Journal of Computer Applications
计算机应用,2021,41(5):1227－1235
ISSN 1001-9081
CODEN JYIIDU
2021-05-10
http://www.joca.cn

文章编号:1001-9081(2021)05-1227-09
DOI:10.11772/j.issn.1001-9081.2020071069

自然语言生成综述

李雪晴[1,2],王　石[2*],王朱君[1,2],朱俊武[1]
(1.扬州大学 信息工程学院,江苏 扬州 225000;　2.中国科学院 计算技术研究所,北京 100190)
(*通信作者电子邮箱wangshi@ict.ac.cn)

摘　要:自然语言生成(NLG)技术利用人工智能和语言学的方法来自动地生成可理解的自然语言文本。NLG降低了人类和计算机之间沟通的难度,被广泛应用于机器新闻写作、聊天机器人等领域,已经成为人工智能的研究热点之一。首先,列举了当前主流的NLG的方法和模型,并详细对比了这些方法和模型的优缺点;然后,分别针对文本到文本、数据到文本和图像到文本等三种NLG技术,总结并分析了应用领域、存在的问题和当前的研究进展;进而,阐述了上述生成技术的常用评价方法及其适用范围;最后,给出了当前NLG技术的发展趋势和研究难点。

关键词:自然语言生成;语言学;自然语言处理;评价方法;文本到文本生成;数据到文本生成;图像到文本生成
中图分类号:TP391　　**文献标志码:**A

Summarization of natural language generation

LI Xueqing[1,2], WANG Shi[2*], WANG Zhujun[1,2], ZHU Junwu[1]
(1. *College of Information Engineering, Yangzhou University, Yangzhou Jiangsu* 225000, *China*;
2. *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract: Natural Language Generation (NLG) technologies use artificial intelligence and linguistic methods to automatically generate understandable natural language texts. The difficulty of communication between human and computer is reduced by NLG, which is widely used in machine news writing, chatbot and other fields, and has become one of the research hotspots of artificial intelligence. Firstly, the current mainstream methods and models of NLG were listed, and the advantages and disadvantages of these methods and models were compared in detail. Then, aiming at three NLG technologies: text-to-text, data-to-text and image-to-text, the application fields, existing problems and current research progresses were summarized and analyzed respectively. Furthermore, the common evaluation methods and their application scopes of the above generation technologies were described. Finally, the development trends and research difficulties of NLG technologies were given.

Key words: Natural Language Generation (NLG); linguistics; natural language processing; evaluation method; text-to-text generation; data-to-text generation; image-to-text generation

0　引言

自然语言生成(Natural Language Generation, NLG)是自然语言处理领域一个重要的组成部分,实现高质量的自然语言生成也是人工智能迈向认知智能的重要标志。作为人工智能和计算语言学的子领域,自然语言生成从抽象的概念层次开始来生成文本[1]。NLG技术具有极为广泛的应用价值,应用于智能问答对话系统和机器翻译系统时,可实现更为智能便捷的人机交互;应用于机器新闻写作[2]、医学诊断报告生成[3]和天气预报生成[4]等领域时,可实现文章报告自动撰写,有效减轻人工工作;应用于文章摘要、文本复述领域时,可为读者创造快速阅读条件等。

按照输入信息的类型划分,自然语言生成可以分为三类:文本到文本生成、数据到文本生成和图像到文本生成。其中,文本到文本生成又可划分为机器翻译[5]、摘要生成、文本简化、文本复述等;数据到文本生成的任务常应用于基于数值数据生成BI(Business Intelligence)报告、医疗诊断报告等;在图像到文本的生成的应用领域中,常见的是通过新闻图像生成标题、通过医学影像生成病理报告、儿童教育中看图讲故事等。国际上对上述技术均进行了多年研究,研究成果主要发表在自然语言处理相关学术会议与期刊上,例如ACL(Annual Meeting of the Association for Computational Linguistics)、EMNLP(conference on Empirical Methods in Natural Language Processing)、NACAL(the North American Chapter of the Association for Computational Linguistic)、CoNLL(Conference on Computational Natural Language Learning)、ICLR(International Conference on Learning Representations)和AAAI(Association for the Advancement of Artificial Intelligence)等。上述每项技术都极具挑战性,在学界和工业界的研究发展中,已经对人们的生活和工作产生巨大的影响。

收稿日期:2020-07-23;**修回日期:**2020-10-21;**录用日期:**2020-10-29。
基金项目:国家自然科学基金资助项目(61872313);国家重点研发计划重点专项(2017YFB1002300, 2018YFC1700302)。
作者简介:李雪晴(1995—),女,江苏泰州人,博士研究生,主要研究方向:自然语言处理;　王石(1981—),男,山东博兴人,副研究员,博士,主要研究方向:语义分析、知识图谱;　王朱君(1996—),男,江苏东台人,硕士研究生,主要研究方向:自然语言处理;　朱俊武(1972—),男,江苏江都人,教授,博士,CCF高级会员,主要研究方向:知识工程、本体论。

NLG 的体系结构可分为传统的管道模型和基于神经网络的端到端(End-to-End, End2End)模型两种。管道模型中的不同模块中包括多个独立步骤,如文本结构、句子聚合、语法化、参考表达式生成、语言实现等。其缺点一是上一步骤结果的好坏会直接影响到下一步骤,从而影响整个训练的结果;二是在于需要耗费大量特定领域的手工标注,难以扩展到新的领域。

随着神经网络研究的发展,研究人员利用端到端的模型进行自然语言处理。端到端的模型处理问题时,不再人为划分子问题,而是将中间的操作包含在神经网络中,省去了代价高而且易出错的数据标注工作。端到端模型通过缩减人工预处理,增加模型的整体契合度,提高系统解决问题的效率。

端到端模型的操作流程:首先,从输入端输入原始数据,然后通过众多操作层进行数据加工,输出端会产生一个预测结果;接着,将预测结果与真实结果相比较得到误差,将误差在端到端模型的每一层反向传播,每一层的表示会根据误差做调整,直到模型收敛或达到预期的效果才结束。端到端模型还可以与基于模板的方法融合以取得更好的效果。2017 年发布的 Task-Completion Bot 方法[6]在 End2End 模型的基础上将基于模板的 NLG 和基于模型的 NLG 进行融合,生成自然语言文本。

下面,本文将介绍一些生成方法和模型,以及分别介绍文本到文本、数据到文本、图像到文本生成。

1 生成方法与生成模型

自然语言生成系统通常在不同阶段使用不同的生成技术达到生成结果符合实际需求的目的。下面介绍几种常用的文本生成技术。

1.1 生成方法

1.1.1 模板生成方法

模板生成方法是最早应用于自然语言生成领域的一种方法[7]。该技术通过将词汇和短句在模板库中进行匹配,匹配后将词汇和短语填入固定模板,从而生成自然语言文本,其本质是系统根据可能出现的几种语言情况,事先设计并构造相应的模板,每个模板都包括一些不变的常量和可变的变量,用户输入信息之后,文本生成器将输入的信息作为字符串嵌入到模板中替代变量。

模板生成方法的优点是思路较简单、用途较广泛,但因技术存在的缺陷使得生成的自然语言文本质量不高,且不易维护。该技术多应用于较简单的自然语言生成环境中。

1.1.2 模式生成方法

模式生成是一种基于修辞谓语来描述文本结果的方法。这种方法通过语言学中修辞谓词来描述文本结构的规律,构建文本的骨架,从而明确句子中各个主体的表达顺序。此方法表示的文本结构中一般包括五种类型的节点:Root、Predicate、Schema、Argument 以及 Modifier。这五种节点中,Root 为结构树的根节点,表示一篇文章位于根节点下有若干个 Schema 节点,Schema 节点表示段落或者句群,位于 Schema 节点下是 Schema 节点或者 Predicate 节点,Predicate 节点代表一个句子,句子是文本的基本组成单位。位于 Predicate 节点下的是 Argument 节点,每个 Argument 节点表示句子中的每一个基本语义成分。如果 Argument 节点有修饰成分,那么子节点 Modifier 就发挥语义成分的修饰的作用。在结构树中,树的叶子节点是 Argument 或 Modifier,树中每个节点都含有若干个槽,槽用来存放标志的各种信息以供文本生成使用。

模式生成技术的最大优点是通过填入不同的语句和词汇短语即能生成自然语言文本,较易维护,生成的文本质量较高。不足是只能用于固定结构类型的自然语言文本,难以满足多变的需求。

1.1.3 修辞结构理论方法

修辞结构理论(Rhetorical Structure Theory, RST)方法来源于修辞结构理论的引申[8],是关于自然语言文本组织的描述性理论。RST 包含 Nucleus Satellite 模式和 Multi-Nucleus 模式[9]两种模式:Nucleus Satellite 模式将自然语言文本分为核心部分和附属部分,核心部分是自然语言文本表达的基本命题,而附属部分表达附属命题,多用于描述目的、因果、转折和背景等关系;Multi-Nucleus 模式涉及一个或多个语段,它没有附属部分,多用于描述顺序、并列等关系。

RST 技术优点是表达的灵活性很强,但实现起来较为困难,且存在不易建立文本结构关系的缺陷。

1.1.4 属性生成方法

属性生成是一项较复杂的自然语言生成方法,其通过属性特征来反映自然语言的细微变化。例如,生成的句子是主动语气还是被动语气,语气是疑问、命令还是声明,都需要属性特征表示。此方法要求输出的每一个单元都要与唯一具体的属性特征集相连,这项技术通过属性特征值与自然语言中的变化对应,直到所有信息都能被属性特征值表示为止。

该方法的优点是通过增加新的属性特征值完成自然语言文本内容的扩展,但需要细粒度的语言导致维护较为困难。

以上四种方法在 NLG 的发展过程中具有十分重要的作用。虽然这些方法存在一定不足,但仍具有较高的应用价值。

1.2 生成模型

1.2.1 马尔可夫链

在语言生成中,马尔可夫链通过当前单词可以预测句子中的下一个单词,是经常用于语言生成的算法。但由于仅注意当前单词,马尔可夫模型无法探测当前单词与句子中其他单词的关系以及句子的结构,使得预测结果不够准确,在许多应用场景中受限。

1.2.2 循环神经网络

循环神经网络(Recurrent Neural Network, RNN)通过前馈网络传递序列的每个项目信息,并将模型的输出作为序列中下一项的输入,每个项目存储前面步骤中的信息。RNN 能够捕捉输入数据的序列特征,但存在两大缺点:第一,RNN 短期记忆无法生成连贯的长句子;第二,因为 RNN 不能并行计算,无法适应主流趋势。

1.2.3 长短期记忆网络

长短期记忆(Long Short-Term Memory, LSTM)网络及其变体能够解决梯度消失问题并生成连贯的句子,旨在更准确地处理输入的长序列中的依赖性,但 LSTM 也有其局限性:LSTM 处理难以并行化,限制了 LSTM 生成系统利用 GPU(Graphics Processing Unit)等现代计算设备的能力。

1.2.4 序列到序列模型

序列到序列(Sequence-to-Sequence, Seq2Seq)模型是由 Google 工程师 Sutskever 等[10]在 2014 年提出,该模型一般是通过 Encoder-Decoder 框架实现,目的是解决大部分序列不等长的问题,如机器翻译中,源语言和目标语言的句子往往并没有

相同的长度。Seq2Seq模型结构如图1所示,该模型更善于利用更长范围的序列全局的信息,并且综合序列上下文判断,推断出与序列相对应的另一种表述序列。

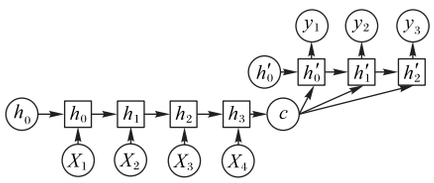

图1　Seq2Seq模型示意图

Fig. 1　Schematic diagram of Seq2Seq model

1.2.5　Attention模型

Attention模型是对人类大脑中的注意力进行模拟,旨在从众多信息中选择出对当前任务更关键的信息。在Encoder-Decoder框架中,Encoder中的每个单词对输出文本中的每一个单词的影响是相同的,导致语义向量无法完全表示整个序列的信息,随着输入的序列长度的增加,解码后的生成文本的质量准确度下降。Attention模型在处理输入信息时,对不同的块或区域采用不同的权值,权重越大越聚焦于其对应的内容信息,Attention模型示意图如图2所示,引入该模型后,能够使得关键信息对模型的处理结果影响较大,从而提高输出的质量。

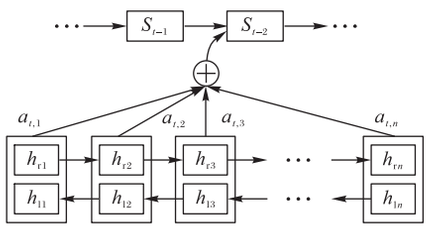

图2　注意力模型示意图

Fig. 2　Schematic diagram of attention model

1.2.6　Transformer模型

Transformer模型在2017年由Google团队[11]首次提出。Transformer是一种基于注意力机制来加速深度学习算法的模型,由一组编码器和一组解码器组成,编码器负责处理任意长度的输入并生成其表达,解码器负责把新表达转换为目的词。Transformer模型利用注意力机制获取所有其他单词之间的关系,生成每个单词的新表示。

Transformer的优点是注意力机制能够在不考虑单词位置的情况下,直接捕捉句子中所有单词之间的关系。模型抛弃之前传统的Encoder-Decoder模型必须结合RNN或者卷积神经网络(Convolutional Neural Network, CNN)的固有模式,使用全Attention的结构代替了LSTM,减少计算量和提高并行效率的同时不损害最终的实验结果;但是此模型也存在缺陷,首先此模型计算量太大,其次还存在位置信息利用不明显的问题,无法捕获长距离的信息。

1.2.7　ELMo模型

2018年,ELMo(Embedding from Language Model)出世。在之前工作中,每个词对应一个vector,处理多义词时会产生偏差。ELMo不同于以往的一个词对应一个固定向量,而是实现了将一句话或一段话输入模型,模型根据上下文来推断每个词对应的词向量。该模型的结构如图3所示,其优点是利用多层LSTM和前后向LSTM,实现结合前后语境对多义词准确理解。

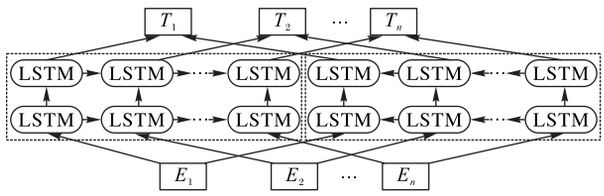

图3　ELMo模型示意图

Fig. 3　Schematic diagram of the ELMo model

1.2.8　BERT模型

BERT (Bidirectional Encoder Representations from Transformers)模型于2018年由Google团队首次提出。在自然语言生成任务中,BERT模型采用双向Transformer,模型的表示在所有层中,共同依赖于左右两侧的上下文。在自然语言生成中,该模通过查询字向量表将文本中的每个字转换为一维向量作为模型输入;模型输出则是输入各字对应的融合全文语义信息后的向量表示。与最近的其他语言表示模型不同,BERT旨在通过联合调节所有层中的上下文来预先训练深度双向表示。此模型在多种NLP任务中取得了先进结果。

1.3　技术对比

在NLG领域中,每种生成方法和模型各具特点。现按时间顺序整理常见的方法模型及其优缺点对比,如表1所示。

表1　常见方法优缺点

Tab. 1　Advantages and disadvantages of common methods

方法	优缺点
模板生成方法	优点:思路简单,用途广泛 缺点:结果质量不高,不易维护
模式生成方法	优点:生成质量较高,较易维护 缺点:用于固定结构,多样性低
RST方法	优点:生成结果的表达灵活性强 缺点:难以建立文本的结构关系
属性生成方法	优点:实现对文本内容的扩展 缺点:细粒度语言,维护困难
马尔可夫链	优点:根据当前状态预测下一步 缺点:忽视其他结构,预测不准
RNN	优点:捕捉输入数据的序列特征 缺点:短期记忆,不能并行计算
LSTM	优点:解决梯度消失 缺点:处理难以并行化
Seq2Seq	优点:处理输入输出序列不等长 缺点:曝光偏差;训练和测试度量不一致
Attention	优点:获取全局与局部关系,更关注关键信息 缺点:无法学习序列中的顺序关系
Transformer	优点:捕捉单词之间关系 缺点:计算量大,位置信息难利用
ELMo	优点:考虑上下文,准确表示多义词的词向量 缺点:利用LSTM提取特征能力弱于Transformer模型
BERT	优点:比RNN更高效地捕捉长距离的依赖 缺点:收敛比left-to-right模型慢

2　文本到文本生成

文本到文本生成技术主要是指以文本作为输入,进行变换处理后,生成新的文本作为输出。此技术包括机器翻译、文本摘要、文本更正和文本复述等。

2.1　机器翻译

在文本到文本生成领域中,机器翻译是使用机器自动地

将一种自然语言文本(源语言)翻译成另一种自然语言文本(目标语言)[12]。在统计机器翻译时期,Brown 等[13]提出基于信源信道思想的统计机器翻译模型,其基本思想是把机器翻译看成是一个信息传输的过程,用一种信源信道模型对机器翻译进行解释。2005 年,Bannard 等[14]使用双语并行语料库来提取和生成释义,基于双语平行语料提出了一种复述模型。该模型设置了一组手动词,利用短语 e_1 和短语 e_2 共有的外文翻译 f 作为“枢轴”,$P(f|e_1)$ 表示 f 是 e_1 的复述的概率,计算 $P(f|e_1)$ 和 $P(f|e_2)$ 的乘积来计算短语 e_1 是 e_2 的复述的概率 $P(e_1|e_2)$,优点是对翻译内容进行细化,并且将上下文信息考虑在内。统计机器翻译的优点是解决了规则法中翻译知识获取的难题,开发周期短,实用性较强。缺陷是模型没有考虑句子的结构信息,模型在句法结构相差加大的语言对中翻译效果不理想。

目前,神经网络机器翻译已经逐渐成为主流方法。相比传统的统计机器翻译而言,使用深度学习神经网络来实现自然语言生成中的机器翻译,不仅适合处理变长的线性序列,而且会根据上下文选择合适的单词。Kalchbrenner 等[15]于 2013 年提出一种用于机器翻译的新型编码器–解码器结构。该模型使用的数据集来自 WMT(Workshop on Machine Translation)公布的新闻部分的 144 953 对长度小于 80 个单词的双语语料库。使用卷积神经网络将给定的一段源文本编码成一个连续的向量,然后再使用循环神经网络作为解码器将该状态向量转换成目标语言。实验结果表明,该模型翻译结果的困惑度比基于对齐的模型低 43%。

如今机器翻译在应用中面临的问题主要是语言数据资源稀缺、缺少平行数据,未来的核心工作是构建高质量的平行数据库,使翻译结果更具有灵活性且贴合语境。

2.2 文本摘要

文本摘要通过分析输入的文本,捕捉原始文本的核心含义,摘取文本中的重要信息,通过提炼压缩等操作,生成篇幅短小的摘要,为用户提供阅读便利。根据实现技术方案的不同,文本摘要可以分为生成式文本摘要和抽取式文本摘要。

生成式文本摘要是一个端到端的过程,首先利用自然语言理解对文本进行语法语义分析,进行信息融合后,再利用自然语言生成技术生成文本摘要。生成式摘要包含新的词语或短语,灵活性较高。随着近几年神经网络模型的发展,带有注意力的序列到序列模型被广泛地用于生成式摘要任务[16]。其优点在于突破了传统模型中固定大小的输入问题,并能从序列中间抓住重点,不丢失重要的信息,从而解决了长距离的信息会被弱化的问题。

抽取式文本摘要是从文档或文档集中抽取其中一句话或几句话,构成摘要。优点是简单实用,不易产生完全偏离文章主旨的点,但是可能伴随生成摘要不连贯、字数不好控制、目标句主旨不明确等缺点,其产生的摘要质量好坏决定于原文。在抽取式方法中,最简单的是抽取文章中的前几句作为文本摘要。常用的方法为 Lead-3,即抽取文章的前三句作为文章的摘要。此方法简单直接,但只适用于单文档摘要。利用 Text Rank 进行文本摘要生成时,将句子作为节点,使用句子间相似度,构造无向有权边。使用边上的权值迭代更新节点值,最后选取 N 个得分最高的节点,作为文本摘要。使用聚类方法实现文本摘要生成时,首先将句子向量化表示,然后利用 K 均值聚类和 Mean-Shift 聚类方法进行句子聚类,接着从得到 K 个类别中,选择距离质心最近的句子,最后得到 K 个句子,作为最终摘要[17]。例如 Jadhav 等[18]直接使用 Seq2Seq 模型来交替生成词语和句子的索引序列来完成抽取式摘要任务,其模型 SWAP-NET(Sentencesand Words from Alternating Pointer NETwork)计算一个 Switch 概率指示生成词语或者句子,最后解码出词语和句子的混合序列,摘要从产生句子的集合选出。

抽取式、生成式摘要各有优点,混合式文本摘要为了结合两者优点,同时运用抽取方式和生成方式进行文本摘要生成。在生成式摘要中,生成过程缺少关键信息的控制和指导,无法很好地定位关键词语,因此一些方法首先提取关键内容,再进行摘要生成。Laha 等[19]将抽取式模型的输出概率作为句子级别的 attention 权重,用该权重来调整生成式模型中的词语级别的 attention 权重,当词语级别的 attention 权重高时,句子级别的 attention 权重也高。此方法使得模型输出的句子级别的权重和词语级别的权重尽量一致,有效定位关键信息。

文本摘要作为传统的自然语言处理任务,核心问题是如何确定关键信息。研究人员发现利用外部知识、关键词信息等方式来更好地辅助摘要的生成,同时要尽量避免出现重复、可读性差这些问题的出现。

2.3 文本复述

文本复述生成技术通过对给定文本进行改写,生成全新的复述文本,要求输出与原文形式差异、语义相同的文本。文本简化是文本复述的一类特殊问题,其目的是将复杂的长句改写成简单、可读性更好、易于理解的多个短句,方便用户快速阅读。在文本简化领域的研究中,Siddharthan[20]于 2014 年发表一篇综述论文,文中使用联想词汇衔接的应用来分析文本的复杂性。在 Coster 等[21]提出的关于句子简化的研究中,将英语维基百科与简单的英语维基百科生成一个平行的简化语料库,使用 Moses 提供初步的文本简化结果,发现在未简化的基础上有 0.005 个 BLEU(BiLingual Evaluation Understudy)改善。

3 数据到文本生成

数据到文本生成也是 NLG 的重要研究方向,以包含键值对的数据作为输入,旨在自动生成流畅的、贴近事实的文本以描述输入数据。数据到文本生成广泛应用于包括基于面向任务的对话系统中的对话动作、体育比赛报告和天气预报等。基于流水线模型的数据到文本生成系统框架,目前广泛应用于面向多个领域的数据到文本的生成系统[22]中。

国内关于数据到文本的生成的研究大多是基于模板,通过人工添加数据进行生成。随着神经网络的发展,数据到文本生成领域中基于神经网络序列生成的方法逐步成为热点。

3.1 基于规则和模板方法

基于规则和模板方法是一种简单实用的自然语言文本生成技术方法[23],其本质是系统根据可能出现的几种语言情况,事先设计并构造相应的规则或模板,其中都包括一些不变的常量和可变的变量,用户输入信息之后,文本生成器将输入的信息作为字符串嵌入到模板中替代变量。

2003 年 Duboue 等[24]提出一种内容选择方法,从文本语料中自动学习内容选择规则和获取相关语义,并用于人物传记的短文本生成。2017 年 Gong 等[25]在基于模板技术的自动生成系统中加以改进,提出了一种基于知识规则的模板自动生成方法,用于从模板集中动态选择模板,实现快速有效地生成海量体育新闻。基于模板集的系统生成的文本灵活性强、内容更加丰富。

模板生成技术是一种简单实用的自然语言文本生成技术[26],该技术通过将词汇和短句在模板库中进行匹配,匹配后将词汇和短语填入固定模板,从而生成自然语言文本,其本质是系统根据可能出现的几种语言情况,事先设计并构造相应的模板,每个模板都包括一些不变的常量和可变的变量,用户输入信息之后,文本生成器将输入的信息作为字符串嵌入到模板中替代变量。

基于规则和模板方法是工业应用中主流的做法,此方法具备可解释性与可控制性,保证所输出文本的正确性;然而方法的劣势较为明显,难以实现端到端的优化,损失信息上限也不高,需要依赖人工干预来抽取优质模板,生成的内容在多样性、流畅度以及连贯性往往会不尽如人意。

3.2　基于神经网络序列生成方法

近年来,随着深度学习技术的推进,研究人员开始使用神经网络序列生成的方法进行数据到文本生成,这种方法称为 data-to-seq 模型。基于神经网络的方法又分为基于神经网络语言模型的方法和基于神经机器翻译的方法。

3.2.1　基于神经机器翻译方法

Mei 等[27]的研究中将数据的文本生成任务视为一个翻译任务,即输入的是结构化数据,输出的是文本。在 Puduppully 等[28]的研究中,为了解决神经系统难以捕获长期结构的问题,提出了一个神经网络架构模型,如图 4 所示。

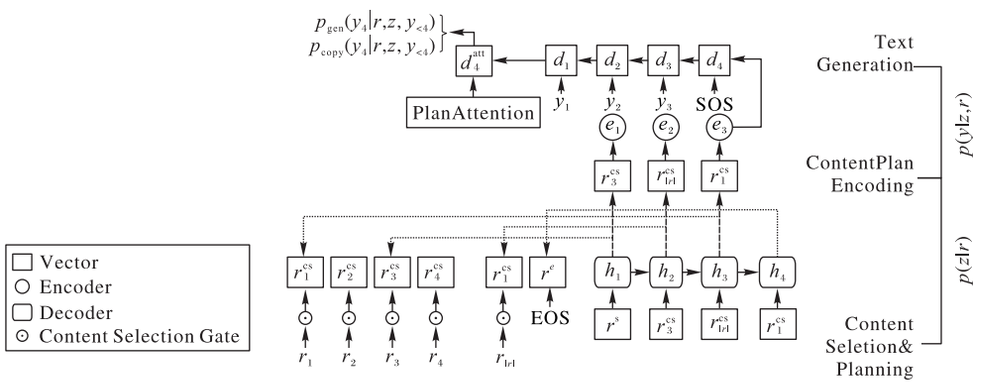

图 4　具有内容选择和规划的生成模型

Fig. 4　Generation model with content selection and planning

模型将任务分解为两个阶段:1)内容选择和规划对数据库的输入记录进行操作,并生成一个内容计划,指定哪些记录将在文档中以及以何种顺序进行语言描述;2)文本生成产生输出文本给定内容计划作为输入;同时加入 copy 机制来提升解码器的效果。实验结果表明,在输出文本中包含的相关事实数量和这些事实呈现出的顺序性,生成质量都得到了提高。

3.2.2　基于神经网络语言模型方法

神经网络语言模型由 Bengio 等[29]于 2003 年提出,模型解决了 n-gram 模型当 n 较大时会发生数据稀疏的问题。基于神经语言模型的方法不需要太多人工干预,易产生丰富流畅的文字描述,不过受限于语料和模型,使用者无法直接控制内容生成,难以确保所输出的文本内容同输入数据中的信息吻合,需要优化来提高实用性。2018 年,在 Yang 等[30]关于 TEG (Topic to Essay Generation)任务的研究中,将知识图谱嵌入当作外部知识辅助自然语言生成。过去 TEG 工作仅仅基于给定的主题去执行文本生成,忽略常识知识所提供的背景知识,常识知识能够有效提高生成文章的新颖性和多样性。Yang 等[30]的实验结果与 BLEU 评分的最佳基线相比,取得了 11.85% 的相对改进,所以通过知识图谱嵌入来辅助自然语言生成,生成的文章新颖多样且主题一致。

3.3　公开数据集

在不同的应用领域,有相关特定的数据到文本生成的数据集,如表 2 所示。在天气预报生成领域中的数据集有 SumTime-Meteo[31]和 Weather Gov[32];体育比赛领域的数据集有 RoboCup[33]、NFL(National Football League)[34]、Rotowire[35];航空领域常见的数据集有 ATIS (Automatic Terminal Information System)[36];人物传记领域常见的数据集有 WikiBio[37]。

表 2　数据到文本生成常见的数据集

Tab. 2　Data-to-text generated common datasets

数据集名	领域	下载地址
SumTime-Meteo	天气预报	http://www.itri.brighton.ac.uk/home/Anja.Belz/Prodigy
Weather Gov	天气预报	http://cs.stanford.edu/~pliang/papers/
RoboCup	体育比赛	http://www.cs.utexas.edu/ml/clamp/sportscasting/
NFL	体育比赛	http://pages.cs.wisc.edu/~bsnyder/
Rotowire	体育比赛	https://github.com/harvardnlp/boxscore-data
ATIS	航天航空	http://www.Ikonstas.net/index.php?page=resources
WikiBio	人物传记	https://github.com/DavidGrangier/wikipedia-biography-dataset

4 图像到文本生成

图像到文本生成是指根据输入的图像信息生成描述图像的自然语言文本，常应用于给新闻图片生成标题、儿童教育中看图讲故事、医学图像报告等。此项技术能够为缺乏相关知识或阅读障碍的人群提供便利。

根据生成文本的长度和内容详细程度分类，可以将图像到文本生成分为图像标题自动生成和图像说明自动生成。图像的文本生成技术主要可分为三类：基于模板的图像描述、检索式图像描述以及生成式图像描述。

4.1 基于模板的图像描述

早期利用流水线模式实现图像到文本生成。在Yao等[38]的研究中，使用句子模板实现生成图像的描述，其模板为四元组形式。此模型在视频监控系统和自动驾驶场景理解系统中解析特定域中的图像视频进行实验，生成有使用价值的文本报告。

基于模板的图像描述方法的优点是能够有效保证生成文本语法的正确性以及内容的相关性。该方法由于视觉模型数量较少，所以存在所生成的句子新颖度和复杂度不高等问题。

4.2 检索式图像描述

检索式图像描述是根据待描述图像，从句子池中检索出一个或一组句子来为图像生成描述[39]。Farhadi等[40]通过建立的三元组〈对象，动作，场景〉实现图像与文本意义的相关联。根据给定的待描述图像，首先利用求解Markov Random Field将其映射到三元组，然后通过Lin相似度来计算图像和句子之间的语义距离，最后选择从句子池中选择语义最相近的句子来实现图像描述的生成。

基于检索式图像到文本的生成方法能够使得生成文本在语法上具有正确性和流畅性[41]，但由于使用句子池中的句子进行图像描述，生成效果欠缺新颖性，在描述复杂场景或包含新颖事物的图片中存在局限性。

4.3 生成式图像描述

从视觉空间或多模态空间中生成图像描述的做法是，在分析图像内容的基础上，使用语言模型来生成图像的描述。因为此方法利用了深度学习技术，所以可以适应为多种的图像生成新的描述的任务需求，生成文本的相关性和准确性较之前方法有所提升。因此，基于深度学习的生成式图像描述是目前研究的热点。

4.3.1 基于多模态空间的图文生成

多模态空间的图文生成框架包含4个部分，框架如图5所示。

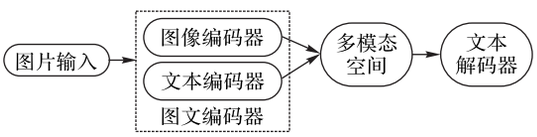

图5 基于多模态空间的图文生成框架

Fig. 5 Framework of image text generation based on multimodal space

图像编码器在深度卷积神经网络的作用下实现图像特征的提取。文本编码器在提取单词特征的基础上学习并更新单词的特征表示，并将其按照上下文顺序馈送给循环神经网络。多模态空间的作用是将图像特征和文本特征映射到空间。然后传至文本解码器，从而生成图像描述。Li等[42]利用知识图谱技术实现医学报告生成，其流程是根据输入的医学图像，先用预训练好的CNN提取出图像特征；然后经过一个图像编码器得到语境向量；接着用句子解码器对语境向量进行解码得到若干个topic；对于每个topic可以用模板库或者生成模式进行强化学习，得到诊断报告。

4.3.2 基于生成对抗网络的图文生成

生成对抗网络（Generative Adversarial Network，GAN）由1个生成网络和1个判别网络组成，在两个神经网络相互博弈中进行学习。输入随机噪声后，生成网络会模拟真实样本进行输出。生成网络的输出作为判别网络的输入，目的是分辨数据来自真实样本还是来自网络生成。在生成网络和判别网络相互对抗中，通过学习调整参数，直到生成结果和真实样本趋于一致。

基于GAN的图像描述方法与传统神经网络模型相比，生成的文本更加贴近人类的描述，更具有多样性。

4.3.3 基于强化学习的图文生成

强化学习是通过Agent与Environment交互的方式来获得奖励，以此来指导Agent的下一步行为。Ren等[43]提出的基于强化学习的图文生成体系结构由“策略网络”和“价值网络”构成。在每个时间步内，两个网络共同计算下一个最佳生成词，该方法借助实际奖励值来衡量图像与句子相似性，并以此评估生成的图像描述文本的正确性。Rennie等[44]提出了一种基于self-critical思想的强化学习方法来训练序列生成模型。此方法没有直接去估算奖励，而是使用测试阶段的输出来归一化奖励而不是评估一个baseline归一化奖励。

基于强化学习的图文生成方法可以优化序列学习中的曝光偏差问题，但也可能存在具有很高方差的问题。

5 评估方法

5.1 BLEU

BLEU是一个双语评估辅助工具，主要用来评估机器翻译的质量。n-gram在自然语言处理中表示多元精度，可以用来评估一个句子是否合理，也可以用来评估两个字符串之间的差异程度。BLEU的核心思想是比较候选文本和参考文本里的n-gram的重合程度，重合程度越高就认为译文质量越高。uni-gram用于衡量单词翻译的准确性，高阶n-gram用于衡量句子翻译的流畅性[45]。实践中，通常是取n=1~4，然后对进行加权平均。它的计算公式如下：

$$BLEU = BP\cdot\exp\left(\sum_{n=1}^{N} W_n \log P_n\right) \tag{1}$$

其中：n表示n-gram，BP为惩罚因子，P_n为多元精度，W_n为多元精度对应的权重。惩罚因子BP具体计算方法为：

$$BP = \begin{cases} 1, & c > r \\ e^{1-r/c}, & c \leqslant r \end{cases} \tag{2}$$

其中：c指候选译文的长度；r指所有参考译文中，其长度与候选译文最接近的长度。惩罚因子主要用来惩罚机器译文与参考译文长度差距过大情况。

5.2 METEOR

METEOR又称显式排序的翻译评估指标[46]，它在基于BLEU的基础上进行了一些改进，其目的是克服一些BLEU标准中的缺陷。使用WordNet计算特定的序列匹配，同义词，词根和词缀、释义之间的匹配关系，改善了BLEU的效果，使其跟人工判别有更强的相关性。计算公式如下：

$$METEOR = (1 - pen) \times F_{means} \tag{3}$$

$$F_{means} = \frac{PR}{\alpha P + (1 - \alpha)R} \tag{4}$$

$$P = m/c \tag{5}$$

$$R = m/r \tag{6}$$

其中:α为可调控的参数,m为候选翻译中能够被匹配的一元组的数量,c为候选翻译的长度,r为参考摘要的长度。pen为惩罚因子,惩罚的是候选翻译中的词序与参考翻译中的词序不同,具体计算方法为:

$$pen = \frac{\#chunks}{m} \tag{7}$$

其中:m是候选翻译中能够被匹配的一元组的数量,$\#chunks$指的是chunk的数量,chunk是既在候选翻译中相邻又在参考翻译中相邻的被匹配的一元组聚集而成的单位。

METEOR主要特点是uni-gram共现统计、基于F值和考虑同义词、词干,常应用于机器翻译和图片说明,因为其依赖于Java才能实现,并且参数较多,需要外部知识源如WebNet的支持,所以它在应用起来有一定的局限性。

5.3　ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)大致分为4种:ROUGE-N、ROUGE-L、ROUGE-W、ROUGE-S。常用的是前两种,ROUGE-N中的"N"指的是n-gram,其计算方式与BLEU类似,只是BLEU基于精确率,而ROUGE基于召回率。ROUGE-L中的"L"指的是Longest common sub sequence,计算的是候选摘要与参考摘要的最长公共子序列长度,长度越长得分越高。

主要介绍ROUGE-N和ROUGE-L的计算公式,ROUGE-N计算公式如下:

$$ROUGE\text{-}N = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \tag{8}$$

其中:n表示n-gram,$Count(gram_n)$表示一个n-gram的出现次数,$Count_{match}(gram_n)$表示一个n-gram的共现次数。

ROUGE-L的计算公式如下:

$$ROUGH\text{-}L = \frac{(1 + \beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{9}$$

$$R_{lcs} = LCS(X, Y)/m \tag{10}$$

$$P_{lcs} = LCS(X, Y)/n \tag{11}$$

其中:X表示候选摘要,Y表示参考摘要,$LCS(X,Y)$表示候选摘要与参考摘要的最长公共子序列的长度,m表示参考摘要的长度,n表示候选摘要的长度。

ROUGE方法的特点是n-gram共现统计、基于召回率(ROUGE-N)和F值(ROUGE-L),常应用于文本摘要。值得注意的是,ROUGE是基于字的对应而非基于语义的对应,不过可以通过增加参考摘要的数量来缓解这一问题。

5.4　Perplexity

Perplexity又称困惑度[47]。它的核心思想是:首先根据参考句子,学习一个语言模型P;然后根据语言模型P,计算候选句子的得分;最后根据句子长度对上述得分进行标准化。计算公式如下:

$$PPL(W) = P(w_1 w_2 \cdots w_N)^{1/N} \tag{12}$$

其中:W是候选翻译,N是候选翻译的长度,P是根据参考翻译得到的语言模型,而$P(w_1 w_2 \cdots w_N)$则是语言模型对候选翻译计算出的得分。

Perplexity这一评估指标是基于语言模型的。困惑度越低、翻译质量越好,经常应用于机器翻译、语言模型。它的缺点是:数据集越大困惑度下降得越快、数据中的标点会对模型的PPL产生影响和常用词干扰。

5.5　CIDEr

CIDEr(Consensus-based Image Description Evaluation)是基于共识的图像描述进行评估,核心思想:把每个句子看成文档,然后计算其TF-IDF(Term Frequency-Inverse Document Frequency)向量的余弦夹角,据此得到候选句子和参考句子的相似度。计算公式如下:

$$CIDEr_n(c, S) = \frac{1}{M} \sum_{i=1}^{M} \frac{\boldsymbol{g}^n(c) \cdot \boldsymbol{g}^n(S_i)}{\left\| \boldsymbol{g}^n(c) \right\| \times \left\| \boldsymbol{g}^n(S_i) \right\|} \tag{13}$$

其中:c表示候选标题,S表示参考标题集合,n表示评估的是n-gram,M表示参考标题的数量,$\boldsymbol{g}^n$表示基于n-gram的TF-IDF向量。

该评估方法主要运用于图片说明,它与ROUGE一样,也只是基于字词的对应而非语义的对应。

5.6　语义命题图像标题评估

语义命题图像标题评估(Semantic Propositional Image Caption Evaluation, SPICE)的核心思想是使用基于图的语义表示来编码文字中的物体、属性和关系。它先将候选文本和参考文本用概率上下文无关法解析成句法依赖关系树,然后用规则法把依存关系树映射成场景图[48],最后计算候选文本中物体、属性和关系中的F-score值。它的计算公式如下:

$$SPICE(c, S) = F_1(c, S) = \frac{2 \cdot P(c, S) \cdot R(c, S)}{P(c, S) + R(c, S)} \tag{14}$$

$$P(c, S) = \frac{\left| T(G(c)) \otimes T(G(S)) \right|}{T(G(c))} \tag{15}$$

$$R(c, S) = \frac{\left| T(G(c)) \otimes T(G(S)) \right|}{T(G(S))} \tag{16}$$

其中:c表示候选文本,S表示参考文本集合,$G(\cdot)$函数表示将一段文本转换成一个场景图,$T(\cdot)$函数表示将一个场景图转换成一系列元组的集合;$\otimes$运算类似于交集,与交集不同的地方在于它不是严格匹配,而是类似于METEOR中的匹配。

SPICE方法的主要特点是使用基于图的语义表示,常应用于图片说明。在评估的时候主要考察名词的相似度,不适合用于机器翻译等任务。

6　发展趋势

借助自然语言生成的演变可以看到,从使用简单的马尔可夫链生成句子到使用注意力机制模型生成更长距离的连贯文本,如今正处于自然语言生成建模的攻坚克难时期。Transformer向真正自主文本生成方向迈出了重要的一步,与此同时,还针对其他类型的内容(例如图像、视频和音频)开发了生成模型。目前在自然语言生成评估标准中,缺乏一个通用的高质量的评估标准,这也是制约NLG发展的一个重要原因,接下来研究的一个热点是整理出一个更好的业内公认的高质量的数据集来制定高质量的评价标准。

参考文献(References)

[1]　REITER E, DALE R. Building natural language generation systems

[J]. Computational Lingus, 2001, 27(2):298-300.

[2] DAY C. Robot science writers[J]. Computing in Science and Engineering, 2018, 20(3): 101-101.

[3] JING B, XIE P, XING E. On the automatic generation of medical imaging reports[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2018: 2577-2586.

[4] GOLDBERG E, DRIEDGER N, KITTREDGE R I. Using natural-language processing to produce weather forecasts[J]. IEEE Expert, 1994, 9(2): 45-53.

[5] ZHANG B, XIONG D, SU J. Neural machine translation with deep attention[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(1): 154-163.

[6] LI X, CHEN Y N, LI L, et al. End-to-end task-completion neural dialogue systems[C/OL]// Proceedings of the 8th International Joint Conference on Natural Language Processing. [2020-05-03]. https://www. aclweb. org/anthology/I17-1074. pdf.

[7] VAN DEEMTER K, THEUNE M, KRAHMER E. Real versus template-based natural language generation: a false opposition[J]. Computational Linguistics, 2005, 31(1):15-24.

[8] CULLEN C, O'NEILL I, HANNA P. Flexible natural language generation in multiple contexts [C]// Proceedings of the 3rd Language and Technology Conference, LNCS 5603. Berlin: Springer, 2009: 142-153.

[9] ZHANG Y, YAO Q, DAI W, et al. AutoSF: searching scoring functions for knowledge graph embedding[C]// Proceedings of the IEEE 36th International Conference on Data Engineering. Piscataway: IEEE, 2020:433-444.

[10] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks [C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 3104-3112.

[11] REITER E. A structured review of the validity of BLEU [J]. Computational Linguistics, 2018, 44(3): 393-401.

[12] 李强,黄辉,周沁,等. 模板驱动的神经机器翻译[J]. 计算机学报, 2019, 42(3): 566-581. (LI Q, WONG F, CHAO S, et al. Template-driven neural machine translation[J]. Chinese Journal of Computers, 2019, 42(3): 566-581.)

[13] BROWN P F, PIETRA S A D, PIETRA V J D, et al. The mathematics of statistical machine translation: parameter estimation [J]. Computational Linguistics, 1993, 19 (2): 263-311.

[14] BANNARD C, CALLISON-BURCH C. Paraphrasing with bilingual parallel corpora[C]// Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2005: 597-604.

[15] KALCHBRENNER N, BLUNSOM P. Recurrent continuous translation models[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2013: 1700-1709.

[16] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. [2020-06-03]. https://arxiv. org/pdf/1409. 0473. pdf.

[17] NARAYAN S, COHEN S B, LAPATA M. Ranking sentences for extractive summarization with reinforcement learning [C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2018:1747-1759.

[18] JADHAV A, RAJAN V. Extractive summarization with SWAP-NET: sentences and words from alternating pointer networks[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2018: 142-151.

[19] LAHA A, JAIN P, MISHRA A, et al. Scalable micro-planned generation of discourse from structured data [J]. Computational Linguistics, 2019, 45(4): 737-763.

[20] SIDDHARTHAN A. A survey of research on text simplification [J]. International Journal of Applied Linguistics, 2014, 165(2): 259-298.

[21] COSTER W, KAUCHAK D. Simple English Wikipedia: a new text simplification task [C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2011: 665-669.

[22] 徐戈,王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报, 2011, 34(8): 1423-1436. (XU G, WANG H F. The development of topic models in natural language processing[J]. Chinese Journal of Computers, 2011, 34(8):1423-1436.)

[23] ORABY S, HARRISON V, EBRAHIMI A, et al. Curate and generate: a corpus and method for joint control of semantics and style in neural NLG[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2019: 5938-5951.

[24] DUBOUE P A, MCKEOWN K R. Statistical acquisition of content selection rules for natural language generation[C]// Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2003: 121-128.

[25] GONG J, REN W, ZHANG P. An automatic generation method of sports news based on knowledge rules[C]// Proceedings of the IEEE/ACIS 16th International Conference on Computer and Information Science. Piscataway: IEEE, 2017: 499-502.

[26] DUŠEK O, NOVIKOVA J, RIESER V. Evaluating the state-of-the-art of end-to-end natural language generation: the E2E NLG challenge [J]. Computer Speech and Language, 2020, 59: 123-156.

[27] 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 North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2016: 720-730.

[28] PUDUPPULLY R, DONG L, LAPATA M. Data-to-text generation with content selection and planning[C]// Proceedings of the 33rd National Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2019: 6908-6915.

[29] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.

[30] YANG P, LI L, LUO F, et al. Enhancing topic-to-essay generation with external commonsense knowledge [C]// Proceedings of the 57th Annual Meeting of the Association for

Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2019: 2002-2012.

[31] REITER E, SRIPADA S, HUNTER J, et al. Choosing words in computer-generated weather forecasts[J]. Artificial Intelligence, 2005, 167(1/2):137-169.

[32] LIANG P, JORDAN M I, KLEIN D. Learning semantic correspondences with less supervision[C]// Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL/ the 4th International Joint Conference on Natural Language Processing of the AFNLP. Stroudsburg, PA: Association for Computational Linguistics, 2009: 91-99.

[33] CHEN D L, MOONEY R J. Learning to sportscast: a test of grounded language acquisition[C]// Proceedings of the 25th International Conference on Machine Learning. New York: ACM, 2008: 128-135.

[34] BARZILAY R, LAPATA M. Collective content selection for concept-to-text generation[C]// Proceedings of the 2005 Human Language Technology Conference/ Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2005: 331-338.

[35] WISEMAN S, SHIEBER S M, RUSH A M. Challenges in data-to-document generation[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2017: 2253-2263.

[36] DAHL D A, BATES M, BROWN M, et al. Expanding the scope of the ATIS task: the ATIS-3 corpus[C]// Proceedings of the 1994 Workshop on Human Language Technology. Stroudsburg, PA: Association for Computational Linguistics, 1994: 43-48.

[37] LEBRET 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. Stroudsburg, PA: Association for Computational Linguistics, 2016: 1741-1752.

[38] YAO B Z, YANG X, LIN L, et al. I2T: image parsing to text description[J]. Proceedings of the IEEE, 2010, 98(8): 1485-1508.

[39] 莫凌波. 基于图像的文本自动生成关键技术研究[D]. 北京: 北京邮电大学, 2019: 6-8. (MO L B. Research on key technologies of automatic text generation based on images[D]. Beijing: Beijing University of Posts and Telecommunications, 2019: 6-8.)

[40] FARHADI A, ENDRES I, HOIEM D, et al. Describing objects by their attributes[C]// Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2009: 1778-1785.

[41] 孔锐,谢玮,雷泰. 基于神经网络的图像描述方法研究[J]. 系统仿真学报, 2020, 32(4): 601-611. (KONG R, XIE W, LEI T. Research on image description method based on neural network[J]. Journal of System Simulation, 2020, 32(4): 601-611.)

[42] LI C Y, LIANG X, HU Z, et al. Hybrid retrieval-generation reinforced agent for medical image report generation[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2018: 1537-1547.

[43] REN Z, WANG X Y, ZHANG N, et al. Deep reinforcement learning-based imagecaptioning with embedding reward[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 1151-1159.

[44] RENNIE S J, MARCHERET E, MROUEH Y, et al. Self-critical sequence training for image captioning[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 7008-7024.

[45] LOPEZ-GAZPIO I, MARITXALAR M, LAPATA M, et al. Word *n*-gram attention models for sentence similarity and inference[J]. Expert Systems with Applications, 2019, 132: 1-11.

[46] DENKOWSKI M, LAVIE A. Meteor universal: language specific translation evaluation for any target language[C]// Proceedings of the 9th Workshop on Statistical Machine Translation. Stroudsburg, PA: Association for Computational Linguistics, 2014: 376-380.

[47] LIN C Y, HOVY E. Automatic evaluation of summaries using N-gram co-occurrence statistics[C/OL]// Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. [2020-05-03]. https://www. aclweb. org/anthology/N03-1020. pdf.

[48] NOVIKOVA J, DUSEK O, CURRY A C, et al. Why we need new evaluation metrics for NLG[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2017: 2241-2252.

This work is partially supported by the National Natural Science Foundation of China (61872313), the Key Project of National Key Research and Development Program of China (2017YFB1002300, 2018YFC1700302).

LI Xueqing, born in 1995, Ph. D. candidate. Her research interests include natural language processing .

WANG Shi, born in 1981, Ph. D., associate research fellow. His research interests include semantic analysis, knowledge graph.

WANG Zhujun, born in 1996, M. S. candidate. His research interests include natural language processing.

ZHU Junwu, born in 1972, Ph. D., professor. His research interests include knowledge engineering, ontology.