题目： Fantastic Expressions and Where to Find Them: Chinese Simile Generation with Multiple Constraints

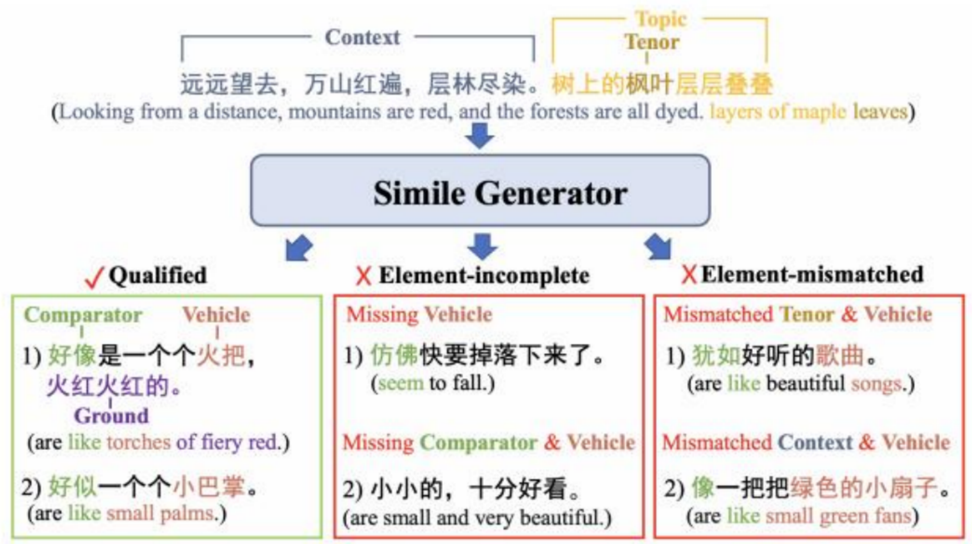
期刊： ACL-2023-1

previous efforts：

- form simile generation as a context-free generation task
  - focus on simile-style transfer (style-transfer-based),改述
    - paraphrase a literal sentence into a simile-style sentence and automatically edits self-labeled similes to their literal version for building pairs of(literal sentence,simile)
  - write a simile from a given prefix(prefix-based)，接明喻句
    - generate the comparator and tenor from a pre-specified tenor

weakness:

- might be undesirable,such as hardly meeting the simile definition



- or difficult to address certain preferences of content as human wishes
  - 例：想对颜色打比方，却生成了对形状打比方的结果

本文主要方法：

Trend: try to incorporate various constaints into simile generation(generating a simile with multiple constraints)

controllable simile generation (CSG)

- generate a simile with multiple simile elements from a given prefix

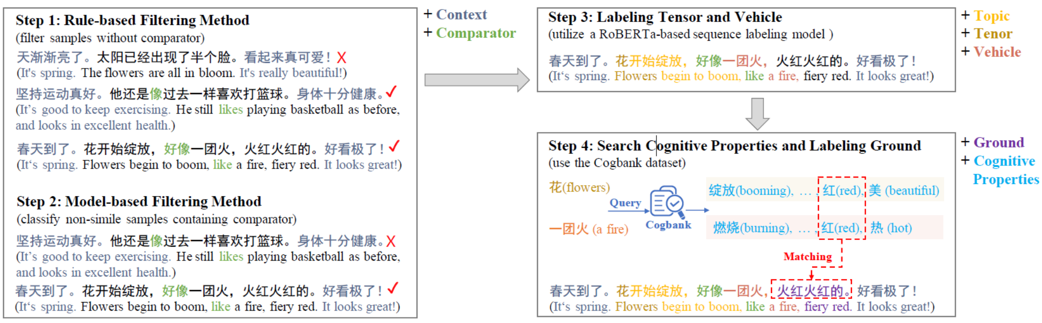
- The controllable simile generation task is formulated as follows: given a *topic* $\boldsymbol{x}$ containing a *tenor* s_t and a variety of pre-specified constraints $\boldsymbol{c}$, the model generates a simile $\boldsymbol{y} = (y_1, y_2, ..., y_N)$ by:

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{c}) = \prod_{n=1}^{N} p(y_n|y_{<n}, \boldsymbol{x}, \boldsymbol{c}; \theta), \qquad (1)$$

where θ are the model parameters. Notably, the constraints $\boldsymbol{c}$ can be freely selected and combined from the candidate set $\boldsymbol{s} = (s_v, s_p, s_c)$, which denote the *vehicle*, *comparator*, and *context*, respectively.

CraCe

- facilitate the CSG work
- dataset creation
    - dataset collection
        - 260k student compositions from the free-access website
        - sentence segmentation and removal of non-Chinese sentences
        - 2 sentences above and below each sample are used as the context element
    - dataset processing



- dataset analysis
    - data quality
        - 3 professional annotators to independently annotate 1000 randomly selected samples from multiple aspects
    - diversity of similes
- Innovation
    - expand three commonly annotated elements (tenor本体，vehicle喻体，comparator比喻词)to eight （topic, tenor, tenor property, comparator, vehicle, vehicle property, ground, context)
    - more collected samples than dataset CMC
    - for explicit ground: annotate to better understand the simile comparision
    - for implicit ground: interpret the relationship between *tenor* and *vehicle* by their cognitive properties

- incorporate position information by adding the punctuation that closely followed the *vehicle*.
- high quality beyond previous Chinese simile datasets

Similor

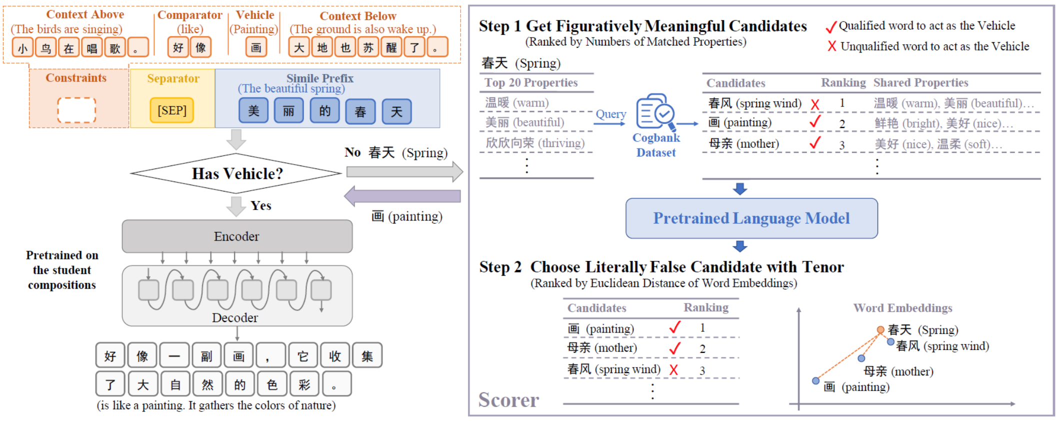
- aim to benchmark CSG
- Step
  - first retrieves *vehicle*(if it is unknown) by the module **Scorer**(a shared cognitive-property-based retrieval method) for the given tenor
  - then incorporates all contrains and input prefix to generate the simile.
  - 

Figure 3: A toy example to elaborate the workflow of Similor and Scorer.

$$
\begin{aligned}
Score_{w_i} &= \mathrm{Rank}(Fig_{w_i}) + \mathrm{Rank}(Lit_{w_i}),\\
Fig_{w_i} &= \mathrm{Match}(w_i, s_t), \text{两项之间共享认知属性(2)的数量}\\
Lit_{w_i} &= \mathrm{EucDist}(w_i, s_t). \text{词嵌入之间的欧几里得距离}
\end{aligned}
$$

- innovation
  - successfully incorporate the contraints in the outputs
  - especially in vehicle-unknown setup, Scorer beats the model-based retrieval method both in automatic and human evaluations without any re-training

CraCe $\xrightarrow{\text{facilitate}}$ CSG $\xrightarrow{\text{model}}$ Similor $\xrightarrow{\text{include}}$ Scorer ⇔ vehicle-retrieval methods without any retraining

(beats model-based retrieval

61.3k simile-element annotated Chinese similes

high quality beyond all other Chinese simile datasets

fine-grained

Similor to benchmark 基准测试

serve as a strong baseline for CSG

Scorer to obtain the explicable comparation for a given tenor

(In the vehicle-unknown situation

是否开源：

是（如来）

Our code and corpus will be released at https://github.com/yangkexin/GraCe.

数据集：

CraCe

实验：

1. evaluate GraCe with previous Chinese simile datasets
    1. prefix generation
    2. human evaluation
2. benchmark the CSG task with different model varieties, then explore the performances of Similor under different combinations of constraints.

Finding：

- ground plays an important role in making the tenor-vehicle pair of simile being easily-understood and figuratively meaningful, yet being ignored in previous datasets
- For GraCe: through the experiments, the paper find(5.1)—实验结果
    - Models finetuned with GraCe outperform other simile datasets in terms of text quality and simile creativity
    - Generative language models tend to produce literal sentences over similes that highlight challenges of simile generation.
- For CSG: through the experiments, the paper find(5.2)—实验结果
    - Both CSG task and models benefit from the pre-training stage, especially for the BART-based backbone.
    - Both $Similor_{CBART}$ and $Similor_{CGPT2}$ can generate similes that correctly incorporate constraints in outputs, which higher text quality than baselines
    - Introducing more simile constrains helps Similor to generate desired similes.
    - Scorer beats model-based retrieval method both in figuratively meaningful and text quality

Contribution：

- A new task setup for simile generation--CSG
- a fine-grained annotated Chinese simile dataset--CraCe, which take the first attempt to expand the elements of simile from the aspect of Cognitive Linguistics. It has 371 patterns comparators of simile.
- tentatively gives a successful implementation of probing simile interpretation from cognitive property
- hope future: may be applied in creative generation, such as puns, hyperbole, and poetry, etc.

个人理解想法：

这篇论文主要聚焦在明喻的生成部分，目标在于生成制定约束的明喻句子。由于明喻需要有明显的比喻词，所以处理起来明显比暗喻要容易的多。在训练模型时，作者直接将“像”换成其他比喻词喂给模型。

文中讨论的约束有

1. 预先指定了喻体——太好了，就不用Scorer生成了，省去Scorer模块，直接进入encoder-decoder
2. 预先指定比喻词

一个重要的点在于如何选取合适的喻体。文中Scorer方法根据本体和喻体（待选）的共享属性数量（越多越好）和他们之间的的欧式距离（越大越好），二者进行权衡。文中使用了两者分别取得分，算排名，然后将排名直接相加的方式。——I think: 可能有比排名简单相加更好的权衡方式？之前学过那么多比喻句，如果需要生成的本体和学过的本体是一样的，那么应该给那个学过的喻体加分，让学过的喻体更容易被生成？

我觉得可以给生成喻体属性约束，给具有预先指定属性的喻体（备选）加分

对文章写作部分：很多句子重复出现，比如”前人工作主要集中在两个方面及解释“在abstract，introduction，related work各出现一次。

在数据集中随机取1000个，使用了三个汉语言文学专业的人来对模型标注的数据进行测评（给的样本是否为明喻句子，分辨出隐喻并将其当做另一种明喻，以及8个元素对应是否正确），在开始之前执行了培训计划和预先标注检查（准确率大于95%）来选取合适的测评者。我理解这个“approval rate”可能是少数服从多数吧，毕竟汉语言文学里面很难有标准的答案。如果是我会选择中小学语文教师作为标注者更为合适（因为绝大多数人的比喻都是他们教的）。

在测试数据集时，从每一种结果随机选出250个样本，使用了3个众包信息评估员（473页）进行测评（fluency, creativity, consistency, overall），为什么这里又不用专业的了？

我觉得根据自身文学底蕴不同，对于明喻的理解（实验中使用到的四个属性）会有差异，进而对标注数据的结果会有一定影响，最好使用更多的人，然后取加权平均。

未理解部分：

没明白475页table 9为什么要把Scorer和LFM做对比，LFM不是Scorer的其中一步吗？这里不是应该比较不同的生成喻体的模型吗？

答：为了证明scorer模块中的step1也是有效的，如果单取LFM(step 2)效果没达到预期。