Fantastic Expressions and Where to Find Them: Chinese Simile Generation with Multiple Constraints Kexin Yang♠

Dayiheng Liu♠ † Wenqiang Lei♢ Baosong Yang♠ Xiangpeng Wei♠ Zhengyuan Liu♣ Jun Xie ♠ ♠Alibaba Group
♢National University of Singapore ♣Institute for Infocomm Research (I2R), A*STAR, Singapore {kexinyang0528, losinuris}@gmail.com

抽象的

明喻出现在通过与另一个概念（即车辆)进行字面错误但具有象征意义的比较来描述概念（即主旨)的创造性背景中。以前的工作将明喻生成形成为上下文无关的生成任务,重点关注明喻风格的转移或从给定的前缀编写明喻。然而,在这种设置下生成的文本可能是不合需要的,例如很难满足明喻定义（例如,丢失车辆)或难以满足人类希望的某些内容偏好（例如,通过明喻描述苹果的颜色）。我们相信,如果与预先指定的约束相结合,明喻可能会更加合格并且以用户为导向。

为此,我们引入了可控明喻生成（CSG）,这是一项新任务,要求模型生成具有多个明喻元素（例如上下文和载体)的明喻。为了促进这项任务,我们提出了 GraCe,包括 61.3k 个明喻元素注释的中文明喻。在此基础上,我们提出了一个 CSG 模型 Similor来对该任务进行基准测试,其中包括一个车辆检索模块Scorer ,用于在车辆未知的情况下获得给定期限的可解释的比较。统计和实验分析都表明,GraCe 的质量高于所有其他中文明喻数据集,在符号元素数量（8 比 3）、Is-明喻准确度（98.9%比 78.7%)以及不断增加的方面都表现出色。不可控和可控明喻生成的模型性能增益。同时,Similor 可以作为 CSG 的强大基线,尤其是 Scorer,它无需任何重新训练即可击败基于模型的检索方法。

1 简介

明喻被广泛使用并激发人们的创造力（Li et al., 2022）。根据修辞学的经典术语（Campbell， 1988）,一个明喻使用比较词（即比较器)在一个概念（即主旨)和另一个概念（即车辆)之间进行字面上错误的比较。它还通过检查它们是否具有共享属性（即地面)来确保该比较对具有象征意义（Tartakovsky 等人， 2019）。值得注意的是， ground 可以以显式或隐式的方式表达（Chakrabarty et al., 2020）。合格样品如图1所示。 “枫叶就像被烧红的火把。”有明确的依据,男高音“枫叶”与车体“火把”的颜色相似,都是“火红”,而“枫叶就像小手掌”。暗示它们具有相似的五角星形状。

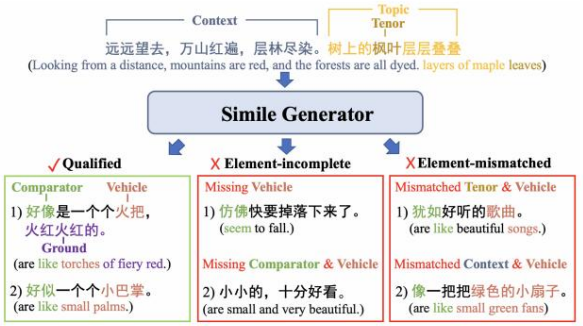


图 1:解释给定前缀的元素不完整和不匹配生成结果的玩具示例。
为非中文使用者提供翻译。

尽管明喻检测已被广泛探索（ Liu et al., 2018; Zeng et al., 2020; Mao and Li, 2021）,但明喻生成仍处于刚刚起步的阶段。现有的工作重点是上下文无关的明喻生成,包括:1)基于样式转移和2)基于前缀的明喻生成。前者将字面句子解释为明喻版本（Chakrabarty et al., 2020； Zhang et al., 2021） ,后者旨在根据预先指定的男高音写明喻（ Li et al., 2022； Chen et al., 2021）。 ., 2022)。

尽管取得了很大进展,但这样的实验设置可能会导致不良结果,例如不合格的明喻或无法满足内容

* 工作是在达摩院实习期间完成的 通讯作者。
†

计算语言学协会第 61 届年会论文集
第一卷:长论文,第 468–486 页
2023 年 7 月 9-14 日

数据集	# 数字	# 平均。	% Is-明喻	主题比较器	男高音车辆	宽/前	宽/前	地面	语境	上方/下方
诗歌（2019b）	43,051	23	-		/	/			/	
歌词(2019b)	246,669	23	-		/	/			/	
客户服务(2021)	5,490,721	61	29.3%		/	/			/	
中央军委(2022)	2,787	35	78.7%		/	/			/	
优雅	61,360	89	98.9%		/	/			/	

表1:现有汉语主要隐喻生成数据集的统计特征和标注信息
以及我们的 GraCe 数据集。表示数据集包含相应项目的注释，

则相反。 # 平均。表示每个句子的平均标记。 W 和 F 表示男高音/车辆词,并且
分别对应的特征词。 % Is-Simile 表示 1000 个明喻的平均百分比
从每个数据集中随机选择样本,并由三位专业注释者进行注释。我们忽略了
诗歌和歌词数据集,因为它们的文本样式与其他数据集不同。

人类愿望的偏好。如图1所示，
前者意味着生成的句子可能会丢失
不可缺少的明喻成分或生成不连贯成分,即生成成分
不完全
或 - 样本不匹配。例如，“枫
叶子小而美丽。”怀念两位男高音
和车辆和“枫叶就像小绿
粉丝们。”车辆“绿色粉丝”不一致
上下文“山是红色的”。当用户希望描述颜色时,可能
会出现第二个问题
枫叶的明喻,但得到“枫叶
就像小手掌一样”,尽管根据明喻的定义它是合格的。

为了解决这些问题,我们探索将各种约束纳入明喻生成
中。
具体来说,我们引入了控制标签明喻生成（CSG)的新任务 生成
明喻
具有来自给定前缀（即主题)的多个明喻元素（例如,车
辆、上下文等）。我们
收集细粒度注释的中文明喻
数据集 (GraCe),包含来自 26 万个经过清理的学生作文
文本的带注释的 61.3k 个相似字符。
如表1所示,我们常见的展开三种
带注释的元素（即男高音、载体和比较器） （Li et al.,
2022)增加到八个,例如可以将每个明喻放入更多的上下
文元素
自然使用情况（Sun et al., 2022）.1 In
为了更好地理解明喻比较,我们对细节进行了明确的注释。
至于隐式的
在此基础上,我们试图通过男高音和载体的认知属性来解
释它们之间的关系。这种属性是一组形容词,描述了相应属
性的显着特征。

名词（Veale和Hao， 2007）,这有助于从认知角度理解比较
语言学（Kövecses， 2010）。以南玻为标杆，
我们构建模型Similor,它首先检索
车辆（如果未知)由模块记分器（a
基于共享认知属性的检索方法
)对于给定的期限,然后结合所有约束和输入前缀（即主
题)来生成
比喻。统计和实验分析均表明 GraCe 的质量高于

以前的中文明喻数据集。同时,Sim ilor 可以成功地
将约束纳入
输出。特别是在车辆未知的设置中，
Scorer 击败了基于模型的检索方法
在自动和人工评估中无需任何
再培训.2

2 相关工作

不同于隐喻（于和万， 2019；
Chakrabarty 等人， 2021a； Stowe 等人， 2021 ）
使用隐式比较器,明喻更容易
被安置在。然而,现有的工作主要集中在明喻检测（Liu et al.,
2018； Zeng et al.,
2020； Mao 和 Li， 2021）,使比喻的产生尚未得到充
分探索。之前关于上下文无关的工作
明喻生成可分为:1)基于风格迁移的明喻生成和2)基于前
缀的明喻生成。第一个将这个任务解释为

将字面句改成明喻句,以及
自动将自我标记的明喻编辑为其文字版本,以构建（文
字句子，
比喻）。例如,SCOPE （Chakrabarty 等人，
2020)使用车辆的常识属性词（Bosse lut et al., 2019)将
其替换为
类似,然后删除比较器以形成
最后的字面句子。 WPS （Zhang 等人， 2021）

[1]详细注释参见附录图4 。

2我们的代码和语料库将在https://github上发布。
com/yangkexin/GraCe.

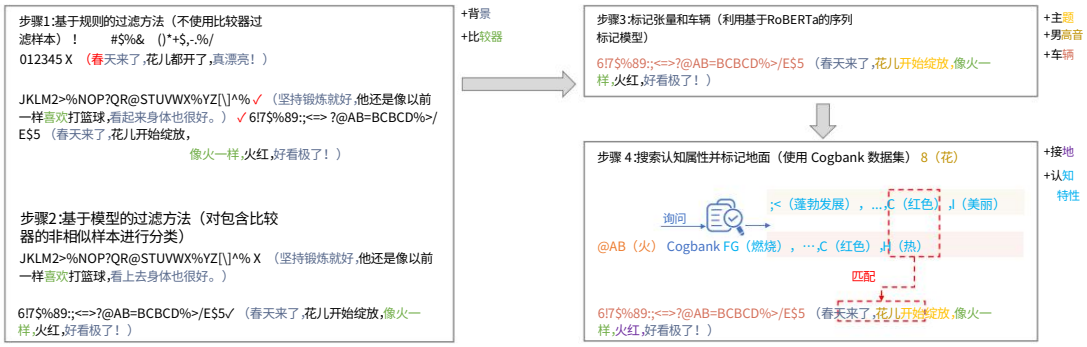


图 2:构建 GraCe 的流程。 “+”表示该元素在相应的步骤中被注释。

从明喻中删除一个范围以获得原义句子。第二个重点是从预先指定的期限生成比较器和期限。

刘等人。 (2019b)使用连续潜变量作为修辞控制器来生成中文诗歌。
CMC （Li et al., 2022)提供了一个多任务框架,利用未标记的数据来提高性能。陈等人。 (2022)使用三个词三元组（主旨、属性、载体)和关系模式来暗示生成明喻的模型。与所有这些不同,我们专注于可控明喻生成
生成具有多重约束的明喻。为了使其成为计算上可行的任务,我们构建了一个高质量的数据集 GraCe 和一个带有 Scorer 的 CSG 模型 Similor,以确保生成的明喻中存在可解释的主旨车辆对。如表1 所示,GraCe在收集样本数（ 61.3k vs 2.7k ） 、明喻质量（98.9% vs 78.7% Is-明喻准确率)和带注释元素的数量（八个与三个）.3

汉语句子,我们得到大约548万个句子。上下最多两句话

每个样本都用作上下文元素。

数据集处理如图2 所示,我们分四个步骤构建 GraCe 数据集。在步骤1中,我们过滤掉不包含比较相关单词的句子。具体来说,我们可以使用Jieba5工具包来标记句子,并过滤掉没有与比较器相关的单词的句子,因为比较器是明喻的标志。比较词多种多样,以确保明喻模式的多样性（例如“希望”、 “仿佛”、“大象”等,都表示“喜欢”）。然而,包含比较词的句子可能不会触发明喻（Liu et al., 2018）。正如步骤一中的例子2， “他还是像过去一样喜欢打篮球。(他仍然像以前一样喜欢打篮球。）” ,这里的“像（as）”意味着同一性而不是比较。因此,步骤 2 的重点是识别包含比较词的非明喻句子。我们训练基于 RoBERTaLarge （Liu et al., 2019a)的二元分类器,置信度分数为 80% 来选择明喻。6不,我们不追求更高的置信度分数,因为它可能面临减少明喻模式的风险。

3 Grace数据集

细粒度带注释的明喻数据集对于训练监督 CTG 模型和探索约束组合都很重要。然而,相关数据集（表1)可能还不够。

因此,我们提出了 GraCe 数据集,并详细阐述了数据集的创建和分析。

3.1 数据集创建

数据集收集我们收集了 260k 学生 com
来自免费访问网站的职位（从小学到高中的年级） ,4确保数据资源接近现实世界的案例。经过句子切分和去除非

经过上述两个步骤,我们得到了没有细粒度标注的明喻数据集。因此,第3步的目的是注释每个明喻的主旨、主题和载体。我们利用基于RoBERTaLarge的序列标记模型来注释每个明喻的主旨和载体。7同时,我们将主题注释为主旨和比较器之间的跨度,它表示主旨及其补充描述。之后,步骤4进一步旨在注释男高音和载体的基础和认知属性。作为解释

3%Is-Simile 的详细信息参见附录C.1。
[4] https://www.hxszww.com/

[5] https://github.com/fxsjy/jieba 6分类器详
细信息参见附录A.1 7标签模型详细信息参见附录A.2

测量	# Nums	# 平均代币数
句子	61,360	89.0
带注释的元素		
主题	61,360	11.4
男高音	61,360	1.9
期限财产	52,474	73.2
比较器	61,360	2.6
车辆	61,360	2.3
车辆财产	61,360	83.0
地面	15,087	8.6
上下文	57,543	39.5

表 2:GraCe 数据集的核心统计数据。这里
ground表示明喻中明确的根据。我们将隐式基础标记为之间的共享属性
男高音和车辆。

测量	价值
% 相似的	98.9
正确的期限 % 正确的车辆 %	95.2
98.2	
% 正确比较器	98.7
正确接地百分比	94.1

表3:随机抽取1000个样本的统计数据
来自三位专业注释者注释的 GraCe 。 98.9%的样本都是明喻。的统计数据
下面的虚线是针对这些比喻计算的。

进行比喻比较（Tartakovsky et al., 2019），
背景对于使明喻的男高音载体对易于理解和理解起着重要作用
具有象征意义（Campbell 和 Katz， 2006；
End， 1986），但在以前的数据集中被忽略。
我们首先查询 Cogbank 数据集8以获得主旨和载体的认知属性。然后，
它们的共同属性用于模糊匹配9
与财产有关的从句以明喻为理由。
最后是我们的 GraCe 数据集的详细统计
如表2所示，部分数据集样本为
见附录A.4。

3.2 数据集分析

数据质量我们邀请三位专业标注员对随机抽取的1000个样本进行多方面独立标注10。

表3中，只有1.1%的样本不是明喻，
这远远超出了其他中国明喻数据集
（见表1）。更重要的是保持高位
即使在明喻重要元素的细粒度注释中，准确性也很高(94.1% - 98.7%)。

[8] https://catalog.ldc.upenn.edu/LDC2020T01
[9] 请参阅附录A.3中的算法详细信息
10人工注释的详细信息参见附录C.1

测量	价值
# 独特的男高音	7,958
# 不同的车辆	5,350 人
# 不同的比较器	第371章

表 4:GraCe 数据集的独特统计数据。

明喻的多样性我们分析了明喻的多样性
进行比喻并在表 4 中呈现统计数据。首先，
期限和载体的丰富性确保了多样化
比喻的内容。而且和刘不同
等人。 （2018）；查克拉巴蒂等人。 （2020)仅使用
数据集中明喻的单一模式比较器
（即“_希望（喜欢）_”），我们构建
比较器作为 371 种填空模板。具体来说，受到 WPS 的启发（Zhang 等人，
2021)明喻的位置信息
上下文是一个强大的功能，我们通过以下方式将其合并
添加紧随其后的标点符号
车辆到我们的模板。如附录所示
图5 “_相似(like) _，”表示明喻部分
出现在中间子句中，车辆后面没有任何描述。如果模板中没有标点符号，
这意味着之后有明确的依据或上下文
车辆来补充内容。

4 可控明喻生成

4.1 任务定义

可控明喻生成任务的公式如下：给定一个包含主旨的主题x
st和各种预先指定的约束c，
模型通过以下方式生成明喻 y = (y1, y2, ..., yN)：

$$p(y|x, c) = \prod_{n=1}^{N} p(in|y<n, x, c; \theta), \quad (1)$$

其中θ是模型参数。值得注意的是，约束c可以自由选择和组合
候选集s = (sv, sp, sc)，表示
分别是车辆、比较器和上下文。

4.2 方法论

我们使用 CSG 模型 Sim 来对这个任务进行基准测试
ilor，其中包含用于车辆未知情况的模块 Scorer 。为了方便演示，我们
从一个玩具示例开始来说明它们。

类似如图3所示，主题“美”
丽的春天(the beautiful spring)” containing the
男高音“春天”首先与
分隔符的可选顺序约束
信号“[SEP]”。如果车辆已预先指定

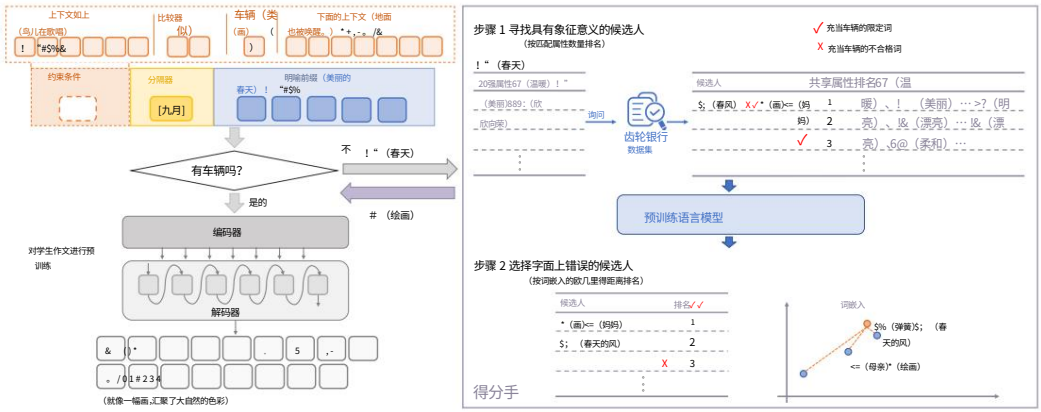


图 3:一个详细说明 Similor 和 Scorer 工作流程的玩具示例。

然后,输入序列被输入到编码器-解码器模型中。之后,模型自动回归生成“想要一件古董，它收集了精美的色彩。 （就像一幅画。它收集了大自然的颜色。）”。我们首先继续使用语言建模对象对收集的 26 万学生作文进行大型中文文本生成模型（例如，ChineseBART （Shao 等人， 2021））的预训练。然后,Similor 被实例化,以便在 GraCe 上进行微调。

记分器如果车辆未知,我们使用
记分器模块检索车辆然后添加

它到输入序列。如图 3 右侧所示,Scorer 包含两个步骤来获得具有象征意义但实际上是错误的男高音车辆对。步骤 1查询 Cogbank 数据集中的“春天”,以获得其前k 个最常用的认知属性。这些属性关系为候选车辆的选择和匹配提供了基础。 Cogbank 数据集（83,017 项)包含的词汇量比现代汉语常用词汇表11 （56,008 项)还要多,可以更全面地检索候选车辆。在实现中,选择认知属性数量与主旨相同的前20个名词作为候选,这确保了比喻意义的充分明喻,因为匹配的属性可以被视为基础。然而,在该步骤中也可以选择一些与文字相关的词语,例如“春风”。为了仅获得比喻项目,步骤 2根据每个项目和主旨之间的词嵌入的欧几里德距离对步骤 1 的候选重新排序。距离较远的候选者排名较高,因为它们在字面上的关联性较小

与男高音。结果， “画”被选为最终的载体。确切地说,给定 Cogbank 数据集中的第 i 个项目wi ,得到 tenor st ,排名分数 Scorecandii 的计算公式为：

$$\begin{aligned} \text{Scorewi} &= \text{Rank}(\text{F igwi}) + \text{Rank}(\text{Litwi}),\text{F} \\ \text{igwi} &= \text{Match}(\text{wi}, \text{st}), \\ \text{Litwi} &= \text{EucDist}(\text{wi}, \text{st})。 \end{aligned} \tag{2}$$

其中Rank(·)表示得到对应分数的排名。 Match(·)表示计算两个项目之间共享认知属性的数量， EucDist(·)表示它们的词嵌入之间的欧几里得距离。值得注意的是,我们使用排名来标准化这些分数,避免不同分数尺度的影响。

5 实验

在本节中,我们首先通过将 GraCe 数据集应用于基于前缀的明喻生成来实验性地评估GraCe 数据集的质量（第5.1 节）。由于这种不可控生成任务的设置不需要对训练样本进行额外注释,因此我们可以将 GraCe 与以前的中文明喻数据集进行比较。在此基础上,我们然后在新的 CSG 任务上评估所提出的类似方案（第5.2 节）。具体来说,我们首先比较受比较器和车辆约束的 Sim ilor 的不同模型品种,然后评估 Similor 在更广泛的约束下的性能。最后,我们探讨 Scorer 是否可以帮助 Similor 在车辆未知的设置中生成明喻。

5.1 GraCe的实验分析

由于统计分析不足以评估 GraCe,我们通过基于前缀的明喻生成来评估它。其中一个简单的管道是训练一个

11http: //www.moe.gov.cn/ewebeditor/uploadfile/2015/01/13/20150113085920115.pdf

数据集	% Comp.	↑明喻Conf. ↑	PPL ↓
主干:中文GPT2			
无 1.4 CS (2021) 46.0		0.3	40.9
CMC (2022) 44.4 GraCe		0.6	43.0
93.5		0.7	30.9
		0.9	10.9
骨干网:ChineseBART			
CS (2021) 65.3 CMC		0.5	33.1
(2022) 56.7 85.3格蕾丝		0.8	33.3
		0.9	28.7

表5:前缀生成的主要结果。 “没有任何”
表示使用主干模型生成句子
如果没有任何持续的培训,我们会忽略“无”
ChineseBART 的流畅度表现较差。 ↑表示
分数越高越好,而↓则恰恰相反。
最高数字以粗体显示。

数据集	Fluen. ↑	创建. ↑	Consi. ↑	总体↑
CS	2.5	1.9	1.9	2.1
CMC 2.2 格蕾丝		2.0	1.9	2.0
3.0		3.2	3.2	2.8

表 6:前缀生成的人工评估。

生成器与语言建模对象
相似数据集。由此推论,该模型要求
生成具有预先指定的基调的明喻。

基线和骨干。我们将提出的 GraCe 与之前的中文明
喻数据集进行比较:
1) CS （Zhang et al., 2021)包含从网络小说中提取的 549
万个类似内容。 2） CMC （李
et al., 2022)包含 2.7k 个隐喻和明喻
来自中国文学语料库。此外,我们利用
两种具有代表性的中文预训练语言
避免从头开始训练的模型:1)中国BART（CBART）
（Shao et al., 2021）： a BARTLarge
模型针对 200GB 中文文本进行了预训练
维基百科和五道语料库。 2)中文GPT2
(CGPT2) （Zhao et al., 2019 ）： GPT2Medium模型
在 CLUECorpusSmall 数据集上进行预训练。

实验设置。我们采用BARTLarge的原始超参数设置

GPT2Medium训练所有模型,使用 BERT 进行
kenizer （Devlin 等人， 2019)处理中文
文本。在推理过程中,我们使用 25 个常见的条件
作为前缀并要求模型继续写入
他们（每人 100 次完成）.12

指标。对于自动评估,我们首先使用
CGPT2 的困惑度 (PPL)来评估
文本质量。对于明喻评估,我们计算
含有比较词的句子比例
用于评估元素不完整的单词(%Comp.)
案例,因为这是明喻的标志。然而,包含比较词的句子可能

不会触发明喻（Liu et al., 2018）。因此,我们
使用明喻会议。评价比喻意义
生成的结果,即元素不匹配
案例。具体来说,我们重用明喻分类器
数据集处理的第 2 步（参见第3.1 节)计算每种方法
的平均置信度得分。
除此之外,我们还进行人工评估
继Chakrabarty 等人之后。 （2020）。 250个样品
从每个生成的结果中随机选择。
然后,三名众包评估员被要求
评分模型结果分为四类:1)流畅度
（流利。）。句子是否流畅、合乎语法； 2)创造力。句子写
得多好
具有象征意义； 3)一致性（Consi.）。
生成的车辆是否与预先指定的期限共享适当的联
系。 4)总体而言。如何
这个比喻总体来说好不好?分数是根据
其结构是否良好、富有创造性和一致性。
分数范围为 1 至 4,越高越好。13

结果前缀生成结果如图所示
表5中列出了人类评估结果,表 6 中列出了人类评估结果。
我们发现:1)模型经过微调
GraCe 在方面优于其他相似数据集
文本质量和比喻创造力。 2)生成性
语言模型倾向于产生字面意思的句子
强调明喻生成挑战的明喻,正如Chakrabarty 等人中也提
到的那样。
（2021b）。尽管模型可以通过前缀生成生成类似明喻的
句子,但这是不希望的
还获得了结果（例如,缺少比较器
并具有不连贯的男高音车辆对）,没有
控制明喻元素。14因此,有必要
探索一种新的明喻生成方法。

5.2 可控明喻生成

我们首先用不同的方法对 CSG 任务进行基准测试
模型品种受限于预先指定的比较器和车辆,然后探索
性能
不同约束组合下的相似性。最后,我们用 Scorer 来评估
Similor
在车辆未知的 CSG 设置中。具体来说,
给定一个包含男高音的主题,男高音载体
要求配对检索方法找到合适的车辆作为约束,然后提示类似

12单词列表和推理设置参见附录B.1

13关于分数定义和内部注释器的详细信息
协议见附录C.2。

14 生成样本如表13 所示。

方法	ROUGE-1/2/L↑	BLEU↑	BERTcore↑	ACC-V↑
CGPT2	20.7/4.2/18.3	0.3	60.6	16.4
CBART	21.3/10.9/20.9	1.7	55.9	71.1
CGPT2FT	22.2/7.6/20.2	3.0	56.8	19.2
CBARTFT	31.4/13.3/26.6	3.0	66.7	54.5
类似CGPT2	37.7/17.4/32.9	3.3	83.8	49.1
类似CBART	56.6/39.6/54.7	19.7	68.9	99.4
类似CGPT2FT	39.5/19.0/34.0	4.0	68.2	84.3
类似CBARTFT	57.3/40.5/55.3	19.9	69.1	99.0

表 7:不同模型的结果,均受到预先指定的车辆和比较器的约束。

约束条件	红色-1/2/L	蓝色	BERTcore	ACC-V	ACC-C
无	29.5/10.4/27.1		63.4	17.9	38.5
上下文	35.4/14.7/32.8		65.4	27.4	42.0
比较器	43.0/23.6/41.5		66.2	30.0	95.9
车辆	51.9/30.6/47.6		68.4	99.0	47.2
车辆 + 比较器	57.3/40.5/55.3		69.1	99.0	99.9
车辆 + 比较器 + 上下文	59.8/41.4/ 57.2		69.9	94.8	98.3

表 8:SimilorCBART 下不同约束和组合的性能。 ACC-C:准确度
如果没有预先指定,比较器是否出现在最终输出中。

生成最终的明喻。

方法。作为明喻生成的新任务,我们
使用 Similor 进行基准测试并评估模型变体,如下所示:1)
ChineseBART (CBART) 和
2) ChineseGPT2 (CGPT2),如第 5.1 节所述。
然而,他们以语言建模为学习对象,不能直接适应新的环境。

任务。跟随何等人。 (2022)使用手册
prompt for simile probing, we use “以_为喻体,
写出比喻句:(意思是用交通工具写一个比 _ 作为
喻: , _ 是预先指定的占位符
textitvehicle)”作为提示。然后,将其与给定主题和比较器连
接起来作为
输入同时生成明喻,这与
情境学习(Brown et al., 2020)。
3)微调 ChineseBART (CBARTFT)和 4)
微调中文GPT2 (CGPT2FT)。我们在收集到的 260k 数据上
对 CBART 和 CGPT2 进行了微调
学生作文与语言建模
分别为对象。微调的目标是
让模型适应作文写作
领域。 5)类似。我们首先实例化Similor
具有 CBART 和 CGPT2,即 SmilorCBART
和SmilorCGPT2,分别。评估增益
继续对学生作品进行微调的表演,Similor 也被实例化为

CBARTFT和 CGPT2FT,即SmilorCBARTFT
和SmilorCGPT2FT ,分别。所有型号
然后通过 GraCe 数据集进行微调。在那之后,
我们评估 Scorer 变体和基线如下:

1)字面错误匹配(LFM)。第二
Scorer 的步骤,旨在通过以下方式对候选人进行排名
词嵌入之间的欧氏距离
候选人和男高音。 2) ANT (陈等人,
2022): BERTLarge的预训练阶段,仅
掩盖 amod 依赖项中的名词或形容词。
继李等人之后。(2022),我们通过 Google 将连接的比较器和主
题翻译成英文
翻译并将其提供给 ANT 以生成车辆。

实验设置。我们随机分割
GraCe数据集分为2000个测试样本,以及2000个
验证样本,其余用于训练。所有模型的训练参数设置

与第 5.1 条相同。由此推断,光束尺寸
和长度惩罚(Wu et al., 2016)设置为 4
和1.2,分别。至于评价Scorer,我们
保留步骤 1 的前 20 个候选者,最后返回用于生成明喻的前
1 个工具。
为了公平比较,所有检索方法都使用
SimlorCBARTFT生成最终结果。

指标。继Chakrabarty 等人之后。 (2020);
张等人。 (2021);李等人。(2022),我们评估了
BERTScore的结果(Zhang et al., 2020),
四克BLEU (Papineni 等人, 2002 年)、 ROUGE 1/2/L
(Lin, 2004 年)。此外,如果车辆或比较器被预先指定为约束,
我们使用
ACC-V或ACC-C评估精度
提供的车辆或比较器出现在输出中。
作为 CSG 中的一种新颖设置,车辆未知的 CSG
旨在找到一个具有象征意义但字面意义的

方法	自动评估					人工评价		
	类似Conf. ↑字面Simi. ↓	PPL ↓	%V ↑	Fluen. ↑	Creat. ↑ Consi. ↑总体↑			
在	0.6	0.003	25.0	42.7%	1.9 100.0% 2.7	1.7	1.6	1.7
线性调频	0.8	-0.020	3.1	28.1	12.8 100.0%	2.3	2.3	2.3
得分手	0.8	0.240				2.5	3.0	2.6

表9:使用不同张量-车辆对检索方法生成明喻的主要结果。 %V代表
其车辆被检索到的样本数占测试样本总数的比例。

自动的	人类评价分数
指标	流动。创造。考虑。全面的
明喻会议0.312	0.634 0.603 0.540
%比较0.351	0.286 0.329 0.324 0.388 0.311
个人PL 0.377	0.321

表 10:自动指标与人类评估分数之间的 Pearson 相关性（p 值 < 0.01）。

false (Goodman, 1979)男高音车辆对有
共享属性来形成地面。因此,为了评估Scorer,我们首先使用
Simile Conf。和第5.1节中提到的Per plexity (PPL)来评估

输出的比喻意义和文本质量，
分别。继Shutova 等人之后。 （2016）;于
和 Wan (2019),计算字面上的错误因子
由字面量 Simi. 表示,表示平均余弦
给定主旨与检索载体的相似度,越低越好。我们使用
SimlorCBARTFT
计算词嵌入。除此之外,我们
进行第 5.1 节中所述的人工评估。

结果。不同模型品种的比较如表 7 所示。我们发现： 1)两者

CSG 任务和模型受益于预训练
阶段,特别是对于基于 BART 的骨干网。 2）
SimilorCBART和 SimilorCGPT2都可以生成正确包含约束的明喻
在输出中,文本质量高于基线。
此外,Similor的表演也有不同
约束条件如表8 所示,它表明： 3)引入更多明喻约束有助于

生成所需的明喻。特别是上下文,Sim ilor 只能生成明喻

上下文（BERTScore 63.4 至 65.4）。最后,作为
如表 9 所示,Scorer在象征意义上和基于模型的检索方法上都优于基于模型的检索方法。
文本质量,保证提供载体
每个测试的男高音。对于字面相似度,LFM
获得最高分但冲浪者得分最低
文本质量,表明在象征意义和字面意义之间存在权衡

生成明喻时的因素。

5.3 进一步讨论

作为明喻生成中的一个新任务,评估
它的方法绝对重要。因此,我们计算了系统级 Pearson 相关性
自动评分和人类对生成的明喻的判断。在表10 中,明喻
Conf。显示了一个
与人类得分有很强的相关性
创造力和一致性,表明它可以
成为评价形象的有效方法
比喻的意义。相反,% Comp。显示了一个
与这两个分数的相关性较差,这表明在判断明喻时仅考虑比较者的局限性。与此同时,PPL

在评估流畅性方面比其他两个指标表现出更高的相关性,但具有显着的效果
与人类分数的差距。为进一步探索
人类在评估明喻时所关心的问题，
我们还计算了人类的内部相关性
分数。如附表11所示,有一个
创造力和一致性之间有很强的相关性。这意味着有基础对于产生创造性的明喻也很重要,说明了

可解释地检索男高音的必要性
在车辆未知设置中配对。

六,结论

在本文中,我们介绍了一种新的任务设置
明喻生成:可控明喻生成
（南玻集团）。为了方便它,我们构建了 GraCe,一个细粒度注释的中文明喻数据集,并且
使用所提出的 CSG 模型对该任务进行基准测试
Similor,其中包括车辆检索模块
得分手。我们的工作首先尝试扩大
从认知语言学的角度分析明喻的要素（Kövecses， 2010）
（即基础和
上下文），并初步给出了从上下文中探究明喻解释的成功实现
认知属性。我们希望这个想法能够提供
对创意一代未来作品的新颖见解,例如双关语、夸张法和诗歌等。

局限性

在本文中，我们探索结合多种
明喻生成的限制并尝试从以下方面解释明喻比较
认知语言学。然而，创意的
明喻是一种主观感觉，很难准确判断，这也是一个很大的问题。
挑战其他类型的创意写作任务。
我们希望这个任务和数据集能够提供新颖的
洞察面向用户的文本生成，并给出
互动和协作的一代更加紧密
以及更细致的探索。

道德声明

我们特此承认所有合著者
这项工作的成员都了解所提供的 ACL 代码
道德和荣誉行为准则。我们详细阐述
对社区的道德考虑如下：

研究中执行的所有程序涉及
人类参与者按照
机构和/或国家的道德标准
研究委员会和 1964 年赫尔辛基
宣言及其后来的修正案或类似的道德标准。本文不包含
由任何机构进行的任何动物研究
作者。研究中的所有个体参与者都获得了知情同意。具体来说，我们
通过以下方式进行所有人工评估
来自中国数据的全职中国员工
注释平台，确保相关工作人员的所有个人信息（例如，用户名、
电子邮件、URL、人口统计信息等）是
被丢弃。同时，我们确保每个样本的报酬高于注释者
当地的最低工资
（大约 0.7 美元/样品）。

参考

Antoine Bosselut、Hannah Rashkin、Maarten Sap、Chai Tanya Malaviya、Asli Celikyilmaz 和 Yejin Choi。2019.COMET：用于自动知识图构建的常识变压器。在2019年亚冠比赛中，第 4762-4779 页。计算协会语言学。

汤姆·B·布朗、本杰明·曼、尼克·莱德、梅兰妮·苏比亚、贾里德·卡普兰、普拉芙拉·达里瓦尔、阿尔文德·尼拉坎坦、普拉纳夫·希亚姆、吉里什·萨斯特里、阿曼达·阿斯克尔、桑迪尼·阿加瓦尔、阿里尔·赫伯特-沃斯、格雷琴·克鲁格、汤姆·海尼汉、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、克莱门斯·温特、克里斯托弗·黑塞、马克·陈、埃里克·西格勒、马特乌斯·利特文、斯科特·格雷、本杰明·切斯、杰克·克拉克、克里斯托弗·伯纳、山姆·麦坎德利什、亚历克·雷德福、伊利亚·苏茨克维尔和达里奥·阿莫代。2020 年。语言模型是小样本学习者。在 NeurIPS 2020。

乔治·坎贝尔。 1988.修辞哲学。西乌出版社。

约翰·D·坎贝尔和阿尔伯特·N·卡茨。 2006.论反转隐喻的主题和载体。隐喻和象征，21（1）：1-22。

Tuhin Chakrabarty、Smaranda Muresan 和 Nanyun 彭. 2020.像一个人一样毫不费力地生成明喻pro：一种用于明喻生成的风格迁移方法。EMNLP 2020，第 6455-6469 页。协会计算语言学。

Tuhin Chakrabarty、张旭瑞、Smaranda Muresan、还有彭南云。 2021a。美人鱼：隐喻具有象征意义和辨别性解码的一代。 NAACL 2021，第 4250-4261 页。协会用于计算语言学。

Tuhin Chakrabarty、张旭瑞、Smaranda Muresan、还有彭南云。 2021b。美人鱼：隐喻具有象征意义和辨别性解码的一代。 NAACL 2021，第 4250-4261 页。协会用于计算语言学。

Weijie Chen, Yongzhu Chang, Rongsheng Zhang, Ji ashu Pu, Guandan Chen, Le Zhang, Yadong Xi, Yi jiang Chen, and Chang Su. 2022. Probing simile来自预先训练的语言模型的知识。在 ACL 2022，第 5875–5887 页。计算语言学协会。

雅各布·德夫林 (Jacob Devlin)、张明伟 (Ming-Wei Chang)、肯顿·李 (Kenton Lee) 和克里斯蒂娜·图塔诺娃。 2019.BERT：预训练用于语言理解的深度双向转换器。 NAACL 2019，第 4171-4186 页。计算语言学协会。

劳尔·J·恩德 (Laure J. End)。 1986.隐喻理解的基础。心理学进展，39：327-345。

约瑟夫·L·弗莱斯. 1971. 衡量许多评估者之间名义规模的一致性。心理通报，76（5）：378。

尼尔森·古德曼。 1979 年。隐喻为兼职。批判性探究，6：125 – 130。

Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and 肖阳华. 2022.预先训练的语言模型能否像人类一样聪明地解释明喻？在 ACL 中 2022 年，第 7875–7887 页。计算语言学协会。

佐尔坦·科韦切斯。 2010.隐喻的新视角认知语言学中的创造力。 21（4）：663-697。

Yucheng Li, Chenghua Lin, and Frank Geurin. 2022. 多任务学习的名义隐喻生成。在 INLG 2022 中。

林钦耀. 2004. ROUGE:自动包装评估摘要。 《文本摘要分支》,第 74-81 页,西班牙巴塞罗那。前交叉韧带。

刘丽珍、胡晓、宋伟、付瑞吉、刘婷、胡国平。 2018.神经多任务学习用于明喻识别。 EMNLP 2018,第 1543-1553 页,比利时布鲁塞尔。计算语言学协会。

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man dar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta:稳健优化的 BERT 预训练方法。 CoRR,abs/1907.11692。

Zhiqiang Liu, Zuohui Fu, Jie Cao, Gerard de Melo, Yik-Cheung Tam, Cheng Niu, and Jie Zhou. 2019b. 现代的修辞控制编码器-解码器中国诗歌一代。 ACL 2019,第 1992-2001 页。计算语言学协会。

毛瑞和小李。 2021.通过基于方面的情感分析和顺序隐喻识别的门控机制桥接多任务学习塔。 AAAI 2021,第 13534-13542 页。

AAAI出版社。

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei Jing Zhu。 2002. Bleu:一种机器翻译自动评估方法。 ACL 2002,第311-318 页。前交叉韧带。

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. CPT:预训练的不平衡变压器中文理解和生成。

CoRR,abs/2109.05729。

叶卡捷琳娜·舒托娃、杜威·基拉和让·梅拉德。 2016.黑洞与白兔:视觉特征的隐喻识别。 NAACL 2016,第160-170 页。

凯文·斯托、图辛·查克拉巴蒂、南云·彭、斯马兰达·穆雷桑和伊琳娜·古列维奇。 2021 年。 使用概念映射生成隐喻。 ACL 2021,第 6724-6736 页。计算语言学协会。

Jiao Sun、Anjali Narayan-Chen、Shereen Oraby、 Shuyang Gau、Tagyoung Chung、Jing Huang、Yang Liu 和 Nanyun Peng。 2022.上下文双关语的生成。 2022 年自然语言处理经验方法会议论文集,EMNLP 2022,阿拉伯联合酋长国阿布扎比,2022 年 12 月 7-11 日,第 4635-4648 页。计算语言学协会。

孙茂松、刘婷、王晓杰、刘志远、刘阳,主编。 2018.中文计算基于语言学和自然语言处理自然标注大数据 - 第十七届全国会议、CCL 2018、第六届国际会议研讨会,NLP-NABD 2018,中国长沙,2018 年 10 月 19-21 日,论文集,计算机科学讲义第 11221 卷。施普林格。

罗伊·塔塔科夫斯基、大卫·费舍洛夫和耶沙亚胡·申。 2019.不像白天那么清楚:论讽刺、幽默和封闭明喻的诗意。隐喻和象征， 34（3）:185–196。

托尼·维尔和郝艳芬。 2007年。学习理解比喻语言:从明喻到隐喻再到讽刺。认知科学学会年会记录,第 29 卷。

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016.谷歌的神经机器翻译系统:弥合人类翻译和机器翻译之间的差距。 CoRR,abs/1609.08144。

于志伟和万晓军。 2019.如何避免句子拼写乏味?走向神经方法无监督隐喻生成。 NAACL 2019,第 861-871 页。前交叉韧带。

Jiali Zeng, Linfeng Song, Jinsong Su, Jun Xie, Wei Song, and Jiebo Luo. 2020. Neural simile recogni tion with cyclic multitask learning and local attention. AAAI 2020,第 9515-9522 页。 AAAI出版社。

Jiayi Zhang, Zhi Cui, Xiaoqiang Xia, Yalong Guo, Yan ran Li, Chen Wei, and Jianwei Cui. 2021. Writing polishment with simile: Task, dataset and A neural approach. In AAAI 2021, pages 14383–14392. AAAI Press.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. 温伯格和约阿夫·阿齐。 2020. Bertscore:使用 BERT 评估文本生成。在 ICLR 中。 OpenRe view.net。

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. UER: an open-source toolkit for pre-training models. In EMNLP 2019, pages 241– 246. Association for Computational Linguistics.

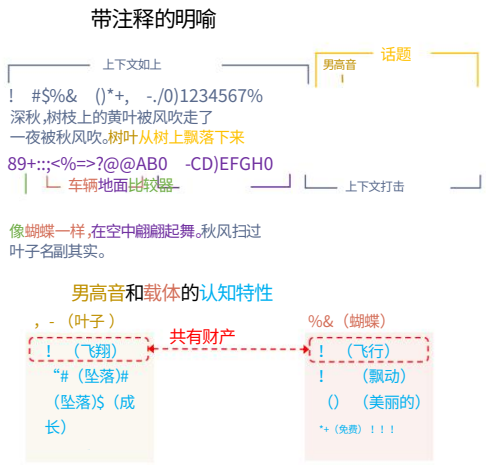


图 4:解释带注释的八个示例
我们的 GraCe 数据集中的明喻元素。翻译
为非华语人士提供。

数据集构建细节

如图4所示,我们常见的展开三个
带注释的元素（即男高音、载体和比较器） (Li et al., 2022)到八个,包括
上下文元素将每个明喻放入更多
自然使用的情况。

A.1 明喻分类

明喻分类器的目的是过滤那些不包含比较词的明喻样本。

这些句子大致可以分为三部分
types: 1) personified sentence, eg, “大树好像在向我们招手。 (The tree seems to be waving to us.)”包含比较词“希望(似乎)”。
2) hyperbole sentence, eg, “这教室静得仿佛掉一根针都能听见。 (The classroom was so silent 就像你能听到一根针掉落的声音。）”包含比较词“仿佛（像）”。 3)字面句子,例如“他
似乎从来没有来过这里。 (He never seems to be here.)” contains comparator word “似乎(seems 到）”。然而,之前的数据集（Li et al., 2022）
只提供不包含比较词的字面句子作为负样本

对于明喻分类器,这可能不能满足我们的要求
设置。

为此,我们收集了一个新的数据集,其中包括
这三类无明喻句的负样本。具体地,收集人
化句 15和夸张句 16来自
网站,并且只保留包含比较词的句子。对于第三类,我们要求三名注释者对从步骤 1 候选者中随机选择的 3000 个样本进行注释。选定了一个句子

如果所有人都认为它是负样本
一个字面的句子。对于正样本,我们
还从作文教学网站上收集明喻,确保风格相似
17 号

给我们的候选人。最后,我们得到新的简单分类数据集并将其随机分为:
训练集 5905 个样本（正:2913 负:
2992）/验证集200个样本（阳性:100
阴性:100）/测试集 200（阳性:100 阴性:100）。

基于这个新数据集,我们微调了中文
RoBERTaLarge模型对步骤 1 候选日期进行分类。为了训练这个模型,学习率为
设置为 5e-5,预热步骤设置为 200。
验证集和测试集的 f1 分数为
分别为 0.85 和 0.82。

A.2 明喻检测

明喻检测的目的是标出主旨和
明喻的载体,即将其形成一个序列
标记任务。在植入中,我们使用最相关的数据集 CCL2018 (2018)来训练序列
标签模型。 CCL2018数据集包含
6554 个训练样本,2038 个测试样本,
1650 个验证样本。基于这个数据集,我们
微调中国 RoBERTaLarge模型以进行标记
GraCe 中的每个样本。为了训练这个模型,
学习率设置为 5e-5,预热步骤
设置为200。验证集和测试集的准确率分数分别为98.47%和98.38%,
分别。

然而,所有样本仅包含一种
比较词（即“像（like）”）,经过训练的
模型不能直接应用于 GraCe
包含各种比较词及其相应的模式。为了解决这个问题,在

推理阶段,我们首先定位并替换每个
比较器模式与包含的模式
比较词“像”,因为它们有相同的
不同词语的含义（都表示类似）。后

15条拟人句:
https://www.t262.com/juzi/nirenju/,
https://wenku.baidu.com/view/a70e349cbbf3f90f76c66137ee06eff9aef84906.html

16个夸张句子:https://www.chazidian.com/zaoju_5/

17https://www.yuwenmi.com/yuwenjichu/biyuju/

算法1模糊匹配

要求： C:以名词为键、以相关认知属性为值的Cogbank 字典

要求： t:需要查询的分词后的单词序列,长度为l， t = {t1, t2, ..., tl}

要求： w:滑动窗口的宽度。

```
w = l
while w > 0如果w
   = l 且 t ∈ C那么
      返回t
   别的
      我 = 1
      while i < l + 1 do
         word = {ti , ..., ti+w}
         if word ∈ C then
            return word

         else i = i +
         1 end if
      结束同时
   万一
   w = w     1
end while
return None 单词映射
```

也就是说,我们使用这个新样本作为模型输入来获得相应的期限和载体。

A.3 Cogbank 数据集的模糊匹配模糊匹配算法如算法1所示。

A.4 明喻样本我们在表12中展

示了一些带注释的 GraCe 样本。

B 实验细节

B.1 明喻生成前缀

We consider 25 commonly used tenors as sentence starters for evaluating different datasets in the Ex periment for prefix generation. The entire set is blow (Translations are provided for non-Chinese speakers.):

“爱(love)”，“时间(time)”，“叶子(leaves)”，“太阳(sun)”，“树叶(leaves)”，“童年(childhood)”，“笑容(smile)”，“落叶(fallen leaves)”，“眼泪(tears)”，“阳光(sunshine)”，“泪水(tears)”，“时光(time)”，“柿子(persimmon)”，“生命(life)”，

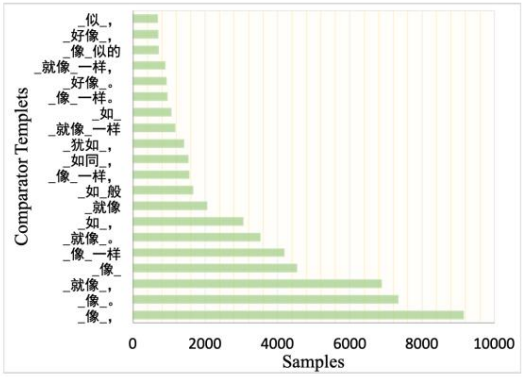


图 5:GraCe 中最常见的 20 个比较器模板,全部表示“喜欢”。 “_”表示占位符,可以填充与基调相关（第一个)和与车辆相关（第二个)的内容。

“记忆(memory)”，“花瓣(petals)”，“天空(sky)”，“目光(gaze)”，“雪花(snowflakes)”，“苹果(apple)”，“青春(youth)”，“枫叶(maple leaves)”，“友谊(friendship)”，“微笑(smile)”，“幸福(happiness)”.

在推论中,我们使用k=10的top-k采样,并将所有模型的随机种子固定为 42 来得到最终结果,而最大生成长度设置为 100。

B.2 前缀生成样本的生成

为了直观地展示数据集的效果,我们在表 13 中展示了一些生成结果。

B.3 可控明喻生成样本的生成

具有不同约束的 Similor 的一些生成结果如表14所示,我们也进行了比较

不同车辆检索方法的Similor结果如表15所示。

C 人类评估的细节

C.1 数据集比较的人工评估

为了将 GraCe 数据集与其他相关数据集进行比较,从每个数据集中随机选择 1000 个样本。同时邀请了三位专业标注员对这些数据样本进行标注。值得注意的是,所有记名者的母语都是中文。专业标注员和众包标注员唯一的区别是,专业标注员是汉语言文学专业，而众包标注员只需要中文相关专业

文学。由于学习的课程包括汉语语法和修辞学,专业注释者
能够验证数据集中的细粒度注释是否正确。

在正式进行之前,我们首先制定评估指南,包括任务背景、要点、详细描述以及不同明喻模式的示例。然后我们

为注释者设置进入壁垒。具体来说,我们组织了一个培训计划和初步注释检查(每个数据集 20 个示例)

选择适当的注释者并获得批准
率高于95%。
分数定义我们首先要求注释者确定给定的样本是否是明喻(1 表示
给定的样本是明喻,0 是相反的)。
值得注意的是,CMC 数据集(Li et al., 2022)也
包含隐喻,请注释者注意
将该案例作为另一种明喻并给它们贴上标签
1. 除此之外,我们进一步检查样本的细粒度注释元素

GraCe 数据集。详细的还询问注释者
判断被注释的元素是否
这些样本是正确的(1 表示是,0 表示
相反),包括主旨、载体、比较器、
和接地。
注释者间协议我们使用 Fleiss
kappa (Fleiss, 1971)测量三个注释者的
可靠性 18.结果是: 1)对于CS数据集:
0.72(显着); 2)对于CMC数据集:0.62(实质性); 3)
对于 GraCe 数据集:0.78(大量)。

C.2 人类评估的细节

对于人类评估,我们首先制定了一个指导方针
评估,其中包括任务背景,
要点、详细描述和示例
评估分数从1到4。然后,我们设置一个
注释者的进入障碍。具体我们整理一下
培训计划和初步注释
考试(每个模型 50 个示例)选择
合适的标注者,通过率较高
超过95%。
分数定义我们定义了四个类别
人工评价如下:

1.流畅(Fluen.)是指句子是否
选项对应的内容流畅、语法合理、结构良好、易于理解。

2、创意(Creat.)是指该选项对应的句子是否具有创意
并且具有象征意义。

3.一致性(Consi.)是指该选项对应的句子是否包含

有意义的男高音载体对。一个有意义的
对表示有一些共享属性
男高音和载体之间,即有
显式/隐式基础。

4.总体是指该句子总体上与该选项的对应程度如何?要求注释者对生成的结果进行评分

基于它的结构良好、创造性和一致性。

注释者间协议我们使用 Fleiss
kappa (Fleiss, 1971)测量三个注释者的
可靠性 19.结果是:1)用于实验
Q1:0.43(中等)2) 对于实验 Q2:0.30
(缓和)。

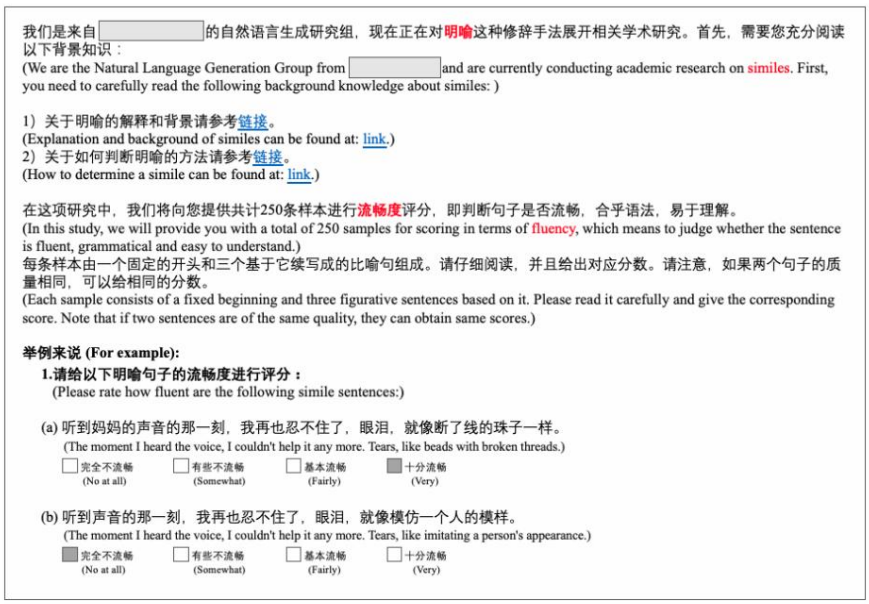
C.3 相关性分析

流动。创造。考虑。全面的	
流动。 0.477 0.482 创建。 0.477 - 0.970缺 点0.482 0.970 - 总体 0.729 0.841 0.843	0.729 0.841 0.843 -

表 11:不同人类之间的皮尔逊相关性
评估分数(p 值 < 0.01)。

18https://www.nltk.org/_modules/nltk/metrics/协议.html

19https://www.nltk.org/_modules/nltk/metrics/协议.html


我们是来自[ ]的自然语言生成研究组，现在正在对**明喻**这种修辞手法展开相关学术研究。首先，需要您充分阅读以下背景知识：
(We are the Natural Language Generation Group from [ ] and are currently conducting academic research on similes. First, you need to carefully read the following background knowledge about similes:)

1）关于明喻的解释和背景请参考链接。
(Explanation and background of similes can be found at: link.)
2）关于如何判断明喻的方法请参考链接。
(How to determine a simile can be found at: link.)

在这项研究中，我们将向您提供共计250条样本进行**流畅度**评分，即判断句子是否流畅，合乎语法，易于理解。
(In this study, we will provide you with a total of 250 samples for scoring in terms of fluency, which means to judge whether the sentence is fluent, grammatical and easy to understand.)
每条样本由一个固定的开头和三个基于它续写成的比喻句组成。请仔细阅读，并且给出对应分数。请注意，如果两个句子的质量相同，可以给相同的分数。
(Each sample consists of a fixed beginning and three figurative sentences based on it. Please read it carefully and give the corresponding score. Note that if two sentences are of the same quality, they can obtain same scores.)

举例来说 (For example):

1.请给以下明喻句子的流畅度进行评分：
(Please rate how fluent are the following simile sentences:)

(a) 听到妈妈的声音的那一刻，我再也忍不住了，眼泪，就像断了线的珠子一样。
(The moment I heard the voice, I couldn't help it any more. Tears, like beads with broken threads.)

☐ 完全不流畅 (No at all) ☐ 有些不流畅 (Somewhat) ☐ 基本流畅 (Fairly) ■ 十分流畅 (Very)

(b) 听到声音的那一刻，我再也忍不住了，眼泪，就像模仿一个人的模样。
(The moment I heard the voice, I couldn't help it any more. Tears, like imitating a person's appearance.)

■ 完全不流畅 (No at all) ☐ 有些不流畅 (Somewhat) ☐ 基本流畅 (Fairly) ☐ 十分流畅 (Very)

图 6:流畅度评分界面。

话题	比较器	男高音		车辆		地面	语境	
		单词	属性词		Property		多于	以下
Sample 1:远看,层林尽染。近看,那深红、浅红、金黄的枫叶,像一只只小手掌在风中摇曳着,似乎在欢迎着我们的到来。片片美丽的叶子像蝴蝶一样飘飞,脚底有树叶轻轻的碎响,秋那厚重的美就久久盘旋心头。 Sample 1: From a distance, the layers of trees are dyed in color. Looking up close, the dark red, light red and golden maple leaves, like small palms swaying in the wind, seem to welcome us. Pieces of beautiful leaves fluttered like butterflies, and the soles of my feet were softly cracking, and the heavy beauty of autumn was circling in my heart for a long time. 片片美_像_一样,叶子飞, 蝴蝶飞, 飘飞远看...似丽的叶乎在欢迎飘落,飞舞, 子着我们的落...美丽...到来。 fluttered From a dis							脚底...盘旋心头。
一片片美丽的叶子	喜欢	树叶	飞翔、坠落、坠落……	蝴蝶飞舞,翩翩起舞,美丽……			坦斯.....似乎欢迎 我们。	而鞋底……在我心里盘旋了很长时间。
Sample 2:当秋姑娘来到了硕果累累的果园时。那一串串紫色的葡萄就像一颗颗紫色的珍珠,真美丽啊!粉红的苹果绽开了笑脸,好像在说:“秋姑娘来了,我们又苏醒了。” 例二:当秋姑娘来到硕果累累的果园时。一串串紫色的葡萄就像一颗颗紫色的珍珠,真漂亮! Pink apple blooming smile, as if to say: Autumn girl came, we woke up again. 那一串_就像_,葡萄水灵灵,珍珠熠熠生辉,无当...果园串紫色时。亮晶晶,晶莹,的葡萄晶莹... 细腻...								真...又苏醒了。”
一串紫色的葡萄	喜欢	葡萄	水润的、闪光的、水晶的……	珍珠	晶莹剔透、晶莹剔透、精致……	没有任何	当……果园。	真的……又醒了。”
Sample 3:透过晶莹的泪珠,我看到了暖洋洋的太阳。爸爸妈妈的爱不就像太阳一样温暖着我吗?那一刻,已成为我人生中最重要的时刻,时时牵动着我的心..... Sample 3: Through the crystal tears, I saw the warm sun. Isn t mom and dad s love warm me like the sun? That moment has become the most important moment in my life, always affecting my heart... 爸爸妈_就像_一爱热烈,太阳温暖,温暖着妈的爱样我吗?甜,光明,温暖... 火红...							透过...太阳。	那一刻,...心.....
爸爸妈妈的爱	喜欢	爱	温暖、甜蜜、 温暖的...	太阳	温暖的,光、火...	温暖的 我	通过...阳光。那个莫	心……心……
Sample 4:到了云锦山庄,我们被眼前的景色迷住了,仿佛走进了仙境。朵朵白云漂浮在碧蓝的天空中,好像一团团棉花。明净的湖面上隐隐约约可以看到蓝天白云的倒影,微风吹过,湖面上泛起了小小的波纹,在阳光的照耀下,变得波光粼粼,真是美丽啊! Sample 4: To the Yunjin Villa, we were fascinated by the scenery, as if in a fairyland. White clouds floated in the blue sky, like clumps of cotton. The reflection of the blue sky and white clouds can be seen faintly on the clear lake. The breeze blows, and there are small ripples on the lake. Under the sunshine, it becomes sparkling, it is really beautiful! 朵朵白棉花柔软,明净的湖云漂浮轻盈,面...美丽在碧蓝啊!蓬松... 的天空中了...white clouds floated in the blue sky,								
	喜欢	白云_	柔软、白色、轻盈……	棉布	柔软、轻盈、蓬松……	没有任何	到...仙境倒影...美丽!	

表 12:GraCe 数据集中带注释的样本。对于每个样本,我们首先给出全文,然后给出所有注释信息。 “无”表示比喻中隐含根据。由于篇幅限制,我们省略了一些属性和上下文的内容。为非中文使用者提供翻译。

Prefix 1:时间(time) CS时间果然很是短暂,不等他做出什么反应时就已经过去了。 (The time was really short, and it passed 在他反应过来之前。) CMC时间是像一条彩带系在心中。 (Time is like a ribbon in the heart.) GraCe时间,如同奔腾的江河,在我面前悄然逝去。 (Time, like a pentium river, quietly flows away in front of 我。)
Prefix 2:叶子(Leaves) CS叶子何时变得这么小,简直就像是一个孩子一般。我的心也被这个情景吓到了。 (When did the leaf 变得那么小,就像个孩子一样。我的心也被这一幕吓到了。) CMC叶子像,就像那空中飞舞的蝴蝶一样,扇动着薄翼,初降人间。 (The leaves are like, just like the 蝴蝶在空中,扇动着薄薄的翅膀,飘落到人间。) GraCe叶子,长得又细又尖,像一把锋利的宝剑。 (The leaves grow thin and pointed, like a sharp sword)
Prefix 3:太阳(Sun) CS太阳点点亮,天边的云层被晒得滚红,就像是要下雨了似的。 (The sun lit up, and the clouds on the 天边都变红了,好像要下雨了。) CMC太阳就是春天的一缕阳光,拂去冬日炎热和干燥。 (The sun is a ray of spring sunshine, brushing away the 冬季炎热干燥。) GraCe太阳火辣辣的照着大地,就像一个巨大的火球在燃烧。 (The sun blazed down on the earth like a huge 火球燃烧。)
Prefix 4: 记忆(Memory) CS记忆果然如同传说中一样美丽迷人。 (The memory is as beautiful and charming as the legend.) CMC记忆,就像那些仙天上的星斗一样美丽。 (Memories are as beautiful as the stars in the fairy sky.) GraCe记忆,是一个五味瓶,酸甜苦辣样样俱全,如同过山车般满载喜怒哀乐。 (Memory is a tasting bottle, 酸的、甜的、苦的、辣的样样都有,就像坐过山车,充满了悲欢离合。)
Prefix 5:雪花(Snowflake) CS雪花着何其美丽,雪花就好像是精灵一样,一点一点融化在天空中。 (How beautiful snowflakes are, 雪花就像仙女一样,一点一点地在天空中融化。) CMC雪花,在空中飘动,就像一个个银色的精灵。 (snowflakes, fluttering in the air, like a silver spirit.) GraCe雪花,飘落的时候就像一只美丽的仙鹤在空中翩翩起舞。 (The snowflake falls like a beautiful crane 在空中跳舞。)

表 13:使用不同数据集生成前缀的示例。表示对应的句子是明喻，while 表示这不是明喻。为非中文使用者提供翻译。

Topic:他唱得并不好听,声音(He did not sing well, and his voice,) Reference:他唱得并不好听,声音像砂纸打磨着耳朵。 (He didn t sing very well, and his voice was like 用砂纸打磨耳朵。)
约束:无 他唱得并不好听,声音是如雷鸣般的响声。 (He did not sing well, and his voice was like thunder.)
Constraint: Context:可他是那样地认真,那样地投入,仿佛在大舞台上唱歌,台下有千万个观众聆听着。(但他是那么认真,那么投入,就好像他在一个大舞台上唱歌,有数百万人在听。) 他唱得并不好听,声音像一把利剑,深深地插进了我的胸膛。 (He did not sing well, and his voice, like a 剑,深深地刺入我的胸口。)
约束:比较器:像...。(像...句号) 他唱得并不好听,声音像蚊子在叫,又像苍蝇在乱飞。 (He did not sing very well, and his voice was like 蚊子的叫声,或者像苍蝇的叫声。)
约束:车辆:砂纸 (就像...句号) 他唱得并不好听,声音,就像砂纸一样沙哑。 (He did not sing very well, and his voice was like the cry of a 蚊子,或者像苍蝇一样。)
约束:上下文+比较器+车辆 他唱得并不好听,声音,像一块干燥的砂纸在微微地颤抖。 (He did not sing very well, his voice, like a 一块干砂纸轻轻颤动。)

表 14:不同约束和组合下的相似样本。为非中文人士提供翻译。

样本1:	
在	风,缓缓吹过,槐花树上的槐花便落下,洁白的花瓣如英亩一般,在空中飞舞。 (The wind, slowly blowing, flowers on the sophora tree fell down, white petals like acres, flying in the air.)
Literal False风,缓缓吹过,槐花树上的槐花便落下,洁白的花瓣,如白色的百合一般,美丽极了。 (The wind, slowly blowing, flowers on the sophora tree fell down, white petals like white lilies, very beautiful.)	
	Shared properties between tenor and vehicle:美丽(beautiful),绽放(blooming),盛开(full-blown),枯萎(withered),美好(good)...
Scorer风,缓缓吹过,槐花树上的槐花便落下,洁白的花瓣,如羽毛一般,轻轻地飞下来。 (The wind, slowly blowing, flowers on the sophora tree fell down, white petals like feathers, gently fly down.)	
	Shared properties between tenor and vehicle:散落(scattered),柔软(soft),美丽(beautiful),轻盈(lithesome), 飘(fluttering)...
样本2:	
在	然后在杯中倒入开水,胖大海立马就浮起来了,还像离开水的小白兔一样。 (Then we pour boiling water into the cup, the sterculia scaphigera floats up immediately like a white rabbit out of water.)
Literal False然后在杯中倒入开水,胖大海立马就浮起来了,我还像一只小刺猬一样蜷缩着。 (Then we pour boiling water into the cup, the sterculia scaphigera floats up immediately and I curl up like a little hedgehog.)	
	期限和车辆之间的共同属性:膨胀（膨胀）
Scorer然后在杯中倒入开水,胖大海立马就浮起来了,还像面包一样膨胀起来。 (Then we pour boiling water into the cup, the sterculia scaphigera floats up immediately, and expands like bread.)	
	Shared properties between tenor and vehicle: 膨胀(intumescent),发开(rasing)
样本3:	
在	没有任何
Literal False老人微眯双眼,眺望着天空中的风筝,眼神祥和宁静,如杰克般飞翔. . . . . (The old man squinted his eyes and looked at the kite in the sky. His eyes were peaceful and quiet, flying like Jack...)	
	男高音和载体的共同属性:忧郁（忧郁）
Scorer老人微眯双眼,眺望着天空中的风筝,眼神祥和宁静,如晨露般滋润着我的心田。 (The old man squinted his eyes and looked at the kite in the sky. His eyes were peaceful and quiet, which moistened my heart like morning dew.) Shared	
	properties between tenor and vehicle:干净(fresh), 清澈(limpid)
样本4:	
在	没有任何
Literal False望着一个个设施,一幅幅画面,从我们的眼前闪过,回忆,像蜡人似的,一个个地浮现在我们眼前。 (Looking at the facilities one by one, a picture flashed from our eyes, memories, like wax dolls, one by one emerged in front of our eyes.)	
	Shared properties between tenor and vehicle: 不真实(unreal)
Scorer望着一个个设施,一幅幅画面,从我们的眼前闪过,回忆,像春花似的,开满了我们的心田。 (Looking at the facilities one by one, a picture flashed from our eyes, memories, like spring flowers, open full of our hearts of the field.)	
	Shared properties between tenor and vehicle: 温暖(warm), 绚烂(splendid)
样本5:	
在	没有任何
Literal False 站在黑板前 , 我忽然有种恍然隔世的感觉 , 尘封已久的记忆如一片平静的太平洋。（站在黑板前,突然感觉好像一代人过去了,尘封的记忆就像平静的太平洋。）	
	男高音和载体的共同属性:深（深）、美丽（美丽）
Scorer站在黑板前,我忽然有种恍然隔世的感觉,尘封已久的记忆如一片大海,宽阔而又神秘。 (Standing in front of the blackboard, I suddenly feel as if a generation has passed, dusty memories are like a sea, wide and mysterious.)	
	Shared properties between tenor and vehicle:深(deep),美丽(beautiful), 悠久(long-standing)

表 15:不同车辆检索方法的相似样本。 “无”意味着没有检索到有效的车辆,我们突出显示张量-车辆对以便更好地查看。为非中文使用者提供翻译。

ACL 2023 负责任的 NLP 检查表

A 对于每次提交：

✓ A1。您是否描述了您工作的局限性？
请参阅限制部分（第 9 页）。

A2。您是否讨论过您工作中的任何潜在风险？
不适用。留空。

✓ A3。摘要和引言是否总结了论文的主要主张？
参见摘要和第 1 节。

A4。您在写这篇论文时使用过人工智能写作助手吗？
留空。

乙 您使用或创造了科学制品吗？

不适用。留空。

B1。您是否引用了您使用的工件的创建者？
不适用。留空。

B2。您是否讨论过任何工件的使用和/或分发的许可或条款？
不适用。留空。

B3。您是否讨论过您对现有工件的使用是否与其预期用途一致（前提是已指定）？对于您创建的工件，您是否指定了预期用途以及是否与原始访问条件兼容（特别是，出于研究目的而访问的数据的衍生物不应在研究环境之外使用）？

不适用。留空。

B4。您是否讨论了检查收集/使用的数据是否包含任何命名或唯一识别个人或攻击性内容的信息所采取的步骤，以及保护/匿名化所采取的步骤？

不适用。留空。

B5。您是否提供了工件的文档，例如领域、语言和语言现象的覆盖范围、所代表的人口群体等？

不适用。留空。

B6.您是否报告了您使用/创建的数据的相关统计数据，例如示例数量、训练/测试/开发拆分的详细信息等？即使对于常用的基准数据集，也要包括训练/验证/测试分割中的示例数量，因为这些为读者理解实验结果提供了必要的背景。例如，大型测试集上的准确性的微小差异可能很重要，而在小型测试集上则可能不那么重要。

不适用。留空。

C ✓您进行过计算实验吗？

参见第 5 节

✓ C1。您是否报告了所使用模型中的参数数量、总计算预算（例如，GPU 时间）以及使用的计算基础设施？
参见第 5 节

ACL 2023使用的负责任的 NLP 检查表采纳自NAACL 2022，还增加了一个关于人工智能写作辅助的问题。

✓ C2。您是否讨论了实验设置,包括超参数搜索和最佳找到的超参数值?

参见附录

✓ C3。您是否报告了有关结果的描述性统计数据（例如,结果周围的误差线、实验组的汇总统计数据）,以及您是否报告最大值、平均值等或仅报告单次运行是否透明?

参见第 5 节

✓ C4。如果您使用现有的包（例如,用于预处理、标准化或评估）,您是否报告了所使用的实现、模型和参数设置（例如,NLTK、Spacy、ROUGE等）?

参见第 3 节和第 5 节

D ✓您是否使用人类注释者（例如众包工作者)或与人类参与者一起进行研究?

参见第 5 节

✓ D1。您是否报告了向参与者提供的说明的全文,包括屏幕截图、
对参与者或注释者等的任何风险的免责声明?
参见附录

✓ D2。您是否报告了有关如何招募（例如,众包平台、学生)和付费参与者的信息,并讨论了考虑到参与者的人口统计（例如,居住国家）,此类付款是否足够?

请参阅附录和道德声明

D3。您是否讨论过是否以及如何获得您正在使用/管理其数据的人的同意?例如,如果您通过众包收集数据,您向众包工作者发出的指示是否解释了如何使用这些数据?

不适用。留空。

✓ D4。数据收集方案是否得到伦理审查委员会的批准（或确定豁免）?
请参阅道德声明

D5。您是否报告了作为数据来源的注释者群体的基本人口统计和地理特征?

不适用。留空。