 基于 GANs 得到的汉语笑点句生成模型在各项指标上均取得了较优的结果. 将笑话属性特点融入到期望奖励函数后, 系统又得到进一步提升, 与基于编码器 – 解码器框架方法相比, 笑话构成的比例提升 6 个百分点, 表明本文所提第 2 种方法行之有效. 尽管从总体来看, 自动生成的笑点句与主体部分构成笑话的比例还偏低, 但本文所做工作使我们对幽默生成问题有了比较深的认识与理解, 促进我们向创造性语言生成的探索迈进了一步, 为后续计算创造研究工作的深入开展奠定了基础.

2 相关研究

语言学中的幽默包含两大类: 言语性幽默 (verbal humor) 与指称性幽默 (referential humor) [8,9]. 其中言语性幽默主要指借助谐音、多义等语言形式或语言特征产生的幽默. 这类幽默在翻译过程中经常因为难以保留引起幽默感的语言特征而失去幽默感. 指称性幽默主要是通过语言产生的意义来实现幽默效果, 翻译后幽默感仍得以保留.

人们在 Aristotle 时代就开始从社会、情绪和认知等不同角度研究幽默的产生机制. 迄今为止, 主流的 3 种理论有: 优越论 (superiority theories)、释放论 (release theories) 与不一致论 (incongruity theories, 又称矛盾论) [10]. 优越论从社会的角度理解幽默, 认为幽默的产生是由于感知者突然意识到自己的超过别人的优越感. 释放论从情绪的角度进行幽默理解, 认为人们之所以意识到是幽默并发笑, 是因为人通过后天教育和社会影响所形成的情绪和心理压抑在幽默世界里得到了释放. 不一致论认为读者

对故事的预期与故事结局存在不一致是幽默产生的关键,这种观点后来被形式化为幽默的语义脚本理论 (semantic script-based theory of humour, SSTH)^[7] 与言语幽默的通用理论 (general theory of verbal humour, GTVH)^[11]. 不一致论被研究者广泛认可,基于不一致论的幽默生成研究主要通过制造上下文的不一致来达成幽默效果.

幽默的生成属于自然语言生成 (natural language generation, NLG) 与计算幽默 (computational humor) 研究领域. 计算幽默是人工智能领域中一个较新的研究分支,主要研究利用计算机对幽默建模的技术与方法,具体包括:幽默识别与幽默生成两类研究. 与幽默识别相比,幽默生成的研究相对较少.

目前大部分幽默生成面向特定类型、特定形式的幽默进行,其核心任务是选择词语填充到固定结构的槽中. 例如: JAPE^[2] 使用少量的模式生成了双关语式的幽默 (punning riddle); HAHAcronym^[3] 对现有的缩写生成新的幽默解释; Petrović 等^[4] 从名词-属性相似度、属性奇异度、属性歧义度、名词差异度 4 个方面提出 4 个假设,使用无监督方法发现可笑的三元组并进行模板填充来生成 “I like x like I like y, z” 形式的幽默; Aggarwal 等^[5] 生成海地语三行诗形式的笑话; Valitutti 等^[12] 尝试利用满足某种约束的单词替换生成幽默文本; Özbal 等^[13] 利用 WordNet 与 ConceptNet 生成了创造性的幽默新词.

少部分学者研究了指称式幽默的生成. Binsted 等^[14] 通过夸张语句生成指称式幽默; Tinholt 等^[15] 故意曲解代词意义来产生幽默; Labutov 等^[16] 通过对 ConceptNet 进行层次聚类作为概念不一致的度量来生成对立脚本; Ren 等^[6] 探索了基于神经网络方法的给定主题后的完整幽默的生成,但取得的效果不是很好,只有 12.1% 的幽默被认为是好的笑话. 还有一些学者尝试在较大的应用背景下进行创造性语言生成的探索. 如: Du 等^[17] 提出娱乐性对话的生成,并且在翻译模型框架下探索了汉语传统的幽默艺术形式——相声的生成,重点关注捧哏的应答生成. 但捧哏的应答并不要求每一句都具有幽默效果,如: 有的应答只是重复上文,起到承接逗哏、烘托逗哏的作用.

与已有的幽默生成研究不同的是,本文尝试汉语开放式笑话 (不限定笑话格式) 的生成,主要关注在给定笑话主体部分的情况下,生成与主体语义连贯的笑点句,从而使整个文本形成一个笑话.

3 任务描述

笑话是语言型幽默的一种主要形式,本文主要通过笑话生成来探索研究幽默的生成. 根据 Raskin^[7] 提出的幽默语义脚本理论,笑话一般由主体与笑点句两部分构成. 其中,主体是笑话的基础,蕴含读者的结果预期;笑点句是笑话中使人发笑的句子或词语,一般是笑话结尾. 表 1 给出了一些笑话示例,其中例 1 与 2 属于一句话笑话 (one-liner joke),最后一个分句为笑话的笑点句,其余部分为笑话的主体部分. 例 3 与 4 是单轮对话式笑话,说话者 A 与 B 的讲话内容分别对应笑话的主体与笑点句部分.

本文提出一个汉语笑话笑点句的生成任务: 给定长度为 m 的主体片段: $s = [w_0, \dots, w_i, \dots, w_{m-1}]$, 生成一个长度为 n 的笑点句片段 $s' = [w'_0, \dots, w'_j, \dots, w'_{n-1}]$, 使 s 与 s' 构成的整个文本形成一个有趣的笑话.

我们认为,上述的笑点句生成任务更适合在当前语言智能技术条件下评价系统的幽默生成能力. 主要原因是: 与完全的开放式幽默生成相比,笑点句生成任务相对简单,但该任务又比特定形式的幽默生成任务具有挑战性. 具体表现为: (1) 幽默大多由短故事构成,但目前的 AI 与 NLG 技术尚不具备高质量故事的生成能力,更别提完整的幽默生成了. 而笑点句的生成无需过多考虑故事结构与情节的设计,因此更加符合当前技术发展情况. (2) 笑点句大多数时候是一个完整的句子,其生成是句子生成,而不仅仅是特定幽默生成中的词语选择或槽填充问题. (3) 一个好笑话要求笑点句同时满足流畅、语

表 1 笑话的构成示例
Table 1 Examples of jokes construction

No.	Joke examples	
1	同样是又馋又懒,	(set_up)
	熊猫和猪告诉我们长相有多么重要.	(punchline)
2	长得丑怎么了, 我自己又看不到,	(set_up)
	恶心的是你们.	(punchline)
3	A: 你觉着我长这样儿是不是应该去整整容了?	(set_up)
	B: 我觉着可以直接人道毁灭了.	(punchline)
4	A: 为什么深海鱼都很丑?	(set_up)
	B: 因为太黑谁也看不见谁, 大家就随便长长喽.	(punchline)

义连贯及滑稽可笑这 3 个条件. 其中, 语义连贯与滑稽可笑是具有竞争排斥性的两个指标. 因为笑点句既要与主体部分语义连贯, 又要通过语义不一致来达成滑稽可笑, 所以要同时满足这 3 个条件并不简单. 除此之外, 笑点句生成的数据集构建起来比较容易, 能够支持复杂模型的训练与学习. 由于笑点句一般在笑话结尾, 通常是笑话的最后一个分句或句子, 因此可以通过去除笑话中最后一个分句或句子, 来构建较大规模的数据集.

4 模型

本文尝试了两种基于深度学习的方法进行笑话笑点句的生成: 基于编码器 – 解码器框架的方法与基于生成对抗网络的方法.

4.1 基于编码器 – 解码器框架的笑点句生成模型

编码器 – 解码器框架由 Cho 等^[18]提出, 框架中的编码器将输入序列转化成一个固定长度的向量后, 解码器负责将该向量转化成输出序列. 该框架已经成功用于解决机器翻译、对话聊天等序列到序列问题.

本文将笑点句生成看作一个序列到序列的自然语言生成问题, 基于编码器 – 解码器框架完成该任务. 其中, 输入序列是笑话的主体部分, 输出序列是笑点句.

图 1 所示为基于该框架的笑点句生成模型. 为了解决长距离依赖问题, 编码器与解码器均采用对文本时序关系具有良好建模能力的 LSTM 模型. 编码器是一个双层的 LSTM 网络, 将前向 LSTM 最后一个隐层状态与反向最后一个隐层状态拼接, 形成输入序列主体的向量表示.

为了避免编码器将整个输入序列信息压缩进一个固定长度的向量所引起的信息表示不完全, 以及先输入的信息被后输入信息稀释掉的问题, 我们使用了注意力 (attention) 机制, 来产生一个“注意力范围”, 使解码器输出的时候可以重点关注输入序列的部分内容, 来产生下一个输出. 具体采用文献 [19] 中的 local attention model 框架. 解码器也是一个双层的 LSTM 网络, 依据笑话主体的向量表示逐渐解码生成笑点句的每一个词.

4.2 基于生成对抗网络的笑点句生成模型

生成对抗网络 GANs 由 Goodfellow 等^[20]于 2014 年提出, 其本质是生成模型, 但生成器的训练

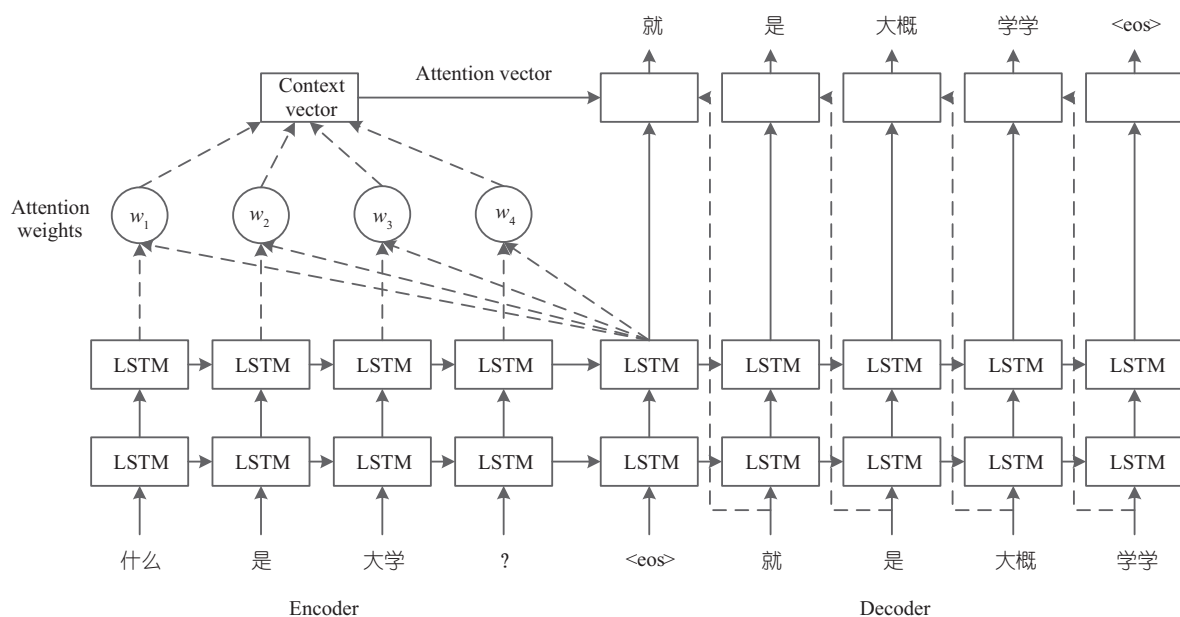


图 1 基于编码器 - 解码器框架的笑点句生成示意图

Figure 1 Punchline generation based on the encoder-decoder framework

学习是通过与判别器的对抗竞争进行的. 训练过程中, 生成器不断捕捉训练集真实样例的概率分布来生成新样例, 并努力使新样例骗过判别器; 而判别器努力学习如何区分真实样例与生成器生成的虚假样例, 并不断向生成器进行反馈, 指导其提升性能. GANs 的目标函数如下:

$$\min_G \max_D J(D, G) = E_{x \sim P_{\text{data}}(x)} [\log D(x)] + E_{y \sim P_{\text{model}}(y)} [\log(1 - D(G(y)))], \quad (1)$$

其中, $D(x)$ 表示判别器认为 x 是真实样本的概率, $G(y)$ 表示生成器在输入 y 时的输出, $1 - D(G(y))$ 表示判别器认为生成器生成的样本为假的概率, P_{data} 表示真实样例的数据分布, P_{model} 表示生成器产生输出的数据分布.

训练过程中, 判别器希望目标函数最大化, 而生成器希望目标函数最小化, 这种“对抗机制”使 GANs 可以免去复杂目标函数的设计, 而通过判别器的评估覆盖高质量生成结果来获得高质量的生成器.

GANs 在图像生成领域取得成功. 为了将 GANs 应用到自然语言中的序列到序列问题, 研究者提出一些方法来克服文本生成中的离散问题, 使判别器的输出可以反向传播到生成器. 如: Lamb 等^[21]提出将生成器的中间隐层向量 (而不是序列输出) 提供给判别器来解决该问题, 之后将 GANs 用在字一级的语言建模、手写体生成中取得了良好的效果. Yu 等^[22]通过增强学习中的策略梯度方法来解决 GANs 框架中的反向传播问题, 在诗歌生成、对话生成与音乐生成中使用 GANs 后取得了效果提升. 随后, 研究者们将 GANs 分别用到对话生成、图像标题生成等任务中^[23,24], 生成效果均得到提升.

本文在幽默生成中采用的 GANs 思想与文献 [23] 类似. 但由于幽默是人类最具创造性的行为之一, 并不遵循通常的语义一致性原则, 导致从有限的幽默数据集中学习生成高质量幽默非常困难. 因此, 本文对生成的笑话是否为好笑话进行显式评价, 将其引入到生成器训练过程中的期望奖励函数中, 通过最大化该函数来指导生成器的学习. 这是本文与文献 [23] 的主要不同.

(1) 笑点句生成的 GANs 模型. 图 2 为本文采用的 GANs 框架. 其中, s 为给定的笑话主体, s' 为

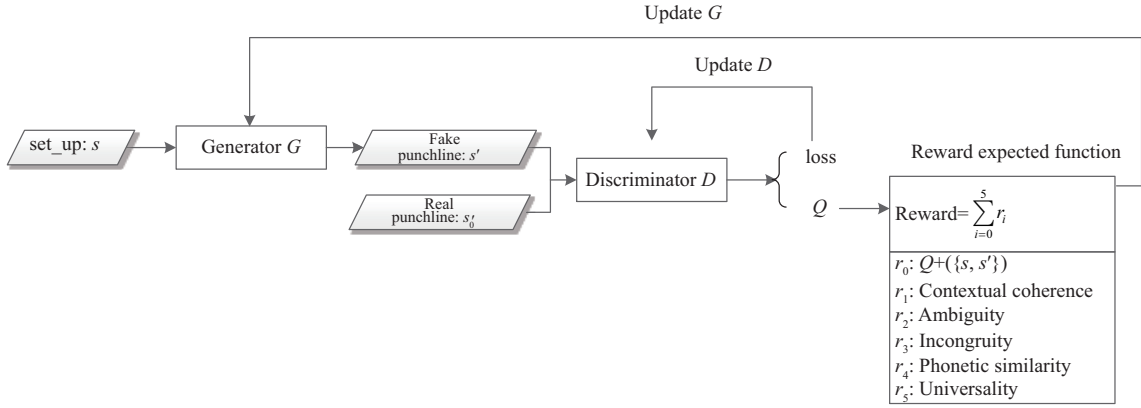


图 2 笑点句生成的 GANs 示意图

Figure 2 Punchline generation based on GANs

生成的笑点句, s'_0 为真实样例中 s 对应的笑点句. 生成器 G 采用与图 1 所示类似的编码器-解码器框架, 负责基于给定的笑话主体 s 生成笑点句 s' . 判别器 D 是一个二元分类器, 负责判别对真实笑点句 s'_0 与生成的笑点句 s' 进行真假样例的判别. 判别器 D 具体使用层次结构的神经网络 (hierarchical neural network) [25], 第 1 层为 LSTM 层, 第 2 层为 RNN 层. LSTM 层由 2 个相互独立的 LSTM 构成, 分别接收真实笑点句和生成器生成的笑点句作为输入, 每个 LSTM 都针对输入得到最后一个隐层状态进行拼接, 然后将结果输入到 RNN 中进行二分类.

与文献 [23] 类似, 模型训练时采用策略梯度方法进行. 训练过程中判别器 D 会将它的梯度回传给 G 和 D . 为了实现梯度回传, 目标函数必须可导. 具体地, 生成器 G 的目标函数及梯度计算公式分别如式 (2), (3) 所示 [23]:

$$J(\theta) = E_{s' \sim p(s'|s)}(Q_+(\{s, s'\})|\theta), \quad (2)$$

$$\begin{aligned} \nabla J(\theta) &\approx [Q_+(\{s, s'\}) - b(\{x, y\})] \nabla \log \pi(s'|s), \\ &= [Q_+(\{s, s'\}) - b(\{x, y\})] \nabla \log \pi(s'_t|s, s'_{1:t-1}), \end{aligned} \quad (3)$$

其中, $Q_+(\{s, s'\})$ 表示在给定主体句 s 的情况下判别器将生成的笑点句 s' 判断为真实样例的概率. π 是生成 s' 的概率, $b(\{x, y\})$ 是基准值, 其目标是减少估值的方差同时使估值无偏, 是一个确保训练稳定的平衡项.

本文的判别器 D 不仅区分生成器 G 的输出与真实样例, 而且还对笑话从语义连贯性、好笑性等方面进行打分, 并将输出作为奖励反馈给生成器, 以指导生成器产生的输出不但接近于人的输出, 且满足语义连贯性、好笑性. 具体的期望奖励函数在论文以下部分详述.

(2) 融入笑话特点的期望奖励函数. 一个好的笑话应该满足流畅性、语义连贯性和好笑性这几个特点. 已有工作表明, 基于编码器-解码器框架的序列到序列模型能够产生流畅的句子, 因此本文假设目前的模型已经满足流畅性的要求, 仅考虑将连贯性与好笑性特点融入期望奖励函数, 实现这两方面的显式评价.

• **语义连贯性.** 受文献 [26] 启发, 本文使用互信息来评估笑话主体 s 与笑点句 s' 的连贯性, 具体计算如下:

$$r_1 = \log P(s'|s) + \log P_{\text{backward}}(s|s'), \quad (4)$$

其中, $P(s'|s)$ 表示在给定笑话主体 s 的情况下, 生成笑点句 s' 的概率. $P_{\text{backward}}(s|s')$ 表示基于笑点句 s' 生成主体 s 的后向概率. $P_{\text{backward}}(s|s')$ 的训练方式与标准序列到序列的模型类似, 都是源与目标交换的方法.

• **滑稽好笑性.** Yang 等^[27] 从歧义性、不一致性、人际关系效应、语音风格等角度分析了幽默的潜在语义结构. 本文参照 Yang 等的分析, 结合汉语幽默特点从歧义性、不一致性、语音相似性、普遍性等来评价笑话的滑稽好笑性.

(i) 歧义性. 歧义是笑话的一个重要组成部分^[28]. 所以本文假设如果有很多歧义词被用在笑话中, 那么这个笑话更有趣可笑. 具体计算如式 (5) 所示. 其中, $\text{senses}(w_i)$ 表示笑话中的词具有不同义项的数量. 我们将关注点放在名词、动词、形容词和副词上, 使用 HowNet 来获得这些词的义项.

$$r_2 = \frac{1}{k} \log \left(\prod_{i=1}^k \text{senses}(w_i) \right). \quad (5)$$

(ii) 不一致性. 在幽默理解的各种理论中, 最普遍接受的理论是不一致性理论. 它认为接收者的期待与结局之间的冲突或矛盾是幽默构成的必要条件. 但迄今为止, 还没有确切的不一致性定义. 本文使用 Word2Vec 对主体与笑点句进行分布式表示, 通过二者之间的语义相似度来衡量不一致性, 并假设如果主体和笑点句不相似, 那么笑话就会更可笑有趣. 主体与笑点句之间的语义相似度计算如下:

$$r_3 = \max_{w_s \in s, w_{s'} \in s'} \{\cos(V_{w_s}, V_{w_{s'}})\}. \quad (6)$$

(iii) 语音相似性. Mihalcea 等^[29] 指出笑话的语音属性是一个重要的特征, 它强化了不协调的效果. 本文主要考虑笑点句中的词与主体中的词押韵的情况, 即: 两个词的尾字韵母相同、音调相同, 且不是同一个字, 则认为这两个词押韵. 具体计算如式 (7) 所示, 其中, $\text{rhymes}(s, s')$ 表示构成笑点句的词与主体中的词押韵的个数, N_s 与 $N_{s'}$ 分别为主体句与笑点句的总词数.

$$r_4 = \frac{1}{N_s + N_{s'}} \log(\text{rhymes}(s, s')). \quad (7)$$

(iv) 普遍性. 本文认为如果词语更常用更普遍, 那么笑话更有趣可笑. 普遍性通过式 (8) 计算, 其中 $P(w_i)$ 和 $P(w_i w_{i+1})$ 分别表示笑点句中的单词和双词在语料库中出现的概率. 实验中, 共收集到将近 20 万条笑话, 其中有 21245 条用于深度学习模型的训练与测试 (数据集具体情况见 5.1 小节), 剩余的近 18 万条用来计算公式中的 n -gram 值.

$$r_5 = \log \left(\sum_i^{n-1} P(w_i) + \sum_i^{n-2} P(w_i w_{i+1}) \right). \quad (8)$$

最终的奖励期望函数是各项奖励的加权总和, 如下所示:

$$r = w_0 \sum_{i=0}^5 r_i, \quad (9)$$

其中, r_0 为 $Q_+(s, s')$, 是判别器在给定主体句 s 的情况下将生成的笑点句 s' 判断为真实样例的概率, $\sum_i w_i = 1$, 实验中设置 $w_i = \frac{1}{6}$.

表 2 本文采用的数据集
Table 2 Datasets used in our models

Dataset	Size	Sources
Training data	500000	480000 pairs of conversation from Sina Weibo and 15000 one-liner jokes or single-turn jokes
Valid data	5000	One-liner jokes or jokes single-turn conversation
Test data	1000	One-liner jokes or jokes single-turn conversation

5 实验结果及分析

5.1 数据集

目前, 无论英文还是中文都没有公开的标准笑话生成评测数据集. 本文首先从多个网站收集笑话数据. 初始数据集中包含近 20 万条笑话, 其中有 83637 个单句笑话, 21089 个单轮对话式笑话和 94554 个多轮对话式或段落式笑话.

为获得较为准确干净的笑话主体 – 笑点句的数据集, 本文通过数据观察后, 主要利用单轮对话式笑话与其他形式的短笑话 (长度限制为 30 个词以内, 且包含多个分句) 构建数据集. 构建时将每个笑话中最后一个句子 (或分句) 作为笑点句, 其余部分作为笑话的主体部分. 最终的主体 – 笑点句数据集包含 21245 条笑话, 部分笑话样例如表 1 所示.

为了克服笑话数据集规模小的问题, 本文采用 48 万对新浪微博原文 – 评论数据^[30] 作为辅助数据进行训练. 该数据集与本文的问题有一定相似性, 主要表现在: (1) 微博的原文 – 评论数据与对话式笑话在形式上类似; (2) 微博与笑话都是短文本, 长度比较接近; (3) 微博的原文 – 评论包含一些具有幽默感的数据, 在内容生成上与笑话具有一定关联.

本文所采用的数据集具体情况如表 2 所示. 其中的测试集是经过人工检查与剔除噪声后, 保留语句通顺的样例得到的.

5.2 评价方法

自动评价. 幽默是人的一种复杂的创造性行为, 目前还缺乏有效的针对幽默生成的自动评价方法与指标. 本文采用大多自然语言生成任务都采用的 BLEU 值来评价生成的笑点句. 该指标对系统输出与参考答案之间的 n -gram 片段进行比较, 匹配片段数越多, 认为系统输出越好. 本文中 n 为 4.

人工评价. 系统生成的笑点句与给定的主体部分形成的文本如果语义连贯且滑稽好笑, 我们就认为构成一个笑话. 构成的笑话比例越高, 系统性能越好. 这里对每个样例是否语义连贯、是否滑稽好笑进行人工评价. 具体评价时, 每个样例文本由 3 个评价者进行判断. 首先对文本是否语义连贯进行评价: 如果样例文本具有语义连贯性, 则标记为 1, 反之则标记为 0. 然后, 对文本是否滑稽可笑进行评价: 如果样例文本滑稽可笑, 则标记为 1, 反之标记为 0. 最后, 选择两个标注者一致认为是笑话的样例作为真正的笑话. 表 3 所示为某个评价者的一个评价样例, 评价者认为该样例的主体部分与笑点句构成了一个笑话.

5.3 实验设置

预训练. 本文首先使用微博语料作为训练集对生成器 G 进行预训练, 从而使生成器 G 能够生成通顺的句子. 然后使用生成器 G 获得部分训练数据的输出, 得到训练判别器 D 所需的虚假样例, 并与

表 3 人工评价示例

Table 3 Examples of human evaluation

Set_up	Punchline	Coherence	Funniness
A: 单身是无罪的.	B: 你是不是说我是无罪的?	1	1

表 4 微博数据对系统的影响

Table 4 Impact of Weibo dataset on systems

Model	BLEU-4
SYS _{encoder-decoder} (with Weibo dataset)	11.79
SYS _{encoder-decoder} (no Weibo dataset)	10.04

笑话数据中的原笑点句 (即真实样例), 一起输入到判别器 D , 对判别器 D 进行预训练.

将训练集更换为笑话数据, 加载使用预训练好的生成器 G 和判别器 D 在对抗学习框架下进行训练. 生成器 G 的隐层大小设置为 200, 学习率为 0.001. 判别器的学习率设置为 0.0001, 并且每 5 个训练批次 (batch) 更新一次判别器 D .

5.4 实验结果及分析

本文在表 2 的测试集上对模型进行测试评价. 自动评价指标 BLEU 值的计算是针对测试集中 1000 个样例 (test-set₁₀₀₀) 进行. 由于人工评价成本比较高, 因此随机抽取 500 个样例 (test-set₅₀₀) 进行人工评价.

实验中, 将文献 [23] 中基于对抗学习的对话生成方法 (Adver-REGS) 作为基线系统, 该系统与文本所提的基于 GANs 并融入笑话属性特征的期望奖励函数的方法最接近. 表 4~6 中的 SYS_{encoder-decoder} 表示本文采用的第 1 种方法 (基于编码器 - 解码器框架的笑点句生成), SYS_{GANs-joke} 表示本文采用的第 2 种方法 (融入笑话属性特点的基于生成对抗网络的笑点句生成). SYS_{GANs-joke}(single) 与 SYS_{GANs-joke}(ensemble) 分别是单模型和集成模型. 单模型基于如式 (9) 所示的 6 项期望函数的加权总和进行期望奖励. 集成模型是同一架构不同参数模型的集成, 具体通过更改训练过程中的超参数, 得到了 6 个性能比较好的单模型, 然后对 6 个模型的结果使用语料库中的 bigram 值进行评分, 以得分最大的结果作为输出得到的一个模型集成.

表 4 给出了基于编码器 - 解码器框架模型在引入 48 万微博数据前后得到的结果, 可以看出: 加入微博数据后, 由于数据集规模加大, 系统性能得到了提升. 由于本文其他生成模型都基于编码器 - 解码器, 因此其他模型的训练也引入微博数据.

表 5 给出了不同模型在 test-set₁₀₀₀ 上获得的 BLEU-4 值情况, 表 6 给出了不同模型在 test-set₅₀₀ 上的人工评价结果. 可以看出: 基于对抗学习框架的模型由于引入了判别器的对抗竞争, BLEU-4 值与幽默构成比例均得到提升. 还可以看出, 在对抗学习框架中融入笑话属性奖励期望函数后, BLEU-4 值与幽默构成比例进一步得到提升. 其中, 幽默构成比例提升了 6 个百分点, 提升效果明显. 表 5 和 6 所示结果充分表明, 本文提出的笑话属性特征奖励期望函数对笑话的建模刻画具有一定效果.

但是, SYS_{GANs-joke}(ensemble) 集成模型取得的指标提升效果不是很明显. 主要原因在于: 目前集成模型是基于 bigram 值进行单模型的选择, 而 bigram 值与笑话生成的可笑性目标不是特别吻合. 未来还需要探索更合理的笑话生成的自动评价指标, 来优化集成模型.

我们选取了人工评价比较好的系统结果示例, 如表 7 所示. 可以看出, SYS_{encoder-decoder} 生成的笑

表 5 不同模型在 test-set₁₀₀₀ 上的结果
Table 5 Results of the different models on test-set₁₀₀₀

Model	BLEU-4
SYS _{encoder-decoder}	11.79
Adver-REGS	12.07
SYS _{GANs-joke} (single)	14.67
SYS _{GANs-joke} (ensemble)	14.8

表 6 不同模型在 test-set₅₀₀ 上的结果
Table 6 Results of the different models on test-set₅₀₀

Model	Ratio of jokes (%)
SYS _{encoder-decoder}	10.2
Adver-REGS	13.4
SYS _{GANs-joke} (single)	16.6
SYS _{GANs-joke} (ensemble)	16.8

表 7 系统的输出示例
Table 7 Examples of the system generation

No.	Set _{up}	Punchline
1	你的数学成绩超过你的体重了吗?	Human: 这真是个血腥的问题.
		SYS _{encoder-decoder} : 没唉.
		SYS _{GANs-joke} (single): 再胖下去也不对吧, 我认了.
2	你觉着我长这样儿是不是应该去整整容了?	Human: 我觉着可以直接人道毁灭了.
		SYS _{encoder-decoder} : 我知道了.
		SYS _{GANs-joke} (single): 我去, 你这辈子不要去医院了.
3	为什么深海鱼都很丑?	Human: 因为太黑谁也看不见谁, 大家就随便长长喽.
		SYS _{encoder-decoder} : 因为你丑.
		SYS _{GANs-joke} (single): 因为他们没结婚.

点句中的用词更普遍一些, 这是由于其使用最大似然估计作为目标函数, 偏向于生成训练集中频繁出现的回应而引起. SYS_{GANs-joke} 生成的笑点句显然更多样化一些, 这也恰恰符合对抗学习框架中利用判别器来实现接近真实样例的这样一个目标.

从总体来看, 系统自动生成的笑点句与给定主体构成笑话的比例还偏低, 仅有 16.8%. 这一结果与文献 [6] 中“只有 12.1% 的笑话为高质量英文笑话”的结论基本一致. 这些结果表明幽默的生成是一个复杂困难的问题. 主要原因为 (1) 幽默是人类一种复杂的创造性行为, 目前对幽默的达成缺乏准确详实的理论解释. 尽管不一致理论被研究者广泛接受, 但该理论仅仅指出不一致是幽默达成的必要条件 [8], 对幽默达成的充分条件、不一致是如何构成的都没有给出明确的形式化定义. 基于该理论的幽默生成系统通过故意制造不一致来达成幽默, 不仅没有生成幽默, 反而生成了许多语义不连贯的“胡言乱语”, 造成了系统输出与预期相差很远的局面. (2) 幽默大多由短故事构成, 而目前的 AI 技术及自然语言生成技术尚不具备高质量的故事生成能力, 那么生成具有创造性特点的幽默就更为困难了. (3) 幽默是主观的, 幽默的达成与接收者的心理预期、理解认知等方面都有关系. 因此, 很容易造成不

同的人对同一幽默的反应不同. 根据 Petrović 等^[4]与 Ren 等^[6]的研究, 仅有约 30% 的笑话被不同的标注者认为是可笑的. 因此, 研究者还需要在幽默理论、故事创作、幽默达成等方面继续深入研究, 才可能解决幽默的生成这一难题.

6 结论

幽默生成不仅可以使计算机具有一定的个性化与创造性特点, 而且还可提升用户体验. 幽默生成是计算创造与计算幽默领域的研究热点. 本文从汉语幽默的主要形式——笑话的生成进行幽默生成的探索性研究. 首先提出一个符合当前自然语言生成技术的笑话生成任务, 然后尝试了基于经典编码器-解码器框架的方法与基于生成对抗网络的方法来完成该任务. 为了克服编码器-解码器框架中没有对幽默特点建模的缺陷, 本文在生成对抗网络中的奖励期望函数中融入了歧义性、不一致性、语音相似性、普遍性等笑话属性特征来评价幽默的生成. 实验结果表明: 在生成对抗网络方法中融入笑话属性特征后, 系统输出构成笑话的比例提升 6 个百分点, 表明这种方法是行之有效的. 尽管从总体来看系统自动生成的笑点句构成笑话的比例还偏低, 但本文通过对幽默生成问题的研究探索, 带动了对于创造性语言生成问题的洞察与理解, 标志着我们向创造性语言生成的探索迈进了一步.

参考文献

- 1 Nijholt A, Andreea I N, Valitutti A, et al. Humor in human-computer interaction: a short survey. In: *Proceedings of Workshop on Designing Humor in Human-Computer Interaction*. Berlin: Springer, 2017. 527–530
- 2 Ritchie G. The JAPE Riddle Generator: Technical Specification. Informatics Research Report EDI-INF-RR-0158. 2003
- 3 Stock O, Strapparava C. HAHAcronym: a computational humorsystem. In: *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, Ann Arbor, 2005. 113–116
- 4 Petrović S, Matthews D. Unsupervised joke generation from big data. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, 2013. 228–232
- 5 Aggarwal S, Mamidi R. Automatic generation of jokes in hindi. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-Student Research Workshop*, Vancouver, 2017. 69–74
- 6 Ren H, Yang Q. Neural Joke Generation. Final Project Reports of Course CS224n. 2017
- 7 Raskin V. *Semantic Mechanisms of Humor*. Berlin: Springer, 1985
- 8 Ritchie G. Can computers create humor? *AI Mag*, 2009, 30: 71–81
- 9 Lin H F, Zhang D Y, Yang L, et al. Computational humor researches and applications. *J Shandong Univ (Nat Sci)*. 2016, 51: 1–10 [林鸿飞, 张冬瑜, 杨亮, 等. 幽默计算及其应用研究. *山东大学学报 (理学版)*, 2016, 51: 1–10]
- 10 Strapparava C, Stock O, Mihalcea R. Computational humour. In: *Emotion-Oriented Systems*. Berlin: Springer, 2011. 609–634
- 11 Attardo S, Raskin V. Script theory revis(it)ed: joke similarity and joke representation model. *Humor-Int J Humor Res*, 1991, 4: 293–348
- 12 Valitutti A, Doucet A, Toivanen J M, et al. Computational generation and dissection of lexical replacement humor. *Nat Lang Eng*, 2016, 22: 727–749
- 13 Özbal G, Strapparava C. Computational humour for creative naming. In: *Proceedings of the 3rd International Workshop on Computational Humor*, Amsterdam, 2012. 15–18
- 14 Binsted K, Ritchie G. Computational rules for generating punning riddles. *Humor-Int J Humor Res*, 1997, 10: 25–76
- 15 Tinholt H W, Nijholt A. Computational humour: utilizing cross-reference ambiguity for conversational jokes. In: *Proceedings of International Workshop on Fuzzy Logic and Applications*, Camogli, 2007. 477–483
- 16 Labutov I, Lipson H. Humor as circuits in semantic networks. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, 2012. 150–155

- 17 Du S K, Wan X J, Ye Y J. Towards automatic generation of entertaining dialogues in chinese crosstalks. 2017. ArXiv:1711.00294
- 18 Cho K, van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the Empirical Methods in Natural Language Processing, 2014
- 19 Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, Lisbon, 2015. 1412–1421
- 20 Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. In: Proceedings of Advances in Neural Information Processing Systems, 2014. 2672–2680
- 21 Lamb A, Goyal A, Zhang Y, et al. Professor forcing: a new algorithm for training recurrent networks. In: Proceedings of the 30th Conference on Neural Information Processing Systems, Barcelona, 2016. 4601–4609
- 22 Yu L T, Zhang W N, Wang J, et al. SeqGAN: sequence generative adversarial nets with policy gradient. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017. 2852–2858
- 23 Li J W, Monroe W, Shi T L, et al. Adversarial learning for neural dialogue generation. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2017
- 24 Shetty R, Rohrbach M, Hendricks L A, et al. Speaking the same language: matching machine to human captions by adversarial training. In: Proceedings of IEEE International Conference on Computer Vision, 2017. 4155–4164
- 25 Serban I V, Sordani A, Bengio Y, et al. Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016. 3776–3783
- 26 Li J W, Monroe W, Ritter A, et al. Deep reinforcement learning for dialogue generation. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2016
- 27 Yang D, Lavie A, Dyer C, et al. Humor recognition and humor anchor extraction. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, Lisbon, 2015. 2367–2376
- 28 Miller T, Gurevych I. Automatic disambiguation of english puns. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, Beijing, 2015. 719–729
- 29 Mihalcea R, Strapparava C. Making computers laugh: investigations in automatic humor recognition. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Vancouver, 2005. 531–538
- 30 Shang L F, Lu Z D, Li H. Neural responding machine for short-text conversation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, Beijing, 2015. 1577–1586

Towards creative language generation: exploring Chinese humor generation

Hongye TAN^{1,2*}, Zhen YAN¹, Ru LI^{1,2} & Yimin JING¹

1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China;

2. Key Laboratory of Ministry of Education for Computation Intelligence and Chinese Information Processing of Shanxi University, Taiyuan 030006, China

* Corresponding author. E-mail: hytan_2006@126.com

Abstract Humor generation is one of the tasks of computational creativity, which can not only make a computer creative and have a personality, but also improve user experiences. This paper explores the generation of Chinese jokes, the main form of humor. In particular, the following task is considered: given the setup of a joke, generate the corresponding punchline that is in line with current natural language generation technologies using one of two approaches. One is based on the encoder-decoder framework and lacks modeling of humor characters. The other is based on generative adversarial networks (GANs), in which four characters (ambiguity, incongruity, phonetic similarity, and universality) are introduced into the reward function to evaluate the generated jokes and supervise the generator. Experimental results show that the GANs approach with joke character rewards obtains promising improvements compared to the encoder-decoder framework, namely, extra six percentage points on the ratio of jokes. While the performance is insufficient, as a first step towards creative language generation, the insights obtained in the exploration will help us in future research.

Keywords generation of humors, generation of jokes, deep learning, encoder-decoder framework, generative adversarial networks



Hongye TAN was born in 1971. She received her Ph.D. degree in computer science from Harbin Institute of Technology. She is an associate professor and a master tutor. Her research interests include natural language processing, question answering, and information retrieval.



Zhen YAN was born in 1997. He is an undergraduate student at Shanxi University with majors in computer science and technology. His research interests include natural language processing and information retrieval.



Ru LI was born in 1963. She received her Ph.D. degree in computer science from Shanxi University. She is a professor and a Ph.D. tutor. Her research interests include natural language processing, question answering, and knowledge graph.



Yimin JING was born in 1996. He is an undergraduate student at Shanxi University with majors in computer science and technology. His research interests include natural language generation.