

基于主题感知的跨模态序列到序列生成模型

张 旭,王旭强,田雨婷,杨 青,孟 洁

(国网天津市电力公司 信息通信公司,天津 300010)

摘 要:结构化数据和非结构化文本被视为两种不同的模态。数据到文本生成是自然语言生成领域中一个重要的跨模态任务,该任务的目标是对于给定的结构化数据,生成一段文本用以描述结构化数据中包含的关键信息。近年的研究工作通常关注于描述性文本的生成,虽然取得了一定的研究进展,但仅能做到信息的传递而不能带来任何增益。为解决这一问题,本研究数据到分析性文本的生成,并针对该任务提出一个基于主题感知的跨模态序列到序列模型。该模型在编码器-解码器结构的基础上,引入数据表的主题信息以保证生成文本与数据表之间的主题一致性,提高生成文本的质量。为验证模型的性能,提出两个真实数据集,并与其他 6 个模型进行对比实验,结果表明,提出的模型取得了最好的性能。

关键词:自然语言生成;结构化数据;分析性文本;主题感知;跨模态

中图分类号:TN929.5

文献标志码:A

Topic-aware based cross-modal sequence-to-sequence generation model

ZHANG Xu, WANG Xuqiang, TIAN Yuting, YANG Qing, MENG Jie

(Information Communication Company, State Grid Tianjin Electric Power Company, Tianjin 300010, China)

Abstract: The structured data and the unstructured text can be regarded as two different modalities. Data-to-text generation is an important cross-modal task in natural language generation field. Given structured data, this task aims to generate the corresponding text which describes the key information of the structured data. Recently, many studies generally focus on the descriptive text generation. Although these studies have achieved great progress, they can only present structured data without any information gain. To deal with this problem, this paper explores the data-to-analysis generation task, and proposes a topic-aware based cross-modal sequence-to-sequence model. Based on the encoder-decoder structure, the model introduces the topic information of the structure data to ensure the topic consistency between the generated text and the structure data and improve the quality of the generated text. To verify the performance of the proposed model, two real datasets are constructed and a series of experiments compared with six baselines are conducted. Experimental results show that the proposed model achieves the best performance.

Key words: natural language generation; structured data; analytical text; topic-aware; cross-modal

随着信息技术的发展,各行各业积累了大量的行业数据。这些数据与人类社会的生产管理息息相关,是各领域分析研究的主要对象。结构化数据格式简单、便于记录与存储,是最普遍存在的数据形式之一,例如公司的财务报表、设备传感器记录等。但结构化数据通常具有很强的领域性,缺乏行业知识的人很难理解其数值与指标背后的含义。因此,如何准确高效地传达结构化数据中的语义信息是一个重要的研究方向。数据到文本的生成逐渐成为自然语言生成领域一个重要的跨模态生成任务,研究成果已广泛应用于天气预报^[1]、新闻媒体^[2-4]等领域。早期的研究^[5-7]主要通过人工规则、模块式系统等方式生成文本。随着深度学习

收稿日期:2020-03-16

基金项目:天津市科技计划项目(18ZXZNGX00310);天津市电力公司科技项目(KJ19-1-38)

作者简介:张 旭(1983—),男,天津人,高级工程师,博士,主要从事电网信息通信技术研究. E-mail: zhangxu_zwhx@126.com

王旭强(1989—),男,天津人,工程师,硕士,主要从事电网信息通信技术研究. E-mail: wangxuqiang_power@126.com

技术的发展,近期的研究主要采用端到端的学习方式^[8-11],通过数据驱动的形式进行学习和训练,避免繁琐的人工规则,并使得生成的结果更灵活多样。

然而,数据到文本生成还有很大的空间亟待探索。其中一个重要原因是,传统的数据到文本生成任务只涉及数据内容的复述,而不涉及深入的分析推理,这在一定程度上限制了数据到文本生成任务的发展。例如,“公司本期流动比率 2.06,去年同期为 1.81”这句话虽然准确传达了数据表中的信息,但对于缺乏专业知识的读者来说仍然无法准确理解其背后的含义。如果文本内容为“公司偿债能力维持稳定,短期偿债能力具有一定的保障”,则更容易被读者理解。因此,单纯的数据描述在很多时候无法满足人们的需求。若对结构化数据表中的内容进一步分析与解读,则可以获得更好的信息传递效果。

本研究关注数据到分析性文本生成任务,提出一种基于主题感知的跨模态序列到序列模型。具体地,将编码器-解码器结构作为基本框架,并引入数据表的主题建模,以保证生成文本和数据表之间的主题一致性。为验证模型效果,构建了 THS 和 IATA 两个真实数据集,并与基于模板的生成模型、基于语言模型的生成模型以及基于神经网络的生成模型等 6 种模型进行了实验对比与分析。实验结果表明,本模型获得最优的性能。

1 相关工作

数据到文本生成旨在基于给定的结构化数据来生成非结构化的文本,是自然语言生成领域的一个重要研究内容,其中,结构化的数据和非结构化的文本被视为两种不同的模态。传统方法^[5-7]通常将该任务分解为内容规划、句子规划以及表层实现三个独立的子任务,并串行地执行这三个子任务以实现从结构化数据模态到非结构化文本模态的生成。这种方法虽然结构简单并且易于理解,但存在传递错误、模型性能严重依赖手工特征的有效性等问题。

近年来,随着深度学习技术的发展,端到端的学习方式逐渐成为数据到文本生成的主流方法。此类方法以数据驱动的形式进行学习和训练,首先采用编码层将结构化数据映射到低维、稠密的语义向量空间,随后采用解码层基于该语义空间生成非结构化的文本,从而实现跨模态的文本生成。上述过程避免了繁琐的人工规则编写,并使得生成结果灵活多样。Lebret 等^[8]利用条件语言模型实现人物传记的生成。Mei 等^[12]使用基于复制机制的序列到序列模型提升内容选择的效果。Li 等^[13]采用两阶段方式生成文本,首先生成文本模板,再采用延迟复制机制填入记录中的数值。Wiseman 等^[14]关注数据到文档的生成,并在 seq2seq 模型中引入复制机制和损失重构机制。Gong 等^[15]采用层次编码的方式学习数据记录的语义。Iso 等^[16]设计了内容追踪模块,在生成文本的过程中通过跟踪数据记录来提升文本的真实性并减少冗余。Puduppully 等^[10]在模型中显式地增加内容选择和内容规划模块,提升模型的内容组织能力。同年,Puduppully 等^[17]还提出了基于实体建模的生成模型。此外,还有部分研究^[18-19]基于半隐马尔科夫模型来实现数据记录到文本的对齐与生成,提升系统的可解释性与可控性。

虽然上述方法在很大程度上提升了文本生成质量,但均仅关注描述性文本的生成,即通过文本对表格的重要内容进行复述,未涉及对表格内容的分析、提炼和推理,这使得生成文本仅能做到信息的传递而不能带来任何增益。针对这一问题,本研究关注数据到分析性文本生成任务,提出基于主题感知的跨模态序列到序列模型来学习如何对表格内容进行分析和描述,保证文本和数据表之间的主题一致性,进而提升生成文本的质量。

2 基于主题感知的跨模态序列到序列模型——EDAT 模型

2.1 问题介绍

给定数据表 $s = \{s^t, s^r\}$, 其中 s^t 为数据表的标题, $s^r = \{r_j\}_{j=1}^{|r|}$ 为数据表的记录集合, $|r|$ 表示记录个数。数据表中的每条记录 r_j 都包含 3 个属性,即 $r_j = \{r_j^e, r_j^t, r_j^v\}$, 其中 r_j^e 代表记录的实体(即行表头), r_j^t 代表记录的类型(即列表头), r_j^v 代表记录值。数据到分析性文本生成任务旨在根据给定的数据表生成对应的分析性文本 $y = \{y_1, y_2, \dots, y_{|y|}\}$, 其中 $|y|$ 代表文本的长度。

2.2 模型结构

为更好地解决数据到分析性文本生成问题,提出基于主题感知的跨模态序列到序列模型——EDAT 模型,其结构如图 1 所示。具体地,首先采用序列到序列的模型框架实现从结构化表格模态到非结构化文本模态的生成,再引入关于数据表类别的主题特征表示,对生成文本的主题进行约束,从而得到更贴合表格内容的分析性文本。

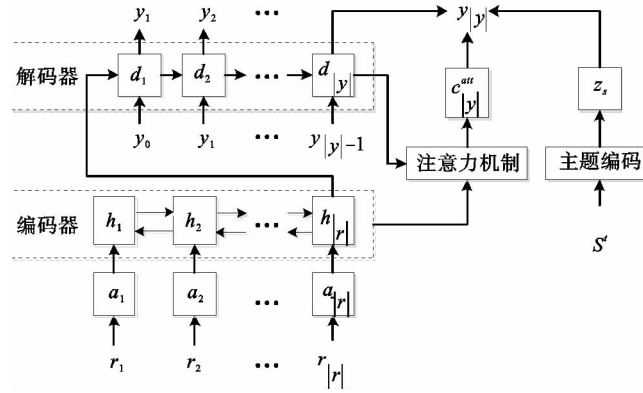


图 1 基于主题感知的跨模态序列到序列模型结构

Fig. 1 Topic-aware cross-modal sequence-to-sequence model architecture

2.2.1 编码层

给定数据表记录集合 s^r 以及数据表标题 s^t , 将 s^r 转化为记录序列 $s^q = \{r_1, r_2, \dots, r_{|r|}\}$, 并构建编码层来建模记录序列的隐藏向量表示 $H = \{h_1, h_2, \dots, h_{|r|}\}$ 以及数据表的主题表示 z_s 。

1) 记录编码

对于记录序列 s^q 中的每个记录 r_j , 编码层将其中包含的 3 个属性 r_j^e, r_j^t, r_j^v 分别映射到低维、稠密的特征向量空间, 得到 3 个对应的向量表示 $r_j^e \in \mathbf{R}^{d_r}, r_j^t \in \mathbf{R}^{d_r}, r_j^v \in \mathbf{R}^{d_r}$, 其中 d_r 表示每个属性特征的维度。

考虑到属性 r_j^v 通常为数值型的记录值, 本节进一步对属性 r_j^v 进行数值编码。由于数值之间的差异程度通常会导致语义的差异, 例如当表示下降幅度时, 数值“0.25”和“15”表达的语义分别为“略有下降”和“大幅下降”, 因此首先根据数据集中数值的分布情况将所有数值划分为不同的区间, 例如 <0 区间, $0 \sim 10$ 区间等, 不同的区间将对应不同的参数设置。在数值编码时, 首先根据 r_j^v 对应的区间范围选择参数 W_k^q 和 b_k^q , 随后将 r_j^v 的具体数值输入到线性变换层得到指示向量 q_j^v , 并通过该指示向量对量化单元的嵌入矩阵进行加权求和得到数值特征表示, 具体计算过程为:

$$q_j^v = W_k^q r_j^v + b_k^q, \quad (1)$$

$$r_j^q = q_j^v \cdot Q. \quad (2)$$

其中, $Q \in \mathbf{R}^{m \times d_q}$ 为量化单元的嵌入矩阵; M 为量化单元个数; d_q 为数值特征的维度。通过上述方法得到的数值特征既考虑到数值本身的大小, 又避免语义的分散, 从而提升模型对数值的理解能力。

通过拼接上述 4 个特征向量可以得到每条记录的向量表示:

$$a_j = [r_j^e; r_j^t; r_j^v; r_j^q]. \quad (3)$$

基于记录的向量表示序列 $A = \{a_1, a_2, \dots, a_{|r|}\}$, 采用长短期记忆网络 (long short-term memory, LSTM)^[20] 编码记录序列的隐藏向量表示。具体地, LSTM 在第 t 个时间步的计算过程为:

$$i_t = \sigma(W_{ii}a_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}), \quad (4)$$

$$f_t = \sigma(W_{if}a_t + b_{if} + W_{hf}h_{t-1} + b_{hf}), \quad (5)$$

$$o_t = \sigma(W_{io}a_t + b_{io} + W_{ho}h_{t-1} + b_{ho}), \quad (6)$$

$$\tilde{c}_t = \sigma(W_{ig}a_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}), \quad (7)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \quad (8)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t). \quad (9)$$

其中: \mathbf{h}_t 为第 t 个时间步 LSTM 输出的隐藏状态; \mathbf{c}_t 为第 t 个时间步的记忆单元状态; $\mathbf{i}_t, \mathbf{f}_t$, 和 \mathbf{o}_t 为 LSTM 中的输入门、遗忘门与输出门; $\sigma(\cdot)$ 和 $\tanh(\cdot)$ 分别为 Sigmoid 与 tanh 激活函数; \odot 表示矩阵元素相乘; \mathbf{W}_* 和 \mathbf{b}_* 为模型参数。

为同时捕获数据记录序列在两个方向上的隐藏特征,使用双向 LSTM 对数据记录序列编码,并将前向和后向编码结果进行拼接作为最终的记录隐藏表示:

$$\vec{\mathbf{h}}_t = \overrightarrow{\text{LSTM}}_{\text{enc}}(\mathbf{a}_t, \vec{\mathbf{h}}_{t-1}), \quad (10)$$

$$\overleftarrow{\mathbf{h}}_t = \overleftarrow{\text{LSTM}}_{\text{enc}}(\mathbf{a}_t, \overleftarrow{\mathbf{h}}_{t-1}), \quad (11)$$

$$\mathbf{h}_t = [\vec{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t]. \quad (12)$$

由此,可以得到记录序列的隐藏表示 $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_{|r|}\}$ 。

2) 主题编码

直观上,在写作过程中,当围绕不同的主题进行叙述时,词汇的使用通常存在较大差异。因此,在数据到分析性文本生成任务中,如何准确学习数据表的主题是一个关键问题。

数据表的主题通常由标题 s^t 标记,首先将标题相同的数据表的文本进行聚合,对其中出现的词进行统计,并构建主题-词的共现矩阵 $\mathbf{U} \in \mathbf{R}^{L \times |D|}$,其中 L 代表主题个数, $|D|$ 代表由数据集中全部词构成的词表的大小,矩阵元素 U_{ij} 代表第 i 种主题对应的第 j 个词的特征值。本节使用词在该主题类别下出现的次数作为特征值。根据共现矩阵,可以选出每个主题下的高频主题词表 $\mathbf{w}_i = \{\mathbf{w}_i^1, \mathbf{w}_i^2, \dots, \mathbf{w}_i^{L_i}\}$,其中 L_i 代表高频词词表的大小。将每个主题下高频词的词向量进行平均作为该主题的特征表示,即:

$$\mathbf{z}_i = \frac{1}{L_i} \sum_{l=1}^{L_i} \mathbf{w}_i^l. \quad (13)$$

其中: \mathbf{z}_i 为第 i 个主题的特征表示, \mathbf{w}_i^l 为第 i 个主题对应的高频主题词表中第 l 个词对应的预训练词向量。由此得到的主题特征集合 $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_L\}$ 。

给定数据表标题 s^t ,根据查表法从主题特征集合中选择对应的主题表示 \mathbf{z}_s 。在解码层,通过引入数据表的主题表示,可以指导生成过程中词项的选择,以获得更好的生成结果。

2.2.2 解码层

基于编码层得到的记录序列的隐藏向量表示 \mathbf{H} 以及数据表的主题表示 \mathbf{Z} ,使用 LSTM 作为解码器生成分析性文本 $\mathbf{y} = \{y_1, y_2, \dots, y_{|y|}\}$ 。

在解码过程的第 t 个时间步, LSTM 单元的输入为上一步预测的词对应的词向量 \mathbf{y}_{t-1} 以及解码器上一步的隐藏表示 \mathbf{d}_{t-1} ,得:

$$\mathbf{d}_t = \text{LSTM}_{\text{dec}}(\mathbf{y}_{t-1}, \mathbf{d}_{t-1}). \quad (14)$$

第 1 个时间步中, \mathbf{y}_0 被初始化为全零向量,并将编码器中前向 LSTM 的最后一个隐藏表示与反向 LSTM 的最后一个隐藏表示进行拼接作为 \mathbf{d}_0 :

$$\mathbf{d}_0 = [\vec{\mathbf{h}}_{|r|}; \overleftarrow{\mathbf{h}}_0]. \quad (15)$$

在生成每个词时,除了考虑上一个生成的词之外,还应该关注到原始数据表中重要的信息。在解码层引入注意力机制,以实现对数据表中信息的选择性关注。给定第 t 个时间步解码器的隐藏表示 \mathbf{d}_t 以及每个数据记录的隐藏表示 \mathbf{h}_j ,可计算注意力权重

$$\beta_{t,j} = \text{softmax}(\mathbf{d}_t^T \mathbf{W}_a \mathbf{h}_j). \quad (16)$$

其中 \mathbf{W}_a 为模型参数。基于 $\beta_{t,j}$,对每个数据记录的隐藏表示进行加权求和,得到上下文向量

$$\mathbf{c}_t^{\text{att}} = \sum_j \beta_{t,j} \mathbf{h}_j. \quad (17)$$

在生成过程中进一步引入数据表对应的主题表示 \mathbf{z}_s ,以保证文本与数据表的主题一致性,提升用词的准确性。将解码器隐藏状态 \mathbf{d}_t 、上下文向量 $\mathbf{c}_t^{\text{att}}$ 以及主题表示 \mathbf{z}_s 进行拼接和非线性变换,得到特征表示

$$\mathbf{d}_t^{\text{att}} = \tanh(\mathbf{W}_d [\mathbf{d}_t; \mathbf{c}_t^{\text{att}}; \mathbf{z}_s] + \mathbf{b}_d) \quad (18)$$

其中, \mathbf{W}_d 和 \mathbf{b}_d 为模型参数。根据该特征表示, 将特征表示映射到词表空间, 进而计算每个词的生成概率:

$$p(y_t | y_{<t}, s) = \text{softmax}(\mathbf{W}_y \mathbf{d}_t^{\text{att}} + \mathbf{b}_y) \quad (19)$$

其中, \mathbf{W}_y 与 \mathbf{b}_y 为模型参数, $y_{<t}$ 代表第 t 个时间步之前生成的词序列, s 代表输入的数据表。在训练过程中, 最小化如下负对数似然函数:

$$l = - \sum_{(y, s) \in D} \log p(y | s) \quad (20)$$

其中, D 表示训练实例集合, $p(y | s)$ 表示正确文本的生成概率。在推理过程中, 对于给定的 s , 预测其对应的分析性文本如下:

$$\hat{y} = \underset{y'}{\operatorname{argmax}} p(y' | s) \quad (21)$$

其中 y' 表示输出文本的候选项。在推理阶段利用集束搜索 (beam search) 来近似地得到最佳生成结果。

3 实验

3.1 数据集

为验证模型的有效性, 分别爬取同花顺网站 (THS) 的上市公司财务诊断以及国际航空运输协会官网 (IATA) 的经济报告, 手工构建了 THS 和 IATA 两个数据集。具体的数据样例如图 2 所示, 其中“|”、“/”以及空格分别用于间隔不同的记录、文本中不同的词以及每个记录中不同的属性。

<p>数据表标题: 盈利能力</p> <p>数据表: 毛利率排名0.0342 毛利率本期82.43 毛利率去年同期50.38 净利率排名0.0881 净利率本期-2.85 净利率去年同期-11.56 总资产净利率排名0.0935 总资产净利率本期-2.42 总资产净利率去年同期-3.53 净资产收益率排名0.0904 ...</p> <p>描述文本: 国农科技/主营/获利/能力/大幅/增强/, /企业/经营/效益/亏损/减轻/, /总资产/收益/能力/有所改善/, /回报/股东/能力/受损/有所/减轻/, /主营业务/利润/贡献/明显/提升/。</p>	<p>数据表标题: Exchange rates</p> <p>数据表: usd 2014 111.3 usd oct-15 119.3 usd nov-15 121.1 usd dec-15 122.4 brl 2014 2.66 brl oct-15 3.84 brl nov-15 3.90 brl dec-15 3.96 mxn 2014 14.75 mxn oct-15 16.53 mxn nov-15 16.60 mxn dec-15 17.20 ...</p> <p>描述文本: the/usd/rose/1.1%/in/dec/, finishing/the/year/up/an/even/10%/yoy/./the/devaluation/of/the/ars/left/it/33%/lower/in/the/month/and/down/53%/yoy/./the/brl/and/cop/ended/the/year/down/49%/and/33%/respectively/vs/the/us/\$</p>
(a) THS数据集	(b) IATA 数据集

图 2 两个数据集的数据样本展示

Fig. 2 The data samples of the two datasets

两个数据集的基本统计情况如表 1 所示, 两个数据集中词在不同主题类型的文本中的分布情况如图 3 所示。

可以看出, THS 数据集中, 不同主题类型下文本用词的差异性很大, 大多数词仅在 2 种类型的文本中出现, 而 IATA 数据集中, 这种差异性同样明显, 有 52% 的词仅在 1 种主题类型中出现。

3.2 实验设置

对于 THS 数据集, 数据表记录中每个属性特征的维度 d_r 以及数值特征表示维度 d_q 均设为 300; 对于 IATA 数据集, 上述维度均设为 100。通过对数据集的分析与统计, 在进行数值区间划分时, 将 THS 数据集划分为 $(-\infty, 0)$, $[0, 1)$, $[1, 10)$, $[10, 50)$, $[50, 100)$, $[100, \infty)$ 6 个数值区间, 量化单元的个数设为 6; 而对于 IATA 数据集, 由于其数值变化范围较大, 因此根据其整数部分划分区间, 量化单元个数设为 20, 且在计算真实数值时利用 \tanh 函数对数值的变化范围进行限制。根据数据集中文本的长度特点, THS 和

表 1 数据集统计

Tab. 1 Statistics of datasets

项目	THS 数据集	IATA 数据集
样本数量	20 190	1 900
平均记录个数	24.45	18.32
平均文本长度	38.17	59.91
主题数	6	10

IATA 数据集的文本最大生成长度分别设为 30 和 50。在两个数据集中,编码器和解码器隐藏状态的维度均设为 300,高频主题词词表的大小 L_t 设为 100。

训练过程中,使用 Adam 优化器优化模型参数,并将批处理大小设为 10,迭代次数设为 60,学习率设为 0.002,dropout 比例设为 0.5。并选取 80% 的样本作为训练集,10% 的样本为验证集,10% 的样本为测试集。

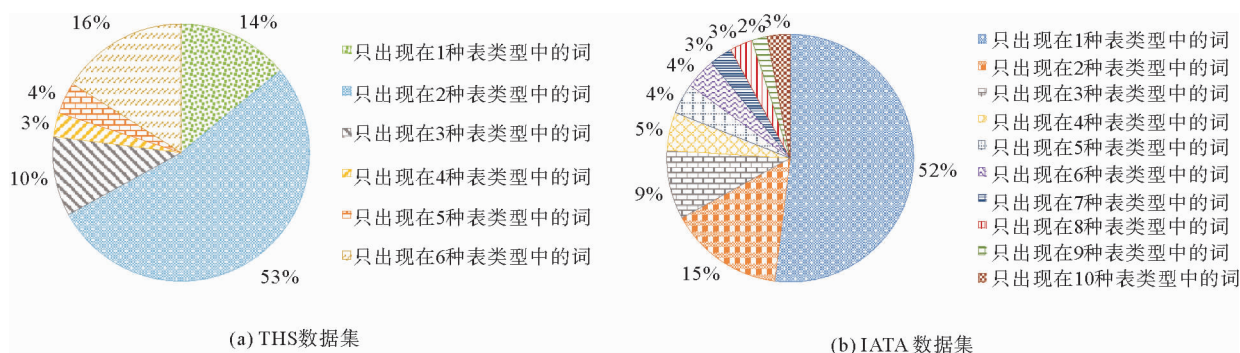


图3 不同主题下文本用词分布情况

Fig. 3 The word distribution on different topics

3.3 对比模型

为了验证提出的 EDAT 模型的效果,与以下模型进行比较:

- 1) KN(Kneser-Ney)模型^[21]:使用 KenLM 工具包训练 5-gram 模型,并且不进行修剪。
- 2) Template:类似于 Wiseman 等^[14]的工作,在训练时,统计不同表类型在各个位置的用词频率,并选择频率最高的词序列构建模板,在推理过程中根据具体的数据表填充模板的空缺处。
- 3) T-NLM:以 Mikolov 等^[22]提出的循环神经网络语言模型为基础,额外输入数据记录的嵌入表示,从而使模型能够利用数据表信息。
- 4) EDA:类似于 Wiseman 等^[14]提出的方法,采用基于注意力机制的序列到序列模型来更清晰地探究主题建模对数据到文本生成的影响。
- 5) EDAT w/o T:在 EDA 模型的基础上仅引入数值编码模块,而不采用主题特征。
- 6) EDAT w/o Q:在 EDA 模型的基础上引入关于数据表类型的主题特征表示,而不采用数值编码。

3.4 评价指标

采用 BLEU^[23] 及 ROUGE^[24] 作为评价指标来判断模型的生成效果。BLEU 是一种基于准确率的相似性度量方法,ROUGE 是一种基于召回率的相似性度量方法。BLEU 和 ROUGE 的值越大,证明生成的结果越符合给定的真实文本。本研究的这两个指标均基于生成文本与真实文本中 4 元组的匹配程度进行计算。

3.5 实验结果

表 2 展示了 EDAT 模型以及对比模型在两个数据集上的生成效果,可以看出本模型在两个数据集的所有指标上均取得了优于对比模型的性能,证明了模型的有效性。

具体地,比较 THS 和 IATA 数据集的实验结果可以看出,THS 数据集上各个方法的生成结果明显优于 IATA 数据集,这是由于 THS 数据集训练样本更多,生成文本的平均长度更短(见表 1)。

THS 数据集上,基于模板的生成模型取得了较好的效果,这是由于 THS 数据集内容的变化性较小,结构更加统一,因此模板可以捕获一定的生成规则,然而其效果依然远低于基于神经网络的模型。IATA 数据集上,模板生成模型的表现比在 THS 数据中更差,且同样低于神经生成模型。这证明了神经生成模

型的优势,也说明 IATA 数据集更加复杂,文本结构的变化性更大。KN 模型两个数据集上表现最差,这是由于该方法在生成过程中仅基于文本中 n 元组的统计信息,而未考虑数据表内容,导致文本无法准确反映数据表的信息。IATA 数据集上 KN 模型在 ROUGE 上效果优于模板模型,这是由于 IATA 数据集中文本变化性较大,模板无法全面覆盖各种情况,而统计语言模型却有更大的覆盖范围。此外,TNLM 方法虽然引入了数据表信息,但是其文本生成的效果比所有基于序列到序列的模型都差,说明序列到序列框架更适用于本文的任务,对记录进行编码有利于生成过程中更有效地利用数据表的信息。

EDAT 模型在两个数据集上的性能均优于 EDA 模型,并且在 ROUGE 指标上优势更加明显,说明本模型生成的结果包含信息更加全面,能够更有效地捕获数据表的主题信息,保持生成文本与原始数据表的主题一致性,从而生成更适合于该主题的主题的文本内容。EDAT w/o T 的性能优于 EDA,说明数值编码能够更好地理解数据表中的数值信息,从而使生成结果更加准确合理。EDAT w/o Q 的性能优于 EDA,EDAT 的性能优于 EDAT w/o T,说明主题建模可以有效保持文本和数据表之间的主题一致性,提升生成效果。从本模型与 EDA、EDAT w/o T 以及 EDAT w/o Q 模型的性能对比可以看出,本模型能够在不同程度上提升分析性文本的生成质量,其效果在 THS 数据集上更加明显。

3.6 主题一致性分析

为验证所提出的模型可以生成更加符合原始数据表主题的主题,对文本生成结果的主题一致性进行分析。首先,通过人工筛选为两个数据集制作能够体现文本主题的中心词词表。随后,计算生成结果在主题一致性方面的得分,具体公式为:

$$TopicScore = \frac{Count(topicw \in Ref \& topicw \in Gen)}{Count(topicw \in Ref)} \quad (22)$$

其中, Gen 表示模型生成的文本, Ref 表示真实文本, $topicw$ 表示中心词。公式(22)的分母代表出现在真实文本中的中心词个数,分子代表生成文本与真实文本中匹配的中心词个数,反映了生成结果对于中心词的覆盖率;分值越大说明生成结果包含了越多正确的中心词,从而与真实文本以及原始数据表具有更高的主题一致性。表 3 显示了不同方法在两个数据集上的得分情况,可以看出,本模型获得了更高的分数。

3.7 样例分析

为更直观地了解本文模型的生成效果,表 4 以 THS 数据集为例,给出 EDAT 模型与对比模型的生成结果。可以看出,本模型具有最优的生成效果。特别地,与 EDAT w/o Q 模型的对比结果显示 EDAT 模型可以有效提升分析结果的准确性;而通过与 EDAT w/o T 的结果示例对比可知,EDAT 模型能够有效地捕获到原始数据表的主体信息,从而生成主题一致的分析性文本。

表 2 EDAT 模型以及对比模型在 THS 和 IATA 数据集上的实验结果

Tab. 2 Experimental results of our model and baselines on THS and IATA datasets

模型	THS 数据集		IATA 数据集	
	BLEU	ROUGE	BLEU	ROUGE
KN	10.47	7.37	8.94	5.73
Temlpate	23.94	24.65	13.37	3.87
T-NLM	43.46	40.94	29.89	22.76
EDA	59.36	67.01	33.61	26.26
EDAT w/o T	63.33	73.80	34.21	26.97
EDAT w/o Q	60.40	68.65	33.93	27.20
EDAT	64.50	75.41	35.23	27.69

表 3 EDAT 模型以及 EDA 模型在 THS 和 IATA 数据集上的主题一致性得分

Tab. 3 The topic consistency scores of our model and EDA model on THS and IATA datasets %

模型	THS 数据集	IATA 数据集
EDA	87.48	66.13
EDAT	89.53	68.14

表4 THS数据集上的样例分析

Tab. 4 The case study on THS

	参考文本	财务状况保持稳定,相对同行业总资产收益能力略有下降,但成长潜力增强。盈利能力维持稳定,主营获利能力保持稳定。
1	EDA	财务状况明显改善,相对同行业企业经营效益下降,但公司规模由扩张变为收缩。盈利能力维持稳定,企业经营效益下降。
	EDAT	财务状况保持稳定,相对同行业总资产收益能力略有下降,但存货变现能力有所增强。盈利能力维持稳定,主营获利能力保持稳定。
	参考文本	主营获利能力削弱非常明显,企业经营效益大幅下降,总资产收益能力大幅提升,回报股东能力大幅增强,主营业务利润贡献略有提升。
2	EDAT w/o Q	主营获利能力削弱非常明显,企业经营效益加倍提高,总资产收益能力大幅提升,回报股东能力增强,主营业务利润贡献略有提升。
	EDAT	主营获利能力削弱非常明显,企业经营效益大幅下降,总资产收益能力大幅提升,回报股东能力明显增强,主营业务利润贡献略有提升。
	参考文本	财务状况明显改善,相对同行业盈利增速步入上升通道,但业绩成长需要其他因素支撑。盈利能力有所加强,主营获利能力大幅增强。成长能力明显改善,营业收入由亏转盈。偿债能力维持稳定,短期债务风险存在但可以控制。
3	EDAT w/o T	财务状况明显改善,相对同行业盈利增速步入上升通道,但回报股东能力修复,盈利能力有所加强,企业经营效益扭亏为盈。成长能力明显改善,盈利增速步入上升通道。偿债能力有所加强,盈利增速略高于债务增速。
	EDAT	财务状况明显改善,相对同行业盈利增速步入上升通道,但业绩成长需要其他因素支撑。盈利能力有所加强,经营扭亏为盈,经营效益改善。成长能力明显改善,营业收入由亏转盈。偿债能力有所加强,盈利增速略高于债务增速。

4 结论

针对数据到分析性文本生成任务开展研究,提出了基于主题感知的跨模态序列到序列模型——EDAT模型。为了实现从结构化数据表模态到非结构化文本模态的转换,采用序列到序列的模型框架,并在此基础上根据主题-词的共现关系学习数据表的主题表示。通过将数据表的主题表示引入解码层,可以有效保证生成文本与数据表之间的主题一致性。为了验证模型的效果,构建了两个真实数据集并进行模型性能验证。实验结果显示,相比其他6个模型,本模型能够更好地捕获不同类型数据表的主题信息,获得最优的性能。

参考文献:

- [1] GOLDBERG E, DRIEDGER N, KITTEDGE R I. Using natural-language processing to produce weather forecasts[J]. IEEE Expert, 1994, 9(2): 45-53.
- [2] VAN DALEN A. The algorithms behind the headlines: How machine-written news redefines the core skills of human journalists[J]. Journalism Practice, 2012, 6(5/6): 648-658.
- [3] CLERWALL C. Enter the robot journalist: users' perceptions of automated content[J]. Journalism Practice, 2014, 8(5): 519-531.
- [4] YOUNG M L, HERMIDA A. From Mr. and Mrs. outlier to central tendencies: computational journalism and crime reporting at the Los Angeles times[J]. Digital Journalism, 2015, 3(3): 381-397.
- [5] REITER E. Building natural language generation systems[M]. United Kingdom: Cambridge University Press, 1996: 248.
- [6] BARZILAY R, LAPATA M. Collective content selection for concept-to-text generation[C] // EMNLP. Vancouver, British Columbia, Canada, 2005: 331-338.
- [7] METEER M W. Bridging the generation gap between text planning and linguistic realization[J]. Computational Intelligence, 1991, 7(4): 296-304.
- [8] LEBRET R, GRANGIER D, AULI M. Neural text generation from structured data with application to the biography domain

- [C]//EMNLP,Austin,Texas,USA,2016;1203-1213.
- [9]BAO J,TANG D,DUAN N,et al.Table-to-text;describing table region with natural language[C]//AAAI,New Orleans,Louisiana,USA,2018;5020-5027.
- [10]PUDUPPULLY R,DONG L,LAPATA M.Data-to-text generation with content selection and planning[C]//AAAI,Honolulu,Hawaii,USA,2018;6908-6915.
- [11]WISEMAN S,SHIEBER S,RUSH A.Learning neural templates for text generation[C]//EMNLP,Brussels,Belgium,2018;3174-3187.
- [12]MEI H,BANSAL M,WALTER M R.What to talk about and how? selective generation using LSTMs with coarse-to-fine alignment[C]//NAACL,San Diego California,USA,2015;720-730.
- [13]LI L,WAN X.Point precisely;towards ensuring the precision of data in generated texts using delayed copy mechanism[C]//COLING,Santa Fe,New Mexico,USA,2018;1044-1055.
- [14]WISEMAN S,SHIEBER S M,RUSH A M.Challenges in data-to-document generation[C]//EMNLP,Copenhagen,Denmark,2017;2253-2263.
- [15]GONG H,FENG X,QIN B,et al.Table-to-text generation with effective hierarchical encoder on three dimensions (row, column and time) [C]//EMNLP,Hong Kong,China,2019;3141-3150.
- [16]ISO H,UEHARA Y,ISHIGAKI T,et al.Learning to select,track,and generate for data-to-text[C]//ACL,Florence,Italy,2019;2102-2113.
- [17]PUDUPPULLY R,DONG L,LAPATA M.Data-to-text generation with entity modeling [C]//ACL,Florence,Italy,2019;2023-2035.
- [18]DOU L,QIN G,WANG J,et al.Data2Text studio;automated text generation from structured data[C]//EMNLP,Brussels,Belgium,2018;13-18.
- [19]QIN G,YAO J,WANG X,et al.Learning latent semantic annotations for grounding natural language to structured data[C]//EMNLP,Brussels,Belgium,2018;3761-3771.
- [20]HOCHREITER S,SCHMIDHUBER J.Long short-term memory[J].Neural Computation,1997,9(8);1735-1780.
- [21]HEAFIELD K,POUZYREVSKY I,CLARK J H,et al.Scalable modified Kneser-Ney language model estimation[C]//ACL,Sofia,Bulgaria,2013;690-696.
- [22]MIKOLOV T,KARAFIÁT M,BURGET L,et al.Recurrent neural network based language model[C]//Interspeech,Makuhari,Chiba,Japan,2010;1045-1048.
- [23]PAPINENI K,ROUKOS S,WARD T,et al.BLEU;A method for automatic evaluation of machine translation[C]//ACL,Philadelphia,PA,USA,2002;311-318.
- [24]LIN C Y.Rouge;A package for automatic evaluation of summaries [C]//In Proceedings of The Workshop on Text Summarization Branches Out.Barcelona,Spain,2004;74-81.

(责任编辑:傅 游)