

文章编号: 1003-0077(2019)05-0066-09

一种使用多跳事实的端到端知识库实体描述生成方法

孟庆松¹, 张翔¹, 何世柱², 刘康², 赵军²

(1. 哈尔滨理工大学 自动化学院, 黑龙江 哈尔滨 150080;
2. 中国科学院自动化研究所 模式识别国家重点实验室, 北京 100160)

摘要: 自动化实体描述生成有助于进一步提升知识图谱的应用价值, 而流畅度高是实体描述文本的重要质量指标之一。该文提出使用知识库上多跳的事实来进行实体描述生成, 从而贴近人工编撰的实体描述的行文风格, 提升实体描述的流畅度。该文使用编码器—解码器框架, 提出了一个端到端的神经网络模型, 可以编码多跳的事实, 并在解码器中使用关注机制对多跳事实进行表示。该文的实验结果表明, 与基线模型相比, 引入多跳事实后模型的 BLEU-2 和 ROUGE-L 等自动化指标分别提升约 8.9 个百分点和 7.3 个百分点。

关键词: 知识图谱; 实体描述; 数据到文本生成

中图分类号: TP391 **文献标识码:** A

An End-to-End Method of Entity Description Generation with Multi-hop Facts on Knowledge Bases

MENG Qingsong¹, ZHANG Xiang¹, HE Shizhu², LIU Kang², ZHAO Jun²

(1. School of Automation, Harbin University of Science and Technology, Harbin, Heilongjiang 150080, China;
2. State Key Laboratory of Pattern Recognition Chinese Academy of Sciences, Beijing 100160, China)

Abstract: Automatic generation of entity description is beneficial to the application of knowledge graphs. Good descriptions are usually written in fluent language, which is an important indicator of text quality. This paper proposes to utilize the multi-hop facts on knowledge graphs to generate entity descriptions, which are expected to match the writing style of human editors and improve the text fluency. Specifically, this paper adopts the encoder-decoder framework and proposes an end-to-end neural network model, encoding multi-hop facts with an attention mechanism in the decoding phase. Experiments show that, compared with the baseline, the proposed model trained with multi-hop facts obtains promising improvement in BLEU-2 by 8.9% and ROUGE-L by 7.3%, respectively.

Keywords: knowledge graph; entity description; data-to-text generation

0 引言

随着互联网技术和 O2O 服务的迅速发展, 网络上积累了大量组织良好的结构化数据。这些数据的组织形式多种多样, 如概念标签体系、商品信息库、微博关系网、知识图谱等, 其中, 以关联互联网中所有知识资源为目标的开放链接数据计划^①为典型的代表项目, 其基本组成部分就是各领域知识图谱。知识图谱包含高质量、组织良好的结构化知识数据,

是人工智能的重要基础设施, 在网页搜索^[1]、推荐系统^[2]、问答系统^[3]等应用中发挥着重要作用。

本文研究如何利用知识图谱中实体的结构化事实信息生成其自然语言描述文本。图 1 展示了在维基媒体基金会所维护的知识图谱(或称知识库) WikiData^[4]上浏览著名作曲家“约翰·塞巴斯蒂安·巴赫”的实体的局部效果^②。除了包含巴赫的

① <http://linkeddata.org/>

② <https://www.wikidata.org/wiki/Q1339>

大量结构化事实,还包含了该人物的一句话描述文本“德国巴洛克作曲家和音乐家”。实体描述生成(Entity Description Generation, EDG),就是给定知识库的实体及其相关的结构化数据,自动地生成其描述内容,一般为一句话的简介。

Johann Sebastian Bach (Q1339)	
German Baroque composer and musician J. S. Bach Bach J.S. Bach J S Bach	
In more languages	
Statements	
instance of	human 1 reference
image	Johann Sebastian Bach.jpg 0 references
sex or gender	male 2 references

图1 知识库中实体的事实及描述示例

尽管知识图谱包含了丰富的对机器友好的语义信息(例如,以机器码和符号描述实体及关系、以三元组形式描述事实,如图1所示),但是这些符号表示的结构化数据难以被人类所阅读和理解,给知识呈现带来了不小的挑战。事实上,在信息交流过程中,大部分人更乐于阅读和表达自然语言。例如,虽然图1的网页中可以找到上百条与巴赫有关的信息(如巴赫创作的众多音乐作品、他的妻子和20个孩子等),但是这些数据的堆砌显然不如文本描述直观。实际上,大部分想要了解该人物的用户首先更乐于阅读文字描述。此外,实体描述还对很多下游任务有帮助,如知识图谱人物关系的解释^[5]、时间线的描述^[6]、问答系统^[7]等任务都能通过文字描述提升用户体验。

然而,完全依靠人工为实体编写描述是难以完成的。表1给出了当前较为流行的几个通用领域知

识库的数据规模。容易看到,虽然 WikiData 等在线知识库允许用户对其做人工修改,但仅仅依靠人力难以给大量实体加上描述信息。相较而言,大部分实体都包含结构化的事实信息。如果能够利用实体已有结构化数据自动生成相对应的实体描述内容,将有利于进一步提升知识图谱的应用价值。因此,如何利用自然语言描述结构化知识图谱中的概念和事实具有重要研究意义和应用价值。

表1 常见知识图谱的数据规模及实体描述数目

知识库	实体数目	实体描述数目
YAGO	1 000 万	—
DBPedia	2 865 万	≈500 万
Freebase	1.25 亿	≈800 万
WikiData	4 596 万	≈1 400 万

实体描述的应用需求是为读者提供该实体的基本信息,自然这个任务就要求模型生成的句子具有功能效用和流畅性。其中功能效用指句子应当满足读者对该实体了解的需要。本文则主要提升句子的流畅性,它要求句子对阅读者自然且不晦涩。NIST 曾在 DUC2007 任务^[8]中列举了对生成句子的一些要求,在流畅性方面则有如语法正确、句子简练等要求。在该任务的最终评测中,所有的句子都会由人工判断是否符合给定标准。

由于流畅性尚无自动化的评价指标,本文尝试对人工构造的实体描述进行观察。以 WikiData 知识库中的实体“Peter Dinklage(彼特·丁拉基)”为例,通过人工构建的维基数据和维基百科之间的对应,得到维基百科对于“彼特·丁拉基”的介绍:“彼特·海顿·丁拉基,美国演员。因出演 HBO 的热门影集《权力的游戏》而名声大噪。”在这个句子中,下划线部分的内容都是与“彼特·丁拉基”有关的信息,从实体的属性里可以直接得到,但是“HBO”则与这个演员并无直接关系。实际上,只有在“权力的游戏”实体的结构化数据中可以获得,即 HBO 是《权力的游戏》的出品公司,并且“权力的游戏”与“彼特·丁拉基”实体有直接关联,可以从该演员的实体出发,在知识图谱上遍历得到。因此,由于实体描述生成的目标接近人工撰写的效果,假若以维基百科的内容为范本,提升实体描述的流畅性,那么从对应实体在知识图谱上的邻居中进一步提取属性值信息将是非常有用的。也就是说,一个流畅的实体描述句子势必需要利用知识库上的多跳属性。

若观察人工撰写的实体描述句子,多跳属性提供的信息一般会以形容词或者定语从句等语法形式出现,例如,上文中提到的“HBO”就是一个形容词作修饰成分。当然,这些内容是可选的,删除以后并不会影响句子表达的语义。表 2 给出了本文用到的数据集中实体的直接属性和多跳属性与目标句子的词汇覆盖程度差异。可以看到,加入多跳属性后词汇量的覆盖会有显著提升。这说明,在大量人工撰写的实体介绍中,多跳属性具有重要作用。因此我们可以认为,直观上流畅的实体描述一定包含了知识库中的多跳属性。

表 2 验证集上多跳属性词汇覆盖占比(%)

领域	直接属性词汇覆盖度	多跳属性词汇覆盖度
建筑	13.2	23.6
人类	19.8	28.1

本文使用深度神经网络中的编码器—解码器框架对实体描述生成进行建模,使用编码器对知识库中与实体有关的事实进行编码,使用关注机制和 LSTM 网络对目标实体描述进行解码。为了解决实体描述标注数据缺少的问题,本文从 WikiData 知识库上整理得到特定领域的子集,进行训练,最终验证了多跳事实对于提升流畅度的重要作用。

1 相关工作

1.1 文本生成任务框架

实体描述的生成任务形式可以归类为数据到文本生成(data-to-text generation)的一种。这类任务在知识库诞生之前就有多种应用,例如,从结构化数据生成天气预报^[9],向病人解释药物的作用^[10],总结数据库表中的统计数据^[11]等。

Reiter 和 Dale 系统性地提出^[12]了数据到文本生成系统的三个主要步骤。第一步是文档规划,主要是先选择用于生成文本的结构化数据,再安排结构化数据之间的逻辑关系,即内容选择、篇章规划两个小步骤。第二步是微观规划,需要完成可能的句子聚合、确定数据具体用词、生成指代表达等过程。最后一步是表述具现化,即用前两步的结果生成具体的符合语法的句子。这三个步骤在传统的生成系统中都有体现,但较新的文献多选择进行联合学习,从而避免传统的管道(pipeline)系统中误差传递的问题。图 2 展示了传统文本生成系统中的这三个步

骤间的顺序关系。

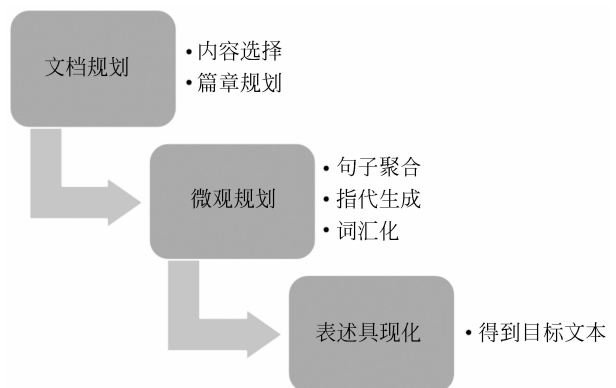


图 2 数据到文本生成系统的主要步骤

1.2 数据到文本生成

数据到文本生成的经典方法对于实体描述生成有不少可借鉴之处。这些方法在很多方面都有所区别,如所针对的数据类型、选用的生成技术、研究的应用场景、使用的评价手段等,所使用的文本生成技术是它们的主要区别^[13]。

基于模板的生成方法是文本生成的主要方法。模板在形式上一般包含了句子的表达模式,在关键的地方留有空缺。当给定一个新的结构化数据,将一些直接属性代入模板的对应空缺之中,就可以得到一个针对该数据的较为流畅的新句子。Angeli 等^[14]使用联合概率建模的方法,加入了多种决策,把文档规划和表述具现化结合起来。对于选择数据记录、选择字段、选择模板等细节,使用人工选定的特征和对数线性模型进行排序和优选。然而他们生成的评论仅限于天气预报生成和机器人球赛两个领域,可以简单且有针对性地构造模板。即只需将特定域的取值从训练文本抽去,再抽去数字就可以构成一个模板。Duma 和 Klein^[15]则在抽取之前先对齐数据和文本,再对抽取的结果使用人工规则进行过滤,最终在 268 篇文章里抽取出了 74 个模板,其中有 43 个模板经过人工审查发现符合语法正确的要求。Saldanha 等^[16]则主要针对公司实体,使用成对的种子词进行 bootstrap,不断执行模式抽取、词对发现两个步骤,再使用打分和删除规则来提取实体描述的模板。

生成语则是自然语言处理中较为重要的方法,具有比模板好得多的泛化能力。Konstas 和 Lapata^[17]在生成描述时使用了概率上下文无关语法(PCFG),每条产生式的概率服从多项分布。他

他们还加上了基于超图的排序,用于优选。Gyawali 和 Gardent^[18]使用基于特征的词汇化树连接语法 (Feature-based Lexicalized TAG) 进行句子生成。他们首先进行了对齐子树结构和知识库的三元组数据,再抽取 FB-LTAG 语法。他们单独训练了一个语言模型,对生成的结果进行排序,选择出最自然的表达。同时为了处理训练集中没有出现过的新模式,他们还设计了基于规则的算法,从训练集上学习到的语法中猜测和扩充新语法。

随着神经网络方法的兴起,数据到文本生成任务中也有一些研究采用深度神经网络来进行的多步骤联合学习。Mei 等^[19]使用双向长短时记忆 (Bi-LSTM) 网络模型,使用预选概率进行粗筛选,再结合关注机制进行精选。他们主要在天气领域的 WeatherGov 数据集上做了生成实验,同时在 RoboCup 的机器人球赛数据上考察了文本生成模型对不同领域的适应能力。

1.3 实体描述生成

实体描述的生成任务大多以神经网络为基础,为了克服模板和生成语法常常限于具体表达的不足,Lebret 等^[20]针对人物领域,为每个人物生成描述信息。他们使用维基百科页面上的信息框 (InfoBox) 作为输入,输出目标则是同页面上的维基百科描述。信息框的内容由人工整理,往往会因人物职业不同而关注不同的属性,例如,针对科学家应标出其专业方向,对球类运动员则会给出他加入的球队;但若人物具有相近职业则可能具有相同属性,不过取值却不同,例如篮球运动员和足球运动员各自的球队名称就有较大差别。基于这两类考虑,他们设计了一个条件概率,以针对信息框中属性、属性值中的词来建立语言模型。他们还使用了拷贝机制^[21],从信息框中直接拷贝未登录词到目标句子里,以解决句子中经常需要生成低频、未登录词的问题。

Chisholm 等^[22]使用神经网络中常用的 Seq2Seq 框架进行人物描述的生成。他们使用 WikiData 作为输入数据,但将结构化数据中每个属性和属性的取值依次相连,从而拼接成序列。借助 Seq2Seq 框架和神经网络的强大拟合能力,他们能学习到结构化数据到实体描述句子之间的映射。例如,对于 WikiData 中的实体“Matias Tuomi”,他们用图 3 所示的方式构造输入数据。同时他们借鉴自编码器 (auto-encoder) 的思想,将目标句子又转

换回原序列,以此扩展原始模型。这样的训练方法可以把文本生成和关系抽取两个任务结合起来进行联合训练。实验效果表明这样的联合训练方式会优于单独使用文本生成任务进行训练得到的模型。

```
TITLE mathias tuomi SEX_OR_GENDER
male DATE_OF_BIRTH 1985-09-03
OCCUPATION squash player
CITIZENSHIP finland
```

图 3 知识库中的事实相连构成的输入序列

实体描述的生成与传统的生成工作相同,因此传统工作中的句子生成技术(如模板、生成语法、神经网络等)可以借鉴到实体描述生成的工作中。但另一方面,传统的自然语言生成工作使用记录、 λ 表达式等作为输入,而实体描述的生成需使用大规模知识图谱。在知识图谱中,基本的组成元素是事实三元组,若从一个实体出发,在图谱上遍历,可能的遍历路径呈指数级上升。因此,如何选择合适的三元组进行句子生成是一个更具挑战性的任务。此外,由于知识图谱往往是开放领域的,所以,其句子表达将更具多样性。

2 利用多跳关系的实体生成模型

本节基于编码器—解码器的神经网络框架,提出一个利用多跳属性的实体描述生成模型,并利用了关注机制。

2.1 形式化记号

首先对记号代表的含义及实体描述生成的任务进行形式化说明。给定一个知识库四元组 $K = (E, R, F, V)$, 其中 E 为实体集合, R 为二元关系集合, V 为字符串、数值组成的常量集合。集合 F 为知识库中的事实的集合, 且每个事实为由头实体、关系、尾实体组成的三元组 $\langle h, r, t \rangle$, 并且 $F = \{ \langle h, r, t \rangle \mid h \in E, r \in R, t \in E \cup V \}$ 。

针对每个实体 $e \in E$, 定义一个函数 $f: E \rightarrow R \times (E \cup V)$, 以三元组形式返回实体相关的属性。并且定义 $f_r: R \times (E \cup V) \rightarrow R$ 和 $f_t: R \times (E \cup V) \rightarrow (E \cup V)$ 两个函数, 分别返回三元组的关系类型及具体的尾实体。

给定数据集 $D = \{ (e_1, s_1), \dots, (e_n, s_n) \}$, 有 n 组输入和输出数据, 每对数据由实体 $e \in E$ 和该实体

对应的句子 $s = (w_1, w_2, \dots, w_L)$ 组成, 其中每个 w_i , $i = 1, 2, \dots, L$ 表示第 i 个词, L 表示句子长度。

实体描述生成就是给定输入数据集 D , 利用函数 f 获取每个实体 e 在知识库 K 上的相关属性, 生成尽可能接近标准句子 s 的输出。

而我们已经提到, 生成流畅的实体描述需要使用多跳属性。因此除了实体 e 的直接数据 $f(e)$ 之外, 我们还可以进行更远的遍历, 以利用多跳数据 $f(f_i(e))$ 。

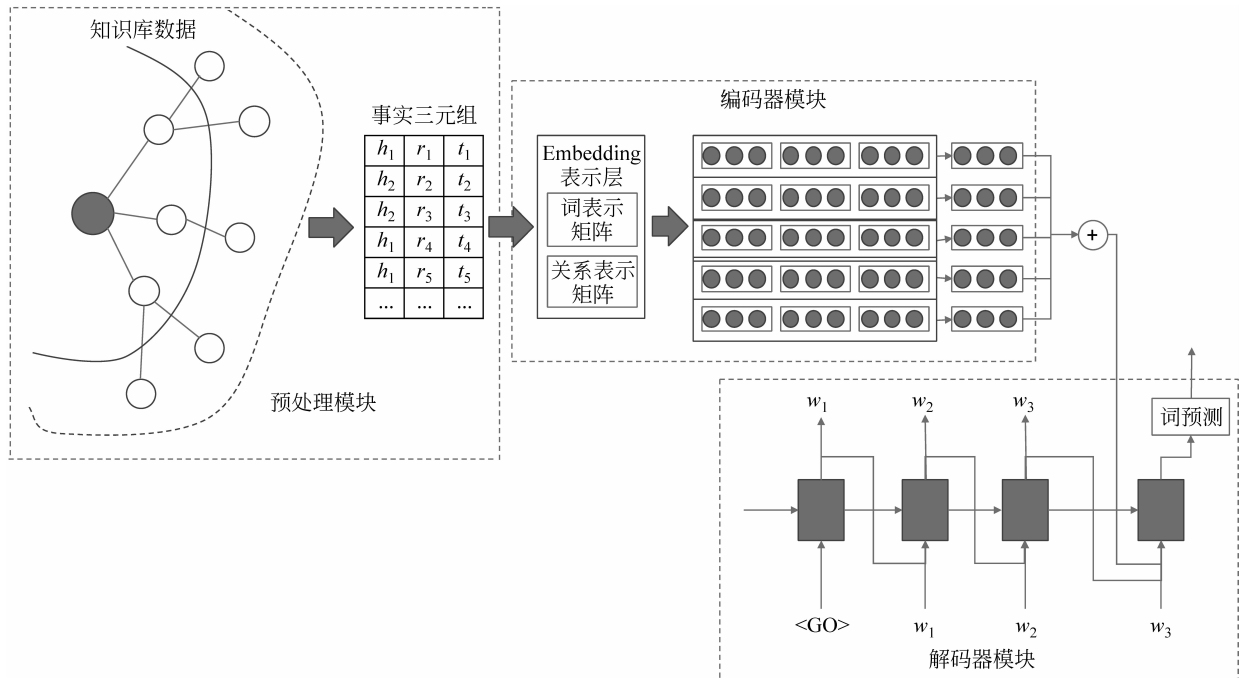


图4 使用多跳属性的实体描述生成模型整体框架

2.2.1 预处理模块：构建事实三元组集合

数据预处理模块需要构建每个样本对应的输入和输出。为了利用多跳属性的数据, 我们使用深度优先搜索算法遍历 WikiData 知识库。由于数据量巨大, 我们限制了所选属性值在从实体出发最远仅两跳路径范围内, 最终得到每个实体相关的事实三元组集合。我们直接使用遍历顺序为这些三元组赋予一个排序。后续模型在关注机制的帮助下, 并不会被这个顺序所影响。

2.2.2 编码器模块

编码器模块为输入的三元组赋予向量表示。我们独立看待每个三元组, 假设它们并没有严格的关联, 编码器独立对待每个三元组的表示。由于目标是生成单词序列, 因此三元组中的符号、关系等取值需要单独设计 embedding。而词序列的 embedding 则包含了头实体和尾实体的取值。这样在解码时可

2.2 模型

本文所提出的模型则使用编码器—解码器的神经网络框架。整体模型结构如图4所示, 主要包括3个模块: ①预处理模块, 针对不同实体从知识库中选择特定的属性, 构建多跳属性集合; ②编码器模块, 负责编码输入的三元组属性集合, 得到每个三元组的向量表示; ③解码器模块, 依次解码句子中的每个词。下面详细说明每个模块。

以避免解码出属性符号, 只得到预测词。

如果三元组的尾部不是实体, 而是其他取值, 例如字符串、数值等, 这个值则会放到词表中。若尾部是另一个知识库实体, 则取该实体的名称加入词表。

每个三元组 $T = \langle h, r, t \rangle$ 通过上述 embedding 层变换为三个向量即 $\langle v_h, v_r, v_t \rangle$ 。我们将三元组的向量视为一个序列, 我们简单地通过求平均的方式获得单个三元组的表示, 在实验中这种方法会比使用一个 LSTM 模型对三个向量进行融合还要好, 具体如式(1)所示。

$$v = \frac{1}{3}(v_h + v_r + v_t) \quad (1)$$

编码器独立地对每个三元组取出 embedding 以及平均操作, 最终得到所有输入的三元组向量序列 $\{v_1, v_2, \dots, v_M\}$ 。

2.2.3 解码器模块

我们使用标准的 LSTM 作为解码器,每个循环单元由式(2)~式(7)所确定。

$$i_t = \sigma(W_{wi}w_t + W_{hi}h_{t-1}) \quad (2)$$

$$f_t = \sigma(W_{wf}w_t + W_{hf}h_{t-1}) \quad (3)$$

$$o_t = \sigma(W_{wo}w_t + W_{ho}h_{t-1}) \quad (4)$$

$$\hat{c}_t = \tanh(W_{wc}w_t + W_{hc}h_{t-1}) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t \quad (6)$$

$$h_t = o_t \odot \tanh(c_t) \quad (7)$$

LSTM 的隐层状态和记忆状态都初始化为零向量。当输入为<GO>标记时开始进行解码,在解码第 i 个词时 LSTM 的隐层状态记为 h_i 。

关注机制是解码器中的重要组成部分。在解码第 i 个词时,需要使用前一步的隐层状态 h_{i-1} 。我们使用了矩阵关注机制。首先以 0 偏置的双线性函数计算对于每个输入的三元组的相似度,如式(8)所示。

$$\alpha_{ij} = h_{i-1}^T M v_j \quad (8)$$

再使用 softmax 函数处理前一步得到的相似度,得到归整化的关注权重,如式(9)所示。

$$\beta_{ij} = \frac{\exp(\alpha_{ij})}{\sum_k \exp(\alpha_{ik})} \quad (9)$$

利用这个权重对所有输入的三元组向量进行加权求和,得到在 h_{i-1} 的情形下对于输入三元组的整体表示,作为当前词的上下文向量,如式(10)所示。

$$u_i = \sum_j \beta_{ij} v_j \quad (10)$$

接下来将输入词通过词表示矩阵后得到的表示 w_i 与上下文向量进行拼接,合为一个输入送入 LSTM 单元中进行计算,如式(11)所示。

$$h_i = \text{LSTM}([u_i; w_i]) \quad (11)$$

最后,为了得到下一个词,我们使用一个简单的 Softmax 层,即用线性映射到词表大小的 logits 向量,再使用 Softmax 函数预测每个词的概率,如式(12)所示。

$$p_i = \text{Softmax}(W h_i) \quad (12)$$

在训练和测试时取其中概率最大的词作为当前步骤的输出。

2.3 训练

我们使用交叉熵作为模型的优化目标,并使用基于随机梯度下降的方法来优化。具体的目标函数

如式(13)所示,其中 $p(w_i | w_1, \dots, w_{i-1}, e)$ 表示模型在给定输入实体和前 $i-1$ 个词时,输出第 i 个词的概率。

$$L = \sum_{(e,s) \in D} \sum_i \log p(w_i | w_1, \dots, w_{i-1}, e) \quad (13)$$

3 实验

3.1 数据集构建

在实际的公开数据中,知识图谱的实体描述都不甚完美。为了保证任务具有良好定义,即通过实体的结构化数据能确实输出其描述,我们利用维基百科和维基数据(WikiData)之间的映射来构建数据集。我们使用了 WikiData 的 20180416 的 dump 数据作为输入。

对于每个实体,我们首先从 WikiData 数据中的“类—实例(P31)”和“类—子类(P279)”两种三元组作为父类别的依据,构建了类别的分类树。我们选择了几个类别来按领域筛选,把类别、子类别、该类别的实体、子类别实体都作为此领域的实体。后续只对分到某个领域中的实体集合进行处理。

WikiData 尽量给出了每个实体和这个实体的维基百科页面的对应关系。但有的实体分类比较细,未必有专门维基百科页面(例如,各种复杂的有机物、罕见元素的同位素等)。此外,我们也希望将来可以针对同一实体进行多语言描述生成工作。基于这两个要求,对于每个领域的实体,我们让实体同时具有中文和英文的对应维基百科页面。由于中文维基百科的覆盖度并不高,这个条件能过滤掉很多的实体。为了保证训练数据规模大小足够用于神经网络训练,我们按照实体数量挑选了一些领域。

由于筛选出的实体均有对应的维基百科页面,我们使用一些过滤规则扔掉维基百科维护和管理所加入的标注句子,将页面正文中开头一定长度的文本(一般为一两个句子)作为训练的目标,得到一组质量相对较好的实体描述标注。

通过分析最终得到的数据量,我们随机地切分了实体集合,构造出训练集、验证集和测试集。表 3 给出了我们的数据集规模。书籍领域由于样本过少,则不用于训练。

表 3 数据集筛选前后规模

领域	总实体数	筛选后实体数	训练集	验证集	测试集
书籍	161 437	2 615	—	—	—
建筑	1 216 713	28 696	23 000	2 000	3 696

3.2 实现细节

我们使用验证集来选择模型的 embedding 维度,从 50、100、300 中选择了效果最好的一组即 300,并将 LSTM 隐层表示的维度也设为 300。LSTM 输出加入了 dropout 并设比率为 0.2。同时由于显存大小的限制,我们设置了每个实体最多能输入 100 个三元组, batch 大小设为 50。我们使用 Adam 进行优化,并使用其推荐的参数配置^[23],将 alpha、beta1、beta2 分别设置为 0.001、0.9 和 0.999。

我们也对三元组的关系进行了一定限制,除了统计关系的出现频次外,也排除了一些无用的属性如书籍的 ISBN 号等,实际数据中只使用出现频次最多的 36 种关系。

3.3 生成效果评价和结果分析

由于我们把维基百科给出的实体描述作为标准句子,这里使用自动化指标 BLEU 值^[24]、ROUGE^[25]值给出评价。

本节实验的主要关注点是引入多跳属性是否能提升模型生成实体描述的能力。基础模型(Base)也是用本文提出的模型结构,但仅使用实体自身的结构化三元组集合,而多跳模型(MultiHop)则使用 3.1 节中构建出的完整三元组集合。二者训练时使用的目标句子均相同。表 4 是两个模型在开发集和测试集上由不同的自动化指标给出的性能。

表 4 验证集和测试集上各项自动评价结果

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Base (dev)	44.4	31.0	22.4	16.7	47.8
Multi-Hop(dev)	50.8	39.1	30.7	24.4	54.1
Base (test)	43.4	29.8	21.0	15.5	46.7
Multi-Hop(test)	50.7	38.7	30.1	24.0	54.0

可以看到,多跳模型无论在验证集还是测试集上,各个自动化指标都有明显的性能提升。这说明我们的模型通过编码多跳数据的表示,能覆盖更多人工描述中所用的词汇,从而贴近人工描述的语言风格,生成更流畅的句子。

在具体实现时,由于多跳模型使用的数据源更大,引入的结构化信息更多,词汇表也不可避免地增大很多,两种情形的模型不可混用。由于词汇表一般由数据中的高频词构成,其他词汇作为未登录词统一看作一个符号。相比之下,训练数据的词汇在多跳模型下未登录词比例会稍高一些。但即便如此,多跳模型依然在评价指标上有大幅提升。这说明了人工编写的实体描述的用词确实更多地存在于多跳属性值中。因此,要提升生成句子的流畅度,也就必须用到多跳数据。

具体地,BLEU 值使用了生成句和标准句之间的 n-gram 重复出现数目进行得分计算,ROUGE 则基于两者的最长公共子串进行计算,它们都可以视为基于词重叠的评价方法,在机器翻译和文摘等任务中都有应用。但是,在生成任务中,由于句子限制不多,各种多样表达都可接受,因此基于词重叠的评价结果与人工评价的结果之间相关性不强^[26],需要人工介入评价。然而,对大量样本做人工评价成本很高,是不实际的。考虑到 BLEU 和 ROUGE 等指标的计算方法,当词重叠度相当高时,生成句已经和标准句几乎相同,可以认为效果较好。

表 5 给出了一些多跳模型生成句子的例子,可以看到所有的标准句和生成句都较为接近。但是,生成句 1 给出了错误的信息,即“奥地利国家图书馆”;生成句 2 也给出了错误的信息,即“地铁 3 号线”;而非正确的“2 号线”。因此,多跳数据虽然能在一定程度上一提升流畅度,但是知识性的提升仍然有待进一步研究。

表 5 多跳模型生成样例

句子属性	句子内容
标准句 1	the university of graz library, in graz, austria is the largest scientific and public library in styria and the third largest in austria.
生成句 1	the <UNK> library in graz, austria is the national library of austria.
标准句 2	<UNK> station, literally <UNK> road station in english, is a station of line 2 western section of the tianjin metro.

续表

句子属性	句子内容
生成句 2	<UNK> station, literally <UNK> road station in english, is a station of line 3 of the tian-jin metro.
标准句 3	<UNK> station is a railway station on the taiwan railways administration yilan line located in <UNK> township, yilan county, taiwan.
生成句 3	<UNK> station is a railway station on the taiwan railway administration yilan line.

4 总结和展望

实体描述是知识库或知识图谱的重要组成部分,本文讨论了自动生成实体描述的必要性,并提出知识库的多跳信息对于实体描述生成非常有益,可以提升句子的流畅度,并接近人工撰写的实体描述。本文的实验则辅助证明了多跳信息的益处。

但是,由于自动指标的 BLEU 值本身只能衡量生成句子与标准句子之间的词汇相似度,不一定能准确体现句子流畅性。如何衡量句子的流畅性将其并用到模型训练中,依然是未来需要研究的问题。另一方面,实体描述的最终目的是满足其功能效用。然而,神经网络模型的可解释性不高,难以解释为何会输出这样的句子,有时会与已知的三元组相矛盾。如何提升神经网络的准确性,与模板或语法系统相结合,也是未来实体描述生成的一个研究方向。

参考文献

- [1] Su Y, Yang S, Sun H, et al. Exploiting relevance feedback in knowledge graph search[C]//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2015: 1135-1144.
- [2] Zhang F, Yuan N J, Lian D, et al. Collaborative knowledge base embedding for recommender systems[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2016: 353-362.
- [3] Yih W-t, Chang M-W, He X, et al. Semantic parsing via staged query graph generation: Question answering with knowledge base[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing: Association for Computational Linguistics, 2015: 1321-1331.
- [4] Vrandečić D, Krötzsch M. Wikidata: A Free collaborative knowledgebase[J]. Commun ACM, 2014, 57(10): 78-85.
- [5] Voskarides N, Meij E, Tsagkias M, et al. Learning to explain entity relationships in knowledge graphs[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing, 2015: 564-574.
- [6] Althoff T, Dong X L, Murphy K, et al. TimeMachine: Timeline generation for knowledge-base entities[C]//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 19-28.
- [7] Haug T, Ganea O-E, Grnarova P. Neural multi-step reasoning for question answering on semi-structured tables[C]//Proceedings of the Advances in Information Retrieval. Cham: Springer International Publishing, 2018: 611-617.
- [8] Over P, Dang H, Harman D. DUC in context[J]. Information Processing and Management, 2007, 43(6): 1506-1520.
- [9] Goldberg E, Driedger N, Kittredge R I. Using natural-language processing to produce weather forecasts[J]. IEEE Expert, 1994, 9(2): 45-53.
- [10] Buchanan B G, Moore J D, Forsythe D E, et al. An intelligent interactive system for delivering individualized information to patients[J]. Artificial Intelligence in Medicine, 1995, 7(2): 117-154.
- [11] Iordanskaja L, Kim M, Kittredge R, et al. Generation of extended bilingual statistical reports[C]//Proceedings of the 14th Conference on Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992: 1019-1023.
- [12] Reiter E, Dale R. Building applied natural language generation systems[J]. Natural Language Engineering, 1997, 3(01): 57-87.
- [13] 张翔. 基于大规模知识库的实体描述生成和应用[D]. 哈尔滨: 哈尔滨理工大学硕士毕业论文, 2018.
- [14] Angeli G, Liang P, Klein D. A simple domain-independent probabilistic approach to generation[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010: 502-512.
- [15] Duma D, Klein E. Generating natural language from linked data: Unsupervised template extraction[C]//

- Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013); Association for Computational Linguistics, 2013: 83-94.
- [16] Saldanha G, Biran O, McKeown K, et al. An entity-focused approach to generate company descriptions [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol 2), 2016: 243-248.
- [17] Konstas I, Lapata M. Concept-to-text generation via discriminative reranking [C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics; Long Papers - Volume 1; Association for Computational Linguistics, 2012: 369-378.
- [18] Gyawali B, Gardent C. Surface realisation from knowledge-bases [C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, 1: 424-434.
- [19] Mei H, Bansal M, Walter M R. What to talk about and how? Selective generation using LSTMs with Coarse-to-Fine Alignment [C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies, 2016: 720-730.
- [20] Lebrecht R, Grangier D, Auli M. Neural text generation from structured data with application to the biography domain [C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 1203-1213.
- [21] Gu J, Lu Z, Li H, et al. Incorporating copying mechanism in sequence-to-sequence learning [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, 1: 1631-1640.
- [22] Chisholm A, Radford W, Hachey B. Learning to generate one-sentence biographies from Wikidata [C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics; Volume 1, Long Papers; Association for Computational Linguistics, 2017: 633-642.
- [23] Kingma D P, Ba J. Adam: A method for stochastic optimization [C]//Proceedings of ICLR, 2015.
- [24] Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002: 311-318.
- [25] Lin C Y. ROUGE: A package for automatic evaluation of summaries [C]//Proceedings of Text Summarization Branches Out, 2004.
- [26] Novikova J, Dušek O, Curry A C, et al. Why we need new evaluation metrics for NLG [C]//Proceedings of Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 2241-2252.



孟庆松(1969—), 博士, 教授, 主要研究领域为控制理论、模式识别等。
E-mail: mqs0530@163.com



张翔(1989—), 通信作者, 硕士研究生, 主要研究领域为自然语言处理、自然语言生成。
E-mail: xiang.zhang@nlpr.ia.ac.cn



何世柱(1987—), 博士, 助理研究员, 主要研究领域为知识图谱、问答系统。
E-mail: shizhu.he@nlpr.ia.ac.cn