基于大纲的条件故事生成

1. 任务描述

故事生成是指给定少量输入信息（如故事开头、关键词等），生成一个完整的连贯合理的故事。故事生成是自然语言处理和人工智能领域的重要前沿课题，对于发展具备语言智能的类人 AI 有重要意义，在娱乐、教育等方面也有广泛的应用价值，近年来受到学术界和工业界的广泛关注。然而开放端故事生成任务，给定的输入信息很少，具有一对多的特性，即对于同一输入，合理的输出可能非常多样，这似的自动评价非常困难。因此本任务给定了一个无序的关键短语集合（即故事大纲）作为输入，要求机器能够合理地利用这些情节，产生一个自然、流畅的趣味故事。

本任务增大了输入的信息量，从而很好地缩减了输出的空间，不仅给自动评价提供了便利，而且也能更好地检验模型生成的可控性，同时还对模型进行情节规划的能力提出了更高的要求。

2. 资源

本任务的数据来源于从网上爬取的中文童话、寓言故事等，我们利用 RAKE 算法从故事中自动抽取得到故事大纲，作为任务输入。我们限制每个故事至多抽取 8 个情节，每个情节不超过 8 个词（使用 jieba[1]来进行分词）。

（1）数据规模

	训练集	验证集	测试集
样本数	1,456	242	729
词表大小	19,320	5,698	12,488
大纲的平均词数	19.20	19.05	19.47
大纲的平均短语数	8.00	8.00	8.00
故事的平均词数	108.91	108.68	109.04

[1] https://github.com/fxsjy/jieba

	训练集	验证集	测试集
故事的平均句子数	7.20	7.11	7.15

（2）数据格式

数据为 csv 格式，包括 title，story，keywords 三列，不同 keywords 以逗号分割。每一行对应一个样本。各字段含义：(1) title: 故事标题；(2) story：故事正文；(3) keywords：从故事正文中抽取的关键词。

示例如下：

title	story	keywords
鹿和马	有一回，鹿和马为了一块草地争吵得不可开交，各人都想将这块草地占为己有。最后鹿仗着自己那对厉害的角，终于战胜了马。这对马来说，简直是无法容忍的。怎样才能重新把鹿赶走呢？马考虑来考虑去；终于想到去求助于人。它找到了一个很强壮的男人。这个人来到草地，同鹿干了一仗，将鹿杀死了。从此，这块引起纠纷的草地，完全归马独自占有了。不过，那位帮助马取得胜利的男人，也已将马占为己有了。马得到了好处，不再贫困，却也失去了自由。	这块草地占为己有；这块引起纠纷；草地争吵；终于战胜；终于想到；来到草地；鹿赶走；鹿杀死

对于训练集和验证集，提供 title, story, keywords。对于测试集，提供 title, keywords，参赛选手需要依次给出每个故事的正文生成结果。

（3）基线模型

我们也将提供一个在 120G 小说语料上预训练的语言模型作为基线。该模型基于编码器-解码器架构，有 3 个不同参数规模的版本，small 版本包括 6 千万参数，base 版本包括 2 亿参数，large 版本包括 10 亿参数。[1]

3. 评价指标

BLEU-1/2：衡量输出的故事和标准答案之间的相似性。[2]

Distinct-1/2：衡量输出的故事的多样性。[3]

Coverage：计算生成的故事和给定的情节序列之间的 Rouge-L recall 分数[4]。衡量输出的故事包含了多少给定大纲中的情节。

Order：用 1 减去逆序（以标准答案的短语出现顺序作为正确顺序）短语对的数目除以总短语对的数目，这里情节的出现位置通过比对最长公共序列来确定。衡量短语出现顺序是否得当。

4. 时间安排

	时间	说明
报名注册	10 月 10 日至 10 月 20 日	发布任务说明，接受参与者报名
发布训练集	10 月 21 日	发布训练数据集
发布测试集	11 月 9 日	发布测试数据集
提交结果	11 月 10 日	提交测试数据集的结果和技术报告
评测会议	11 月 20 日	发布比赛最终结果，进行会议

5. 提案发起单位

发起单位：清华大学

评测委员会：

黄民烈，清华大学

关健，清华大学

联系方式：j-guan19@mails.tsinghua.edu.cn

6. 反作弊声明

(1) 参与者禁止注册多账户报名，经发现将取消成绩并严肃处理。

(2) 参与者禁止在指定考核技术能力的范围外利用规则漏洞或技术漏洞等不良途径提高成绩排名，经发现将取消成绩并严肃处理。

(3) 可以接触到赛题相关数据的人员，其提交结果将不计入排行榜及评奖。

7. 交流平台

主办方建立技术讨论群（微信群），供选手讨论、沟通，主办方也将安排工作人员定期在群内答疑，且后续的相关活动信息均会在群内发布。

8. 参考文献

[1] Guan et al. LOT: A Benchmark for Evaluating Chinese Long Text Understanding and Generation. 2021.

[2] Papineni et al. Bleu: a method for automatic evaluation of machine translation. ACL 2002.

[3] Li et al. A diversitypromoting objective function for neural conversation models. NAACL 2016.

[4] Lin et al. Rouge: A package for automatic evaluation of summaries. ACL 2004.