计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(5):246−253 [doi: 10.15888/j.cnki.csa.008480]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

结构化数据到数值型分析文本生成模型①

杨子聪[1,2], 焦文彬[1], 刘晓东[1], 汪　洋[1]

[1](中国科学院 计算机网络信息中心, 北京 100190)
[2](中国科学院大学, 北京 100049)
通信作者: 焦文彬, E-mail: wbjiao@cnic.cn

摘　要: 基于结构化数据的文本生成是自然语言生成领域重要的研究方向, 其可以将传感器采集或计算机统计分析得到的结构化数据转化为适宜人阅读理解的自然语言文本, 因此也成为了实现报告自动生成的重要技术. 研究基于结构化数据到文本生成的模型为报告中的各类数值型数据生成分析性文本具有重要的实际应用价值. 本文针对数值型数据的特点, 提出了一种融合 coarse-to-fine aligner 选择机制和 linked-based attention 注意力机制的编码器-解码器文本生成模型, 考虑了生成数值型数据的分析性文本过程中内容过度分散、无法突出描述的问题, 另外也将数值型数据具体所属的域进行了关系建模, 以提高生成文本中语序的正确性. 实验结果表明, 本文提出的融合两种机制的模型, 比仅使用传统的基于内容的注意力机制和在前者基础上增加使用 linked-based attention 注意力机制的模型, 以及基于 GPT2 的模型在指标上都具有更好的表现, 证明了本文提出的模型在生成数值型数据的分析性文本任务中具有一定的效果.

关键词: 结构化数据; 数值型数据; 文本生成; 报告自动生成; 深度学习

引用格式: 杨子聪,焦文彬,刘晓东,汪洋.结构化数据到数值型分析文本生成模型.计算机系统应用,2022,31(5):246–253. http://www.c-s-a.org.cn/1003-3254/8480.html

Generation Model from Structured Data to Numerical Analysis Text

YANG Zi-Cong[1,2], JIAO Wen-Bin[1], LIU Xiao-Dong[1], WANG Yang[1]

[1](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Text generation based on structured data is an important research direction in the field of natural language generation. It can transform structured data collected by sensors or statistically analyzed by computers into natural language texts suitable for human reading and understanding, thus becoming an important technology for automatic report generation. It is of great application value to study models of generating texts from structured data for the generation of analytical texts from various types of numerical data in reports. In this study, we propose an Encoder-Decoder text generation model incorporating the coarse-to-fine aligner selection mechanism and the linked-based attention mechanism, which matches the characteristics of numerical data, and consider the problems of excessive content dispersion and failure to highlight descriptions in the process of generating analytical texts from numerical data. In addition, we also model the relationship between the domains to which the numerical data specifically belong in order to improve the correctness of the discourse order in generated texts. Experimental results show that the model proposed in this study, which incorporates both mechanisms, has better performance in terms of metrics than the traditional model based on the content-based attention mechanism only, the model based on both the content-based attention mechanism and the linked-based attention mechanism, and the GPT2-based model. This proves the effectiveness of the proposed model in the task of

① 基金项目: 中国科学院信息化专项 (XXH13510-03)
收稿时间: 2021-08-07; 修改时间: 2021-09-13; 采用时间: 2021-09-18; csa 在线出版时间: 2022-04-11

generating analytical texts with numerical data.

Key words: structured data; numerical data; text generation; automatic report generation; deep learning

随着自然语言处理技术的快速发展, 越来越多报告的生成实现了自动化, 例如财报的自动生成、体育赛事报道的自动生成和医学报告的自动生成等. 由于报告具有描述总结数据的天然特征, 基于结构化数据的文本生成便成为实现报告自动生成的核心内容, 报告自动生成系统只有拥有了高性能的结构化数据到文本生成模型, 才能产出高质量的分析报告. 而报告多以数值型数据和对应的描述分析性文本组成, 因此研究基于数值型结构化数据的文本生成模型具有重要意义.

基于结构化数据的文本生成的主要技术框架有两种: 基于规则和模板化的传统方法, 和数据驱动的端到端的深度学习方法[1]. 传统方法虽易于控制和改进, 但费时费力且无法迁移[2]. 而基于深度学习的方法可控性虽表现还不如人意, 但可通过不断的训练和优化模型来逐步提升, 且迁移性强. 基于深度学习的方法主要使用 Encoder-Decoder 训练框架, 该框架为 2014 年 Cho 等人在 Seq2Seq 循环神经网络中首次提出, 最早被用来进行机器翻译模型的训练, 后广泛用于文本生成领域内的各项任务[3]. 在结构化数据到文本生成的任务中, Encoder 和 Decoder 部分使用的深度学习网络主要分为循环神经网络 (RNN) 及其变种长短时记忆网络 (LSTM) 和 Transformer 网络两大类[4], 从指标和效果上看, 仍旧是前者具有更好的表现, 且该类神经网络适合处理序列化的数据, 另外其在可控性上也有更多优化空间. 在使用 RNN 和 LSTM 实现 Encoder-Decoder 训练框架的基础上, 大多数模型往往会在 Encoder 部分使用 attention 机制[5], 通过计算隐含层状态和更新后的解码器状态的相似度, 获得每个隐含层在汇总成中间语义向量时的权重, 使得模型能够有重点地关注输入. 另外考虑到输入的结构化数据中出现的词语通常可用来生成句子, 但由于频率太低经常被忽略的问题, 因此在 Decoder 部分常常使用 copy 机制[6].

考虑到目前的结构化数据到文本生成技术主要应用于人物生平介绍、餐馆信息描述和商品介绍的生成, 而本文处理的结构化数据相较于以上应用场景中处理的结构化数据的一大特点是数值型数据较多, 甚至全部是数值型数据, 这样的情况常常导致生成的文本无法捕捉重点信息, 训练上出现困难. 基于此, 本文提出的模型融合了 coarse-to-fine aligner 选择机制[7], 在使用传统的基于内容的注意力机制计算隐含层权重的基础上, 另外赋予隐含层一个被选择的概率, 通过计算各个隐含层被选择的概率和当前时刻获得的注意力权重两者的乘积确定最终的注意力权重, 从而达到对结构化数据 [field, content] 中的 content 进行选择性描述的目的. 另外由于报告中对数据的分析性文本常常要求逻辑明确、语序正确, 因此本文的模型也融合了 linked-based attention 注意力机制, 通过对结构化数据 [field, content] 中的 field 进行关系建模, 模拟不同的 Field 之间在文本中出现的先后关系[8]. 最终本文的模型采用 LSTM 实现了 Encoder-Decoder 框架, 在 Encoder 部分使用基于内容的注意力机制, 在 Decoder 部分使用 copy 机制, 并在此基础上根据数值型数据的特点和报告中分析性文本的应用要求融合了 coarse-to-fine aligner 选择机制和 linked-based attention 注意力机制. 通过使用 A 股的市场数据和对 A 股的每日播报资讯作为模型训练和测试的数据集, 并与仅使用基于内容的注意力机制的模型和在前者基础上增加使用 linked-based attention 注意力机制的模型进行对比, 显示了本文提出的模型具有较好的效果.

1 基本训练框架和机制研究

本节简要介绍下基于结构化数据的文本生成所使用的基本训练框架, 以及解决该类任务时在该框架中常使用的两种机制.

1.1 Encoder-Decoder 框架

Encoder-Decoder 是一种训练框架, 分为编码器和解码器. 编码器的功能是将现实问题转化为数学问题, 例如将输入的文本、图片或音频表征成向量. 解码器的功能是基于编码器的结果求解数学问题, 并转化为现实世界的解决方案.

而编码器和解码器功能的实现均需要依靠深度学习网络, 具体选择则根据应用场景需要而定. 由于 LSTM 具有处理序列数据的优势, 且解决了 RNN 在面临长序列时产生的梯度消失和梯度爆炸的问题, 本文选择 LSTM

构建编码器-解码器模型.

输入的序列x_i经过 LSTM 进行编码, 得到隐含层状态h_i, t时刻隐含层的状态由当前输入x_t和$t-1$时刻的隐含层状态h_{t-1}决定, 计算过程如下:

$$h_t = f(h_{t-1}, x_t) \tag{1}$$

得到隐含层状态后, 将其汇总即可得到中间语义向量C, 计算过程如下:

$$C = LSTM(h_1, h_2, h_3, \cdots, h_n) \tag{2}$$

解码的过程则依据中间语义向量C和输出序列$Y_1, Y_2, Y_3, \cdots, Y_{t-1}$来预测下一个时刻的输出$Y_t$, 计算过程如下:

$$Y_t = \arg\max P(Y_t) = \prod_{t=1}^{T} p\left(Y_t \mid \{Y_1, \cdots, Y_{t-1}\}, C\right) \tag{3}$$

在 LSTM 中, 式(3)可表示为:

$$Y_t = g(Y_{t-1}, s_{t-1}, C) \tag{4}$$

其中, g表示非线性激活函数, 实质上相当于经过 LSTM 网络作用后再经过 Softmax 函数处理, Y_{t-1}表示$t-1$时刻的解码器输出值, s_{t-1}表示$t-1$时刻解码器的状态, C则表示中间语义向量. Encoder-Decoder 框架图见图 1.

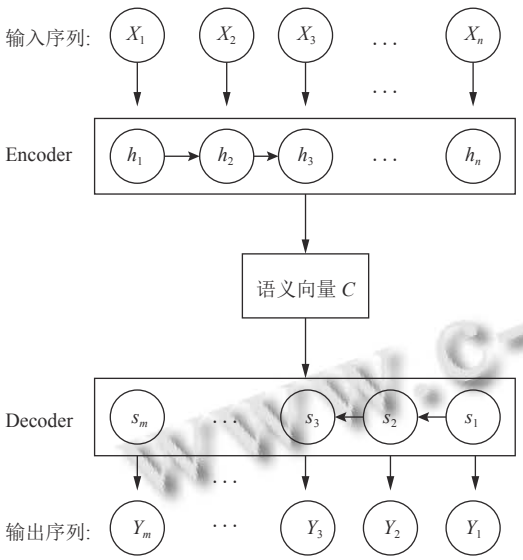

图 1 Encoder-Decoder 框架图

1.2 基于内容的注意力机制

在使用 Encoder-Decoder 框架时, 常常会在汇总隐含层形成中间语义向量时使用注意力机制, 这样的做法解决了 Encoder 部分必须将整个输入序列的信息都压入到一个固定长度的 context 中, 从而也解决了输入序列过长可能的信息缺失和输入序列过短可能的信息冗余问题, 同时还可以对输入的内容分配不同的关注度, 最终充分利用信息.

注意力机制首先由 Bahdanau 等人提出, 且在自然语言处理领域应用广泛[9]. 注意力机制有很多种, 本文提出的模型采用的是基于内容的注意力机制, 基本思路是计算各个时刻的隐含层的输出值h_i与t时刻解码器的s_t状态的相似性, 来确定隐含层各部分在汇总到中间语义向量过程中所占的权重, 具体计算方法使用的是 Luong attention[10], 其计算方式如下:

$$\alpha_{t_i} = \frac{e^{g(s_t, h_i)}}{\sum_{j=1}^{N} e^{g(s_t, h_j)}} \tag{5}$$

$$a_t = \sum_{i=1}^{L} \alpha_{t_i} h_i \tag{6}$$

其中, α_{t_i}表示t时刻隐含层状态h_i的权重, s_t表示t时刻解码器端隐含层的状态, g函数是计算s_t和h_i的相似度.

在得到各个隐含层h_i在t时刻的权重之后, 将权重分别与对应的隐含层相乘, 最终求和便可得到基于内容的注意力机制下的中间语义向量.

1.3 Copy 机制

在使用 Encoder-Decoder 框架时, 还存在另一个问题, 即无法充分利用结构化数据中的词语. 具体来说, 结构化数据中的很多词汇十分适合用于最后生成的文本当中, 但由于其出现频率较低, 常常被忽略. 因此编解码器模型中常常使用 copy 机制解决这一问题, 它使得模型结合 generate 和 copy 两种方式, 模型在解码阶段会选择是从词汇表中按照概率选择要生成的词还是直接从输入的数据中复制[11]. 其中复制的概率计算方法如下:

$$s_t^{\text{copy}}(w) = \sigma\left(h_i^{\mathrm{T}} W_c\right) h_t' \tag{7}$$

其中, h_i表示输入数据中的单词经过编码阶段得到的结果, h_t'表示解码器的状态. 另一部分则需要计算从词汇表中生成的概率, 计算方式如式(8), 其中, h_t'表示 t 时刻解码器的状态. 最后, 将两部分概率相加, 并且经过 Softmax 函数处理形成一个将输入数据中的词汇c扩充进原词汇表v的新的词汇概率分布, 具体计算方式如式(9) 和式 (10).

$$s_t^{\text{LSTM}} = W_s h_t' + b_s \tag{8}$$

$$s_t(w) = s_t^{\text{LSTM}}(w) + s_t^{\text{copy}}(w) \tag{9}$$

$$p_t(w) = Softmax(s_t(w)) = \frac{\exp\{s_t(w)\}}{\sum_{w' \in v \cup c} \exp\{s_t(w')\}} \tag{10}$$

2 结构化数据到文本生成模型研究

为了更针对性地解决生成数值型数据的分析性文本这一任务, 本文提出的模型融合了 coarse-to-fine aligner 选择机制和 linked-based attention 注意力机制. 这两种机制均作用于生成中间语义向量的过程中, 其中, coarse-to-fine aligner 选择机制在模型使用基于内容的注意力机制基础上, 增加了对结构化数据 [field, content] 中 content 部分的预选功能, 优化了生成的文本中对描述内容的选择. 而 linked-based attention 注意力机制则是对结构化数据 [field, content] 中 field 部分进行关系建模, 使得模型可以在生成文本时保持一个合理的描述顺序.

2.1 融合 coarse-to-fine aligner 选择机制

在基于结构化数据生成文本时, 无论是使用传统的基于规则的模板方法还是数据驱动的端到端的深度学习方法, 优化之处均是相同的 3 部分: ① 内容规划, 即选择结构化数据中需要描述的 field 和 content; ②句子规划, 即确定选择的描述内容在生成的文本中的描述顺序; ③ 句子实现, 即基于前两步的规划生成对应的文本. 在内容规划部分, 基于结构化数据到文本生成的模型往往仅使用基于内容的注意力机制, Mei 等提出了一种 coarse-to-fine 的选择机制[7], 在计算每部分隐含层注意力权重的基础上赋予一项选择该部分的概率, 从而实现了对内容选择的优化.

在处理数值型结构化数据时, 数据中的 field 部分和其他类型数据的处理方式没有不同, 而 content 部分大多是数值型的数据, 甚至全部是数值型数据, 在训练过程中无法使得模型对某几项数据进行重点关注和描述, 仅使用基于内容的注意力机制已经无法满足此类场景下的应用需要. 因此, 本文提出的模型融合了 coarse-to-fine 选择机制 (见图 2), 在基于内容的 coarse 程度的注意力机制基础上, 赋予每个隐含层被选择的概率, 并通过计算选择每部分的概率和每部分基于内容的注意力机制下的权重的乘积, 最终获得 fine 程度的注意力权重. 基于数值型数据的 content 部分的特点, 本文在实现这种机制的过程中做了适应性的改动, 首先将 field 和 content 两部分的 Embedding 也即 f_i 和 c_i 进行 concatenation, 得到 $r_i:[f_i;c_i]$, 在 r_i 经过 Encoder 得到隐含层状态后, 隐含层将会进入 pre-selec. 预选模块获得一项被选择的概率.

Coarse aligner 根据 field 部分的 embedding 也即 f_i, 和 r_i 的编码向量 h_i 来分别计算两部分的注意力权重 $\widetilde{\alpha_{t,i}^{(f)}}$ 和 $\widetilde{\alpha_{t,i}^{(c)}}$, 计算方式如式 (11) 和式 (12), 最终联合两部分权重做 Softmax 处理得到 coarse aligner 的输出 $\alpha_{t,1:N}^{\text{content}}$, 计算方式如式 (13).

$$\widetilde{\alpha_{t,i}^{(f)}} = f_i^{\mathrm{T}} \cdot \left(W^{(f)} y_{t-1} + b^{(f)}\right) \tag{11}$$

$$\widetilde{\alpha_{t,i}^{(c)}} = h_i^{\mathrm{T}} \cdot \left(W^{(c)} y_{t-1} + b^{(c)}\right) \tag{12}$$

$$\alpha_{t,i}^{\text{content}} = \frac{\exp\left\{\widetilde{\alpha_{t,i}^{(f)}}\widetilde{\alpha_{t,i}^{(c)}}\right\}}{\sum_{j=1}^{N}\exp\left\{\widetilde{\alpha_{t,j}^{(f)}}\widetilde{\alpha_{t,j}^{(c)}}\right\}} \tag{13}$$

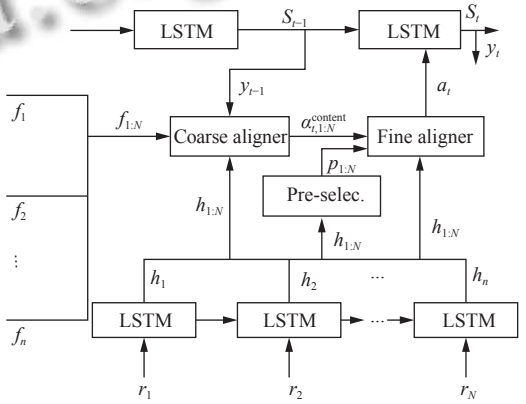

图 2 Coarse-to-fine aligner 选择机制

在预选模块 pre-selec.中每部分隐含层被选择的概率计算方式如下:

$$p_i = Sigmoid(Ph_i) \tag{14}$$

在 fine aligner 中最终经过 coarse-to-fine aligner 选择机制优化的基于内容的注意力机制计算方式如下:

$$\alpha_{t,i}^{\text{content(coarse-to-fine)}} = \frac{p_i \alpha_{t,i}^{\text{content}}}{\sum_{i=1}^{N} p_i \alpha_{t,i}^{\text{content}}} \tag{15}$$

最终的 fine aligner 输出的中间语义向量计算方式如下:

$$a_t = \sum_{i=1}^{N} \alpha_{t,i}^{\text{content(coarse-to-fine)}} h_i \tag{16}$$

2.2 融合 linked-based attention 注意力机制

在第 2.1 节中介绍了基于结构化数据到文本生成任务的 3 个优化点, linked-based attention 注意力机制即是针对句子层面的规划提出的注意力机制. Sha 等基于 LSTM 实现 Encoder-Decoder 框架, 并根据句子规划的思路, 提出了一种基于链接的混合注意力机制, 将其

应用在 Encoder 部分, 模拟不同领域之间的关系, 明确地对这类信息进行建模[8]. 由于针对数值型数据生成的分析性文本对语序及内容的描述顺序有较高的要求, 因此本文提出的模型也将这种机制设计进入了模型当中.

Linked-based attention 机制的基本思路就是对要描述的 field 进行关系建模, 使得模型捕捉到 field 部分使用的词汇在文本中出现的先后顺序, 其实现是使用一个$N_f \times N_f$大小的链接矩阵来存储 field 部分的词汇的先后关系, N_f表示 field 部分的种类数, $L\left[f_j, f_i\right]$表示在f_i被提及后f_j被提及的可能性. Linked-based attention 的计算方式如式 (17), 其中, $\alpha_{t-1,j}$表示上一时刻使用的注意力权重.

$$\begin{aligned}\alpha_{t,i}^{\text{link}} &= Softmax\left\{\sum_{j=1}^{N}\alpha_{t-1,j}\cdot\mathcal{L}\left[f_j,f_i\right]\right\}\\ &= \frac{\exp\left\{\sum_{j=1}^{N}\alpha_{t-1,j}\cdot\mathcal{L}\left[f_j,f_{i'}\right]\right\}}{\sum_{i'=1}^{N}\exp\left\{\sum_{j}\alpha_{t-1,j}\cdot\mathcal{L}\left[f_j,f_{i'}\right]\right\}}\end{aligned} \tag{17}$$

本文设计的模型中使用相同的链接矩阵对 field 部分的内容进行关系建模, 但在计算上一时刻使用的注意力权重$\alpha_{t-1,j}$时, 则要联合融合了 coarse-to-fine aligner 选择机制的基于内容的注意力机制下的权重$\alpha_{t,i}^{\text{content(coarse-to-fine)}}$和 linked-based attention 注意力机制下的权重$\alpha_{t,i}^{\text{link}}$两部分, 最终的注意力机制下的权重使用$\alpha_{t,i}^{\text{hybrid}}$表示, α_t^{hybrid}的计算方法如下:

$$z_t = \sigma\left(w^{\mathrm{T}}\left[h'_{t-1};e_t^{(f)};y_{t-1}\right]\right) \tag{18}$$

$$\alpha_t^{\text{hybrid}} = \widetilde{z}_t\cdot\alpha_t^{\text{content(coarse-to-fine)}} + (1-\widetilde{z}_t)\cdot\alpha_t^{\text{link}} \tag{19}$$

其中, z_t表示一种使用自适应门的门限控制函数, 其中h'_{t-1}表示$t-1$时刻解码器的隐含层状态, e_t表示 field 部分的编码结果和t时刻 linked-based attention 注意力机制下的注意力权重加权求和的结果, y_{t-1}表示上一时刻的生成词. 最后通过两部分注意力机制下的权重计算最终的混合注意力机制权重, 其示意图如图 3.

2.3 训练函数

该模型的优化函数也即训练目标为训练集中每个句子$y_0y_1y_2\cdots y_t$的最大似然估计.

$$J = -\sum_{t=1}^{T}\log p\left(y_t \mid y_0\cdots y_{t-1}\right) \tag{20}$$

其中, $p\left(y_t \mid y_0\cdots y_{t-1}\right)$为解码器部分得到的全部词汇的生成概率$p_t(w)$(见式(10)). 整个模型通过反向传播进行端到端的训练, 在文本生成中, 其通过贪心算法选择t时刻概率最大的单词, 并且当生成词为特殊符号<EOS>时, 表示解码结束, 文本生成完毕.

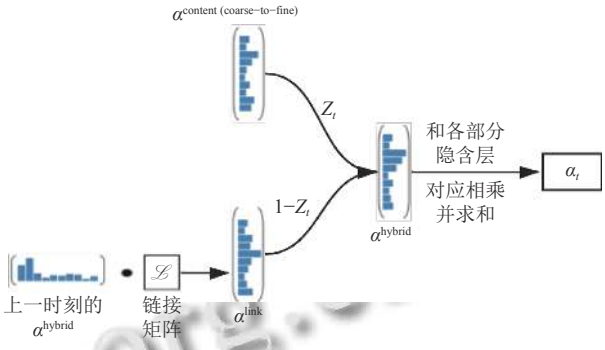

图 3 混合注意力机制示意图

3 实验分析

3.1 数据集的构建

从寻找数值型结构化数据以及数据获取的便捷性出发, 我们通过财经金融网站提供的数据下载接口获取了 2020 年 1 月 14 日至 2021 年 5 月 26 日的 A 股市场数据, 并将其整理成模型训练所需的标准格式 [field: content] 作为最终的数值型结构化数据, 并且联合每日股市点评的摘要及 A 股每日播报的资讯作为已有结构化数据对应的分析性文本, 同时使用了中文分词工具 jieba 对获取的文本进行了分词处理[12].

3.2 评价方法

我们选择了 *BLEU* 和 *ROUGE-L* 作为评估本文提出的生成数值型数据分析性文本模型的评价指标[13,14]. 其中 *BLEU* 是文本生成任务中通用的评价方式, 其基本思路是比较机器生成的文本和参考文本中 n-gram 的重合度, 且其是一种基于准确率的评价指标, 计算方法如式 (21)[15]:

$$BLEU = BP\times\exp\left(\sum_{n=1}^{N}W_n\times\log P_n\right) \tag{21}$$

其中, *BP* 表示惩罚因子, 其考虑了机器生成的文本长度小于参考文本长度时评分却较高的情况, 具体计算方式如式 (22) 和式 (23). 其中, lc表示机器生成的文本的长度, lr表示最短的参考文本的长度. P_n表示准确率, 计算方式如式(24), 其中, $h_i(C)$表示第i个匹配到的 n-gram在C里面出现的次数, $h_i\left(S_j\right)$表示第i个匹配到的 n-gram在第j个参考译文出现的次数. W_n指n-gram的权重, 且n一般计算到 4.

$$BP=\begin{cases}\exp(1-lr/lc), & lc\leqslant lr\\ 1, & lc>lr\end{cases} \tag{22}$$

$$P_n=\frac{\sum_{i\in\text{n-gram}}\min\left(h_i(C),\max\limits_{j\in m}h_i(S_j)\right)}{\sum_{i\in\text{n-gram}}h_i(C)} \tag{23}$$

虽然 *BLEU* 的使用极为广泛, 且计算速度较快, 但是其并不考虑语义和句子结构, 因此, 我们又选择了基于召回率的 *ROUGE-L*, 其属于 ROUGE 评价方法中的一种, 而 ROUGE 的基本思路是比较参考文本的 n-gram 和机器生成文本的重合度. 与此同时, *ROUGE-L* 对进行比较的 n-gram 进行了优化, 将其改为了最长公共子序列. 通过比较最长公共子序列的重合度, 能更好地评价生成文本的内容, *ROUGE-L* 的计算方式如下[16]:

$$R_{LCS}=\frac{LCS(C,S)}{len(S)} \tag{24}$$

$$P_{LCS}=\frac{LCS(C,S)}{len(C)} \tag{25}$$

$$ROUGE\text{-}L=\frac{(1+\beta^2)R_{LCS}P_{LCS}}{R_{LCS}+\beta^2P_{LCS}} \tag{26}$$

其中, $LCS(C,S)$表示机器生成的文本与参考文本的最长公共子序列的长度, $len(S)$和$len(C)$分别表示参考文本和机器生成文本的长度.

3.3 实验结果与分析

表 1 为实验使用的结构化数据和本文模型生成的描述性文本示例. 我们在自行构建的数据集上进行了模型效果的对比分析, 通过把模型生成的文本与测试集中的参考文本进行对比, 得到了在 *BLEU* 和 *ROUGE-L* 评价方法下的评价结果, 见表 2.

其中参与对比的第 1 个模型为 baseline, 此模型为本文在基本训练框架和机制研究中介绍的基础模型, 即通过 LSTM 实现 Encoder-Decoder 框架, 并且使用了基于内容的注意力机制和 copy 机制的结构化数据到文本生成模型, 第 2 个模型为在 baseline 基础上使用了 linked-based attention 注意力机制的模型, 第 3 个模型为基于 OpenAI 开发的 GPT2 预训练模型的 multi-conditioned Transformer[17], 第 4 个模型即为本文提出的融合了 coarse-to-fine aligner 选择机制和 linked-based attention 注意力机制的模型.

表 1 结构化数据和本文模型生成的对应描述性文本

结构化数据	机器文本
指数: 沪指, 收盘价: 2 863.567 3, 最高价:2 870.493 9 最低价: 2849.238 1, 开盘价: 2 851.015 8, 涨跌额: −33.858 0 涨跌幅: −1.168 6, 成交量: 169 895 363.0 成交金额: 191 000 000 000.0 指数: A股指数 收盘价: 3 255.410 2 最高价: 3 276.749 5 最低价: 3 254.141 4 开盘价: 3 269.933 9 涨跌额: −9.171 涨跌幅: −0.280 9 成交量: 229 805 949 成交金额: 271 431 254 181.0 指数: 创业板指数 收盘价: 1 922.557 3 最高价: 1 945.766 7 最低价: 1 922.130 8 开盘价: 1 941.123 9 涨跌额: −12.514 3 涨跌幅: −0.646 7 成交量: 2 410 886 066 成交金额: 46 864 268 374.9 指数: 深证成指 收盘价: 10 988.767 1 最高价: 11 086.808 6 最低价: 10 983.415 开盘价: 11 074.886 9 涨跌额: −51.434 4 涨跌幅: −0.465 9 成交量: 16 469 717 711 成交金额: 222 275 061 483.0	中国股市周四收盘上涨, 因全球投资人信心, 沪指强势回升. 上证指数收报动能, 涨−0.85%, 深证成指收报10 829.045 4, 涨−0.500 6%, 创业板指收报创业板, 涨–0.234 9%, 全天成交额.

表 2 本文模型和其他模型的结果对比

模型	*BLEU*	*ROUGE-L*
Baseline (LSTM+content-based Attention)	23.469	4.44
Baseline+linked-based attention	23.904	14.65
Multi-conditioned Transformer (GPT2 based)	0.00	0.00
本文	24.255	17.93

从表 2 中的结果看, 相较于仅仅使用基于内容的注意力机制的模型, 增加使用 linked-based attention 注意力机制的模型在 *ROUGE-L* 指标上有较大的提升, 说明生成的文本在语义和内容上有很大的改进, 同时其在 *BLEU* 指标上也有一定提升. 而本文提出的模型在使用 linked-based attention 注意力机制的基础上, 还融合了 coarse-to-fine 选择机制, 该模型在 *BLEU* 和 *ROUGE-L* 指标上均进一步获得了提升. 而基于 GPT2 预训练模型的 multi-conditioned Transformer 在两个指标上均未得分, 可以见得生成数值型数据的描述性文本时基于 Transformer 的 GPT2 等预训练模型直接根据语义信息生成文本的方法无法生成有效的文本, 说明解决此类任务按照内容规划和句子实现的思路进行仍是最稳妥

有效的方案. 最终, 通过与不同类型模型 (如基于 GPT2) 的横向比较和与同类型模型的纵向比较, 说明本文提出的模型在解决针对数值型结构化数据生成分析性文本这一特定领域的结构化数据到文本生成任务时具有很好的适配性, 能够比已有的模型更好地解决这类问题.

4 总结与展望

基于结构化数据的文本生成是自然语言生成领域重要的研究方向, 其是新闻自动报道和报告自动生成等领域的关键技术. 从为报告中的数值型数据自动生成分析性文本出发, 本文提出了一种融合 coarse-to-fine aligner 选择机制和 linked-based attention 注意力机制的编码器-解码器文本生成模型, 通过在自行构建的数据集上进行训练和测试, 并通过和已有的模型进行性能对比, 说明了该模型在生成数值型数据的分析性文本这一特定领域的结构化数据到文本生成任务上具有更好的表现.

生成数值型数据的分析性文本是实现报告自动生成的核心内容, 但目前解决该类任务的文本生成技术仍有较大提升和改进的空间, 包括生成的文本的长度、内容合理性、数据的正确性及对数据的统计分析水平. 后续的工作将会重点加强文本生成的可控性, 提高生成文本的逻辑性及对各类数据描述的严谨性, 使得模型在实际应用中具有更好的鲁棒性[18]. 而随着人工智能的不断发展, 深度学习网络也将逐渐增加更多的逻辑推理能力, 基于结构化数据的文本生成技术也将逐渐具有更多的统计和推理能力, 并生成更智能的分析性文本[19–21].

参考文献

1 曹娟, 龚隽鹏, 张鹏洲. 数据到文本生成研究综述. 计算机技术与发展, 2019, 29(1): 80–84, 89. [doi: 10.3969/j.issn.1673-629X.2019.01.017]

2 Gong JP, Ren W, Zhang PZ. An automatic generation method of sports news based on knowledge rules. Proceedings of the 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). Wuhan: IEEE, 2017. 499–502. [doi: 10.1109/ICIS.2017.7960043]

3 Cho K, van Merriënboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: ACL, 2014. 1724–1734. [doi: 10.3115/v1/D14-1179]

4 Tran VK, Nguyen LM, Tojo S. Neural-based natural language generation in dialogue using RNN encoder-decoder with semantic aggregation. Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. Saarbrücken: ACL, 2017. 231–240. [doi: 10.18653/v1/W17-5528]

5 Ding XF, Jiang WJ, He JW. Generating expert's review from the crowds ': Integrating a multi-attention mechanism with encoder-decoder framework. Proceedings of the 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). Guangzhou: IEEE, 2018. 954–961. [doi: 10.1109/SmartWorld.2018.00170]

6 Niranjan A, Shaik MAB. Improving grapheme-to-phoneme conversion by investigating copying mechanism in recurrent architectures. Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Singapore: IEEE, 2019. 442–448. [doi: 10.1109/ASRU46091.2019.9003729]

7 Mei HY, Bansal M, Walter MR. What to talk about and how? Selective generation using LSTMs with coarse-to-fine alignment. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: ACL, 2016. 720–730. [doi: 10.18653/v1/N16-1086]

8 Sha L, Mou LL, Liu TY, *et al*. Order-planning neural text generation from structured data. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AIAA, 2018. 5414–5421.

9 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv: 1409.0473, 2014.

10 Luong T, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015. 1412–1421.

11 Bao JW, Tang DY, Duan N, *et al*. Table-to-text: Describing table region with natural language. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AIAA, 2018. 5020–5027.

12 Chang EN, Shen XY, Zhu DW, *et al*. Neural data-to-text generation with LM-based text augmentation. Proceedings of the 16th Conference of the European Chapter of the

Association for Computational Linguistics: Main Volume. Online: ACL, 2021. 758–768.
13 Celikyilmaz A, Clark E, Gao JF. Evaluation of text generation: A survey. arXiv: 2006.14799, 2020.
14 Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language. Proceedings of the 9th Workshop on Statistical Machine Translation. Baltimore: ACL, 2014. 376–380. [doi: 10.3115/v1/W14-3348]
15 Papineni K, Roukos S, Ward T, *et al*. BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia: ACM, 2002. 311–318. [doi: 10.3115/1073083.1073135]
16 Lin CY. ROUGE: A package for automatic evaluation of summaries. Proceedings of ACL Workshop on Text Summarization Branches Out. Barcelona: ACL, 2004. 74–81.
17 Lee J. Transforming multi-conditioned generation from meaning representation. Proceedings of the International Conference on Recent Advances in Natural Language Processing. Online: INCOMA Ltd., 2021. 805–813.
18 李锦乾, 张冬茉, 姚天方. 自然语言生成中的句子结构优化处理. 计算机应用研究, 1998, (1): 54–58.
19 康波, 孟祥飞, 夏梓峻. 应用驱动的大数据与人工智能融合平台建设. 数据与计算发展前沿, 2019, 1(1): 35–45. [doi: 10.11871/jfdc.issn.2096-742X.2019.01.005]
20 廖方宇, 洪学海, 汪洋, 等. 数据与计算平台是驱动当代科学研究发展的重要基础设施. 数据与计算发展前沿, 2019, 1(1): 2–10. [doi: 10.11871/jfdc.issn.2096-742X.2019.01.002]
21 孙哲南, 张兆翔, 王威, 等. 2019 年人工智能新态势与新进展. 数据与计算发展前沿, 2019, 1(2): 1–16.