文章编号: 1000-5641(2019)05-0113-10

基于自注意力机制的冗长商品名称精简方法

傅 裕, 李 优, 林煜明, 周 娅

(桂林电子科技大学 广西可信软件重点实验室, 广西 桂林 541004)

摘要: 大部分电子商务网站为了吸引用户的关注, 通常将商品的很多属性也纳入到商品名称中, 使得商品名称中包括了冗余的信息, 并产生不一致性. 为解决这一的问题, 提出了一个基于自注意力机制的商品名称精简模型, 并针对自注意力机制网络无法直接捕捉商品名称序列特征的问题, 利用门控循环单元的时序特性对自注意力机制进行了时序增强, 以较小的计算代价换取了商品命名精简任务整体性能的提升. 在公开商品短标题数据集 LESD4EC 的基础上, 构造了商品名称精简数据集 LESD4EC_L 和 LESD4EC_S, 并进行了模型验证. 一系列的实验结果表明本, 所提出的自注意力机制冗长商品名称精简方法相对于其他商品名称精简方法在效果上有较大的提升.

关键词: 自注意力机制; 商品名称精简; 门控循环单元

中图分类号: TP391 **文献标志码**: A **DOI:** 10.3969/j.issn.1000-5641.2019.05.009

Self-attention based neural networks for product titles compression

FU Yu, LI You, LIN Yu-ming, ZHOU Ya

(*Guangxi Key Laboratory of Trusted Software*, *Guilin University of Electronic Technology*, *Guilin Guangxi* 541004, *China*)

Abstract: E-commerce product title compression has received significant attention in recent years, since it can facilitate more specific information for cross-platform knowledge alignment and multi-source data fusion. Product titles usually contain redundant descriptions, which can lead to inconsistencies. In this paper, we propose self-attention based neural networks for this task. Given the fact that self-attention mechanism networks cannot directly capture sequence features of product names, we enhance the mapping networks with a dot-attention structure, which was computed for the query and key-value pairs by a gated recurrent unit (GRU) based recurrent neural network. The proposed method improves the analytical capability of the model at a lower relative computational cost. Based on data from LESD4EC, we built two E-commerce datasets of product core

收稿日期: 2019-07-28

基金项目: 国家自然科学基金(U1501252, U1811264, 61562014); 广西自然科学基金重点项目(2018GXNSFDA281049); 桂林电子科技大学研究生优秀论文培养项目(17YJPYSS17); 广西可信软件重点实验室研究课题(kx201916)

第一作者: 傅 裕, 男, 硕士研究生, 研究方向为海量数据管理. E-mail: fuzzyu@foxmail.com.

通信作者: 李 优, 女, 副教授, 硕士生导师, 研究方向为Web数据分析、观点挖掘.
E-mail: liyou@guet.edu.cn.

phrases named LESD4EC_L and LESD4EC_S; we subsequently tested the model on these two datasets. A series of experiments show that the proposed model achieves better performance in product title compression than existing techniques.

Keywords: self-attention mechanism; product titles compression; gated recurrent units

0 引 言

在真实的购物体验中, 消费者通常仅使用几个关键词对商品进行检索; 对于商品知识本身来说, 即可以使用更加精简的商品关键词汇对商品进行知识表示. 为了让商品尽可能地被搜索引擎检索到, 商品提供商通常使用丰富的商品描述词汇组成商品标题, 虽然这种商品名称存在信息冗余, 但得益于电脑端良好的可视化环境, 长文本商品标题依然可以很好地向用户提供商品信息. 然而, 对于不同的电商平台, 受限于商品平台知识体系的不同, 同一商品在不同的商品平台的描述体系中存在不一致性. 这种差异不利于进一步构建跨平台商品知识对齐、多源商品的数据融合. 因此, 有必要消除商品名称中的冗余词汇, 这项工作被称为冗长商品名称精简. 图 1 展示了同一商品在不同电商平台下的显示情况, 标号 ① 为商品在京东[1]平台显示情况, 标号 ② 为在天猫[2]平台显示情况.

图 1 商品在两种不同平台下的显示情况

Fig. 1 A sample product display on different e-commerce platforms

商品名称精简归属于短文本信息提取工作, 然而不同于短文本信息提取, 商品名称表现出以下特点: ① 商品名称长度相对稳定, 维持在一定长度以内; ② 商品标题内部语义依赖关系薄弱; ③ 商品名称词性单一, 通常由名词及少量的形容词构成. 目前已提出的短文本关键信息提取可以归纳为基于词频共现关系的方法和基于强语义依赖关系的方法这两种. 在处理商品名称精简时这两种方法表现出以下不足: ① 基于词频共现关系的方法所提取的目标特征单一, 需要构建全局关联矩阵来计算其属性特征, 在处理大规模数据时存在时间上和空间上的计算开销过大; ② 基于强语义依赖关系的方法并不适用于商品名称精简, 因为真实的商品名称往往并不具备基本的语法依赖关系, 甚至仅仅由一些简单的名词拼凑而成, 无法通过其依赖关系进行进一步分析.

针对商品名称的特点以及现有方法的不足, 本文提出了基于自注意力机制[1]的冗长商品名称精简方法, 主要贡献如下.

[1] https://www.jd.com/
[2] https://www. tmall.com/

(1) 首先提出了一种端到端的基于自注意力机制的冗长商品名称精简方法, 命名为ERS-NET, 该方法具有简洁高效的特点, 并且不需要依赖外部数据.

(2) 根据商品名称精简的特点, 提出了使用基于门控循环单元的神经网络来解决自注意力机制无法直接采集商品时序信息的问题, 并以此取代了自注意机制网络中的三元前馈网络, 以较小的代价换取了模型整体性能的提升.

(3) 在 LESD4EC 数据集[2]的基础上, 生成了商品名称精简数据集LESD4EC_L和 LESD4EC_S, 并以此为基础进行了冗余商品名称精简的验证. 实验结果表明, 本文提出的 ERS-NET 网络模型优于已提出的商品名称精简的基准模型.

1　相关工作

对于冗长商品名称精简, 主要基于自然语言处理中的文本摘要生成及关键词提取任务的研究. 研究者们希望在原始文档中生成包含文档关键内容的摘要, 减少信息观察者的信息观测量, 提升文档信息检索效率[3]. 传统的提取式方法主要依赖于对文本特征的选择. Rose 等人提出了基于词频共现特征的文本关键词生成方法 RAKE (Rapid Automatic Keyword Extraction)算法[4]. Mihalcea 等人提出使用图结构重新组织文本特征, 利用 PageRank 算法来提取摘要信息, 并命名为 TextRank[5]. Zhao 等在TextRank 方法的基础上, 考虑文本的主题及上下文信息特征, 提出了 TopicPhraseRank 模型[6], 用于 Twitter 文本的摘要信息归纳. 无论是基于图结构特征还是词频特征的方法, 在构建全局关系矩阵时都面临着计算空间开销过大的问题, 所以此类方法并不能很好地应对大规模文本信息的摘要生成任务. 随着神经网络技术在自然语言处理领域不断取得的显著成就, 研究者们根据短文本摘要任务的特性, 采用合适结构的网络进行文本摘要生成任务的研究工作, 此类方法通常采用带有长短期记忆网络[7](Long Short-Term Memory, LSTM)或门控循环单元[8](Gated Recurrent Units, GRU)的神经网络模型, 配合注意力机制[9-10], 不需要过多的特征选择, 直接通过学习文本的词嵌入特征进行摘要生成. Nallapati 等提出了使用基于 Seq2Seq(Sequence to Sequence) 模型[11]的神经网络进行文本摘要生成, 将其命名为 TextSum 模型[12]; 文献 [13]在此基础上, 精简了该模型的网络结构, 提出了基于概率预测的 SummaRuNNer 网络模型来预测文本摘要信息. See 等对基于 Seq2Seq 的摘要生成模型进行了改进[14], 引入指针 (Pointer Networks) 机制[15]进行文本摘要的生成. 冗长商品名称精简均是基于以上工作进一步展开的.

现有的研究工作主要分为抽象式方法和提取式方法. 抽象式方法不局限于原文本中的词汇, 通过获取原文本中的语义特征, 从词汇表中生成语义连贯的摘要. Zhang 等提出了基于GAN(Generative Adversarial Networks)网络[16]思想的 MM-GAN 模型[17], 对商品标题的信息进行压缩, 该方法使用 Seq2Seq 模型作为 GAN 的生成器, 抽象生成商品短标题, 并使用另外一个基于 LSTM 的 RNN(Recurrent Neural Networks)对生成的信息质量进行判别. 提取式方法是从原文本中提取目标词汇, 重新组合成文本摘要. Wang 等提出使用查询日志对商品信息进行外部语义增强[18], 进一步对商品标题做短标题生成. Gong 等在现有工作的基础上使用外部词频特征(Term Frequency–Inverse Document Frequency, TF-IDF)以及命名实体特征(Named Entity Recognition, NER)来对商品信息进行特征增强[2], 提出了基于 LSTM 的神经网络模型——FE-NET 模型, 以此来处理商品短标题生成, 并提供了基于淘宝网站[1]的商品短标题数据集 LESD4EC. 本文工作在提取式商品短标题生成的基础上, 弱化问题本身对外

[1] https://www.taobao.com/

部特征数据的依赖, 进一步对商品名称精简技术展开研究.

2 基于自注意力的商品名称精简方法

2.1 问题定义

本文将冗余商品名称精简转化为序列二分类问题. 给定商品名称序列 $T=\{t_1, t_2, \cdots, t_n\}$, 以及商品 NER 属性$E=\{e_1, e_2, \cdots, e_n\}$, 其中 t 代表商品名称中的单词, e 代表商品名称单词 NER 标记, n 代表商品名称长度. 使用$Y=\{y_1, y_2, \cdots, y_n\}$表示商品名称精简序列标记, 对应生成的商品精简名称 S, 其计算公式为

$$S = T \cap Y, \quad y_i \in \{0,1\} \wedge |S| \leqslant \gamma, \tag{1}$$

其中, γ 为精简词汇上限个数, 且$\gamma \leqslant n$.

冗长商品名称精简的目标在于给定名称序列 T 以及商品 NER 属性 E, 使其精简词汇预测结果 Y' 与真实的标记结果 Y 间的对数似然损失最低, 目标函数定义为

$$L = \min \sum_i^n \varphi\left(P(y_i'|t_i, \mathrm{e}_i), P(y_i|t_i, \mathrm{e}_i)\right), \tag{2}$$

其中φ用于计算对数似然损失.

2.2 ERS-NET 模型

相对于 RNN 和 CNN(Convolutional Neural Networks)网络, 自注意力机制网络具有良好的并行性, 并且可以捕捉句子全局依赖关系[1], 不足之处在于浅层自注意力机制网络无法快速获取文本的时序关系. 已提出的自注意力机制网络使用一个前馈神经网络, 将输入的文本信息映射成为三元关系矩阵, 查询矩阵 $Q \in \mathbf{R}^{n\times d}$, 键值对矩阵 $K \in \mathbf{R}^{n\times d}$ 和 $V \in \mathbf{R}^{n\times d}$, n 代表输入序列长度, d 代表网络维度. 然后通过公式

$$\mathrm{ATT} = \mathrm{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{3}$$

计算缩放点积注意力ATT得分.

本文采用基于自注意力机制的网络来解决冗长商品名称精简, 所提出的网络模型整体架构如图 2 所示.

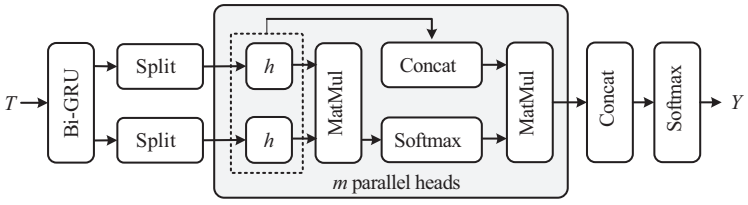

图 2 基于自注意力机制的冗长名称精简模型

Fig. 2 Overview of the ERS-NET model

首先通过一个词嵌入网络层, 将输入商品名称序列 T 映射为词嵌入表示 T'. 为改善自注意力网络对于商品名称时序关系的捕捉能力, 本文提出使用双向 GRU 网络取代自注意力网络中的三元前馈网络, 将双向 GRU 网络的输出结果分别封装成查询矩阵和键矩阵, 值矩阵由双向 GRU 网络的输出结果拼接得到, 然后再计算点积注意力得分, 计算公式分别为

$$\overrightarrow{h} = \overrightarrow{\mathrm{GRU}}\left(T'\right), \tag{4}$$

$$\overleftarrow{h} = \overleftarrow{\mathrm{GRU}}\left(T'\right). \tag{5}$$

运用 GRU 网络的时序捕捉能力, 在得到双向 GRU 网络的隐层输出表示$\overrightarrow{h}$ 与$\overleftarrow{h}$ 后, 通过相反方向的隐层表示计算每个商品词的全局依赖关系, 再使用 Softmax 方法进行非线性激活, 得到商品标题内部的注意力分布矩阵, 最后将的 GRU 隐层表示$[\overrightarrow{h} : \overleftarrow{h}]$拼接并与注意力分布矩阵做乘运算, 得到目标点积注意力ATT, 所对应的公式分别为

$$\mathrm{dot} = \left(\frac{\overrightarrow{h} \times \overleftarrow{h}^{\mathrm{T}}}{\sqrt{d_h}}\right), \tag{6}$$

$$\mathrm{ATT} = \frac{\exp\left(\mathrm{dot}_i\right)}{\sum_i^n \exp\left(\mathrm{dot}_i\right)} \cdot \left[\overrightarrow{h_i} : \overleftarrow{h_i}\right]. \tag{7}$$

本文采用 Vaswani 等提出的多头并行方案[1], 将所得到 GRU 隐层输出结果通过 m 个并行通道进行点积注意力计算. 每一个通道的点积注意力计算方式为

$$\mathrm{Head}_i = \mathrm{ATT}\left(\overrightarrow{h}W_i^Q, \overleftarrow{h}W_i^K\right), \tag{8}$$

其中, W 代表网络的训练参数矩阵, 对于每个并行通道, $W \in \mathbf{R}^{n\times(d/m)}$.

利用上述过程得到的多头矩阵拼接成为一个完整的矩阵, 通过全连接网络层实现目标矩阵的一个线性映射, 并使用 Softmax 分类器进行商品词标记预测, 最后使用最小交叉熵对其预测结果进行损失计算, 所对应的公式分别为

$$H = [\mathrm{Head}_i : \cdots : \mathrm{Head}_h]W^o + b^o, \tag{9}$$

$$Y = \mathrm{Softmax}\left([H : E]\right), \tag{10}$$

其中, E 代表商品名称中的 NER 属性, W 和 b 是神经网络的训练参数. ERS-NET模型的算法流程如算法 1 所示.

算法 1　ERS-NET商品名称精简算法

输入: 训练集商品名称 T_train, 训练集商品标记 Y, 测试集商品名称 T_test, 模型迭代次数 Steps(st), 自注意计算头数 m

输出: 测试集精简商品短标题 S

```
train:
    for all 网络参数 do
        使用截断正态分布随机初始化参数
    end for
    /*读取数据*/
    T' ← 词嵌入网络(T_train)
    for st ← 0 to steps do
        for t in T'
            h_t ← BiGRU(t,h_{t−1})
        end for
        for i ← 0 to m–1 do
            使用公式(6)计算商品词语义关系
            使用公式(8)计算单头注意力得分
```

```
        end for
        H ← concat (ATT)
        Y′ ← Softmax(H)
        L ← −∑ ylogy′, y ∈ Y 且 y′ ∈ Y′
        if st == 0 or L < minL
            model.save()
            minL ← L
        使用 Adam 方法最小化交叉熵损失函数 L 来更新网络参数
    end for
test:
    model.load()
    Y_ predict ← model(T_test)
    S ← Y_ predict∩T_test
```

2.3 算法复杂度分析

在冗长商品名称精简中, 本文使用计算复杂度为 $O(n \cdot d)$、序列操作为 $O(n)$ 的 GRU 网络, 对原有的自注意力机制网络进行数据的序列增强, 并取代生成查询矩阵及键–值矩阵的前馈神经网络, 计算复杂度为 $O(n \cdot d)$, 其中 n 为序列长度, d为计算维度. 为了保证 GRU 网络层不对整体的网络模型增加额外的计算量, 需要对 GRU 网络的计算规模进行约束. 在 ERS-NET 网络模型中, 多头点积注意力网络的值矩阵是由双向 GRU 网络层的输出结果拼接得到, 为保证自注意力机制网络的计算一致性, 需要将 GRU 网络的输出维度降低为原前馈网络的1/2, 这样可以在一定程度上减少模型的计算开销. 此外, 还可通过控制 GRU 网络的输入计算维度进一步对模型进行优化. GRU 网络的输入信息由词嵌入层提供, 也就是说, 减小 GRU 网络的输入计算维度需同步减小词嵌入层的计算维度. 这种调整的优点是可以同时减少词嵌入层和 GRU 网络层的计算开销. 实验证明, 通过以上调整策略可以控制 ERS-NET 网络模型的整体计算开销.

3 实　　验

在实验过程中, 本文分别随机从 LESD4EC_L 和 LESD4EC_S 数据集中抽取 500 000 条商品数据作为训练数据, 50 000 条作为测试数据. 商品名称被映射到 300 维度的向量空间中进行词嵌入操作, 并使用截断正太分布 U(–0.01, 0.01)对其进行初始化, 词嵌入内容将在训练过程中进行优化. 整个模型的隐层维度设置为 512 维, 使用 8 个头的并行通道进行点积注意力计算, 使用学习率为 0.001 的 Adam[19] 优化方法对网络参数进行更新.

3.1 数据集生成

本文所使用的原始数据集 LESD4EC 包含 6 481 623 条数据, 并提供了商品名称(title)、商品命名实体标注信息(NER)、商品摘要数据(Y). 由于标注工作者背景、所具备的知识等因素, 使得标注结果存在一定的差异性, 即同一商品标题可能存在多个不同的标记, 导致了标记结果的不一致. 并且随着商品名称的不断更新, 一些原始标记已不存在于当前商品名称之中. 因此本文对原始 LESD4EC 数据集进行了分析, 合并了相同商品标记数据, 生成了长文本商品标记数据集 LESD4EC_L, 并在数据集 LESD4EC_L 的基础上重新统计了商品标注中标注频率, 即依据奥卡姆剃刀原则, 剔除标注频率低的商品词汇, 最终生成了短文本商品数据集 LESD4EC_S. 考虑到商品信息的实用性特点, 在剔除低频词之前, 会考虑原始数据集

LESD4EC 提供的 NER 属性以及原始商品词的位置特征, 例如相对于商品 NER 属性“普通词”, 本文更倾向于“品牌”、“品类”这类更具有商品特性的词汇, 并且认为商品原始数据中位置靠前的词汇更具有代表性. 具体生成过程如算法 2 所示. 最终生成了分别包含2 031 353 条数据的长文本商品名称精简数据集 LESD4EC_L 和短文本商品名称精简数据集 LESD4EC_S, 商品名称合并样例如表 1 所示.

算法 2　商品名称精简数据集合并算法

输入: 商品数据集 LESD4EC, 关键词上限个数 γ

输出: 商品名称精简数据集 LESD4EC_L, LESD4EC_S

```
for title in LESD4EC
    repeat
        将与 title 相同 ID 的商品标记 Y 合并;
        更新至LESD4EC_L;
    until 无可更新数据;
    根据 NER 信息为 title 标记分配权重, 根据权重计算标记得分 Score;
    根据 Score 得分, 在 title 内部进行词汇排序, 并剔除 Score 为 0 的词汇;
    if 有效标记词汇数 < γ
        将有效标记词汇更新到 LESD4EC_S;
    else
        将前γ 个标记词汇更新到 LESD4EC_S;
end for
```

表 1　商品数据集合并样例析

Tab. 1　Example of product title datasets union

商品名称	标注者 1	标注者 2	标注者 3	标注者 4	标注者 5
谭木匠, 新品, 礼盒, 小, 可爱, 木梳, 年轻款, 送, 孩子, 女生, 儿童节, 礼品	小, 可爱, 木梳, 礼盒	谭 木 匠, 小, 可爱, 木梳	谭 木 匠, 可 爱, 木梳	谭木匠, 可爱, 木梳	卡通, 趣味, 小, 可 爱, 木梳
商品名称	标注者 6	标注者 7	标注者 8	LESD4EC_L	LESD4EC_S
谭木匠, 新品, 礼盒, 小, 可爱, 木梳, 年轻款, 送, 孩子, 女生, 儿童节, 礼品	礼盒, 可爱, 木梳	礼盒, 小, 可爱, 木梳	谭 木 匠, 可 爱, 木梳	谭木匠, 礼盒, 小, 可爱, 木梳	谭木匠, 礼盒, 木梳

3.2　实验基准模型

在对比实验中, 本文选取了编码–解码模型 Seq2Seq[11]、自注意力机制模型 Self_ATT 以及基于多特征输入的 FE_NET 模型[2]作为基准测试模型. Seq2Seq 是一种广泛应用于文本解析的深度学习模型. Self_ATT 源于 Transformer 模型[1], 应用于机器翻译. 为了适用于商品名称精简, 本文从原有的 Transformer 模型中提取了完整的自注意力机制网络结构, 并构造了自注意力网络 Self_ATT 模型. FE_NET 模型是专门用于商品短标题生成的模型.

3.3　实验评估标准

本文采用 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)指标[20]来对模型生成的商品精简标题进行评估, 该指标目前已被广泛地应用于 DUC(Document Understanding Conference)摘要评测[1]. 为了更全面地衡量所生成商品短标题的质量, 同文献 [3] 一样, 本文也在

[1] https://duc.nist.gov/

ROUGE 基础上引入了精确率(Precision)、召回率(Recall)及 F1 分数(F1-score)来进一步分析模型性能. 联合评估计算的公式分别为

$$\text{ROUGE\_}P = \frac{|S \cap S_h|}{|S|}, \tag{11}$$

$$\text{ROUGE\_}R = \frac{|S \cap S_h|}{|S_h|}, \tag{12}$$

$$\text{ROUGE\_F1} = 2 \times \frac{\text{ROUGE\_}P \times \text{ROUGE\_}R}{\text{ROUGE\_}P + \text{ROUGE\_}R}, \tag{13}$$

其中, S_h 表示人工标注的商品标题, S 表示 ERS-NET 预测的商品标题.

3.4 实验结果及分析

本文分别在长文本商品标记数据集 LESD4EC_L 和短文本商品标记数据集 LESD4EC_S 上对 ERS-NET 模型及其他基准模型进行了测试验证, 实验结果如表 2 所示.

表 2 在不同数据集上的商品名称精简实验结果

Tab. 2 Results on different datasets

模型	LESD4EC_L			LESD4EC_S		
	ROUGE_P/%	ROUGE_R/%	ROUGE_F1/%	ROUGE_P/%	ROUGE_R/%	ROUGE_F1/%
Seq2Seq	60.22	72.71	65.88	76.40	78.37	77.37
Self-ATT	73.36	74.10	73.73	79.18	82.17	80.65
FE-NET	73.02	74.84	73.92	81.73	85.30	83.48
ERS-NET	**74.86**	**76.62**	**75.55**	**82.94**	**86.85**	**84.85**

从表 2 中可以看出, 在商品名称精简即商品短标题生成中, Seq2Seq 模型明显逊色于其他模型. 相比于其他基准模型, Seq2Seq 模型由编码和解码两部分组成, 两段式的数据交互不可避免地会引起信息损失, 尽管在实验的过程中, 使用了 Luong 注意力[10]来对 Seq2Seq 模型增强, 依然无法改善该模型的 ROUGE_F1 . FE-NET 模型作为专有的商品短标题生成模型, 依托于丰富的外部商品特征信息, 在实验中取得了较好的成绩. Self-ATT 可以很好地捕捉句子的全局依赖特性, 在冗余商品名称精简中, 分别在 LESD4EC_L 数据集和 LESD4EC_S 数据集上取得了 ROUGE_F1 为 73.73% 和 80.65%. 本文在此基础上, 使用小规模 GRU 网络对其进行语义时序增强, 最终得到的 ERS-NET 模型, 相对于 FE-NET 模型, 它弱化了模型对全局特征的依赖, 仅仅考虑了数据本身的属性特征, 其 ROUGE_P、ROUGE_R、ROUGE_F1 在不同数据集上均取得了最优, ROUGE_F1 分别取得了 1.63 个百分点和 1.37 个百分点的提升.

为了进一步调整 ERS-NET 模型对商品名称的精简能力, 本文使用 Sigmoid 激活层取代了原有的 Softmax 分类层, 通过调节分类阈值 τ 来确定最终的分类结果, 其 ROUGE_P、ROUGE_R、ROUGE_F1 的变化如图 3 所示. 在商品名称精简时, 当阈值τ=0.4 时可以取得 ROUGE_F1 的最优, 分别达到了 76.34% 和 85.14%.

在商品名称精简中, 虽然本文使用了较小规模的基于 RNN 结构的 GRU 网络对自注意力机制网络进行时序增强, 但依然要防止较高复杂度的 GRU 网络产生的额外计算开销, 故需要对 GRU 的计算维度 d 进行约束. 在实际的实验中, 本文约束 GRU 的计算维度 $d = 50$, 并且将 ERS-NET 网络部署到两种规格的英伟达图形处理器(Graphics Processing Unit, GPU)上进行训练测试, 分别是 Nvidia Quadro P600[1]、Nvidia GTX 1060[2]. 表 3 展示了每万条数据在两种规格 GPU 上训练所花费的时间, 单位为秒(s). 从实验结果中可以看出, 通过约束 GRU 网络的计

[1] https://www.nvidia.cn/content/dam/en-zz/Solutions/design-visualization/ documents/Quadro-P600
[2] https://www.nvidia.cn/geforce/products/10series/geforce-gtx-1060/

算维度, 可以有效地防止 GRU 网络产生额外的计算开销, 从而保证了 ERS-NET 模型整体的计算效率.

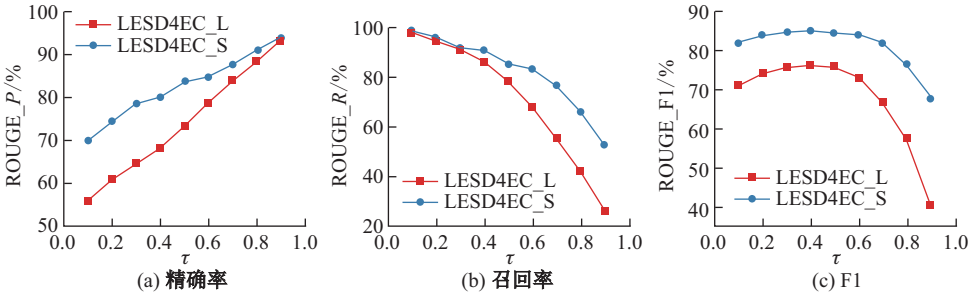

图 3　不同阈值下商品名称精简任务的精确率、召回率、F1 变化情况

Fig. 3　ROUGE performance on product title compression with different thresholds

表 3　不同 GPU 下每万条数据执行时间

Tab. 3　Computational time per ten thousand data items on GPU

GPUs	花费时间/s	
	Self-ATT	ERS-NET
Quadro P600	11.179	10.365
GTX 1060	3.957	3.831

4　案　　例

本文选取同时出现在两个数据集中的两个商品, 通过具体案例进一步分析 ERS-NET 模型在商品名称精简上的优越性. 如表 4 所示, 其中后缀 L 和 S 分别代表在数据集LESD4EC_L 和 LESD4EC_S 上进行的实验, LESD4EC_S 数据集在标签长度上做了上限为 3 的约束. 可以看出, 案例 1 中, 模型通过在不同数据集上进行学习, 可以很好地根据各自数据集的标签特性对原商品名称进行精简. 若以案例中的标签为评判基准, ERS-NET 模型在案例 2 上的预测结果逊色于案例 1; 但从精简的结果来看, 在 LESD4EC_L数据集上, 其预测的商品精简标题“*捷波朗运动心率无线蓝牙防水耳机*”并没有影响大众对商品的理解, 并且在 LESD4EC_S 上的预测结果“*jabra 蓝牙耳机*”要比原标签“*jabra 捷波朗耳机*”具有更好的可读性. 从具体的实验案例可以看出, 在真实的应用场景下, ERS-NET 模型依然具有良好的商品名称精简能力.

表 4　两个真实应用场景下的冗余商品名称精简案例

Tab. 4　Case study with real application scenarios

案例 1	标签_L	精简预测_L	标签_S	精简预测_S
geras, 童装, 男童, 圆领, 套头, 卫衣, 套装, 春秋, 新品, 儿童, 运动, 纯棉, 两件套	geras, 男童, 圆领, 套头, 卫衣, 套装	geras, 男童, 圆领, 卫衣, 套装	geras, 卫衣	geras, 卫衣
案例 2	标签_L	精简预测_L	标签_S	精简预测_S
jabra, 捷波朗, elite, 运动, 臻, 跃, 心率, 无线, 蓝牙, 跑步, 防水, 耳机, 新品	jabra, 捷波朗, elite, 运动, 心率, 无线, 耳机	捷波朗, 运动, 心率, 无线, 蓝牙, 防水, 耳机	jabra, 捷波朗, 耳机	jabra, 蓝牙, 耳机

5 结 论

冗长商品名称精简旨在精炼商品命名信息, 为构建跨平台商品知识体系、多源商品数据融合提供必要的数据支持. 本文在商品数据集 LESD4EC 的基础上构造了商品精简标记数据集 LESD4EC_L 和 LESD4EC_S, 并在此基础上针对商品名称精简进行了实验分析, 提出了基于自注意力机制商品名称精简模型 ERS-NET, 该模型使用较小规模的 GRU 神经单元对商品名称进行语义时序增加. 实验结果表明, 本文提出的 ERS-NET 模型在两种规格的商品名称精简中相对于已提出的商品名称精简模型均取得了最优. 在之后的工作中, 会以本文提出的 ERS-NET 模型为基础, 继续研究在多源商品数据情况下商品实体的解析.

[参 考 文 献]

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[2] GONG Y, LUO X S, ZHU K Q, et al. Automatic generation of chinese short product titles for mobile display[J]. arXiv preprint arXiv:1803.11359, 2018.

[3] LIU Z Y, HUANG W Y, ZHENG Y B, et al. Automatic keyphrase extraction via topic decomposition[C]//Proceedings of the 2010 conference on empirical methods in natural language processing. Association for Computational Linguistics, 2010: 366-376.

[4] ROSE S, ENGEL D, CRAMER N, et al. Automatic keyword extraction from individual documents[M]//Text mining: Applications and theory. Hoboken: A John Wiley and Sons, Ltd., 2010: 1-20.

[5] MIHALCEA R, TARAU P. Textrank: Bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[6] ZHAO W X , JIANG J , HE J , et al. Topical keyphrase extraction from Twitter[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2011: 379–388.

[7] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[8] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1724-1734.

[9] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[10] LUONG T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1412-1421.

[11] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]//Advances in Neural Information Processing Systems. 2014: 3104-3112.

[12] NALLAPATI R, ZHOU B W, DOS SANTOS C, et al. Abstractive text summarization using sequence-to-sequence RNNs and beyond[C]//Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. 2016: 280-290.

[13] NALLAPATI R, ZHAI F F, ZHOU B W. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents[C]// Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-17). 2017:3075-3081.

[14] SEE A, LIU P J, MANNING C D. Get to the point: Summarization with pointer-generator networks[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1073-1083.

[15] VINYALS O, FORTUNATO M, JAITLY N. Pointer networks[C]//Advances in Neural Information Processing Systems 28(NIPS 2015). 2015: 2692-2700.

[16] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems 27(NIPS 2014). 2014: 2672-2680.

[17] ZHANG J, ZOU P, LI Z, et al. Multi-modal generative adversarial network for short product title generation in mobile e-commerce[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers). 2019: 64-72.

(下转第 167 页)

5 结　论

随着 NoSQL 型数据库在各行各业应用的增长, 基于 NoSQL 型数据库的二级索引也将迎来越来越多的应用. 本文基于 LevelDB 和 R-tree, 设计了支持 LSM 结构的空间二维坐标数据索引. 通过实验证明, 本文所设计的二级索引具备可用性, 对二维数据常用的 10-NN 最近邻查询有较好的支持. 由于时间有限, 本文并未对引入该索引后 LevelDB 本身的插入查询性能变化有较多的探讨, 因此该研究后续仍有极大优化的空间.

[参 考 文 献]

[1] Google Inc. LevelDB [EB/OL]. [2019-06-20]. https://githbu.com/google/leveldb.

[2] BECKMANN N, KRIEGEL H P, SCHNEIDER R, et al. The R*-tree: An efficient and robust access method for points and rectangles [C]// ACM Sigmod Record. ACM, 1990, 19(2): 322-331.

[3] LUO C, CAREY M J. LSM-based storage techniques: A survey [J/OL]. arXiv preprint, arXiv: 1812.07527, 2018.

[4] WU L, LIN W, XIAO X, et al. LSII: An indexing structure for exact real-time search on microblogs [C]// 2013 IEEE 29th International Conference on Data Engineering (ICDE). IEEE, 2013: 482-493.

[5] KHODAEI A, SHAHABI C, LI C, et al. Hybrid indexing and seamless ranking of spatial and textual features of web documents [C]// International Conference on Database and Expert Systems Applications. Berlin: Springer, 2010: 450-466.

[6] QADER M A, CHENG S, HRISTIDIS V. A comparative study of secondary indexing techniques in LSM-based NoSQL databases [C]// Proceedings of the 2018 International Conference on Management of Data. ACM, 2018: 551-566.

[7] TAN W, TATA S, TANG Y, et al. Diff-Index: Differentiated index in distributed log-structured data stores [C]// EDBT. 2014: 700-711.

[8] DSILVA J V, RUIZCARRILLO R, YU C, et al. Secondary indexing techniques for key-value stores: Two rings to rule them all [C] // Proceedings of the 20th International Conference on Extending Database Technology and 20th International Conference on Database Theory 2017. 2017: 21-24.

(责任编辑: 林　磊)

(上接第 122 页)

[18] WANG J G, TIAN J F, QIU L, et al. A multi-task learning approach for improving product title compression with user search log data[C]//32nd AAAI Conference on Artificial Intelligence. 2018: 451-458.

[19] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv: 1412.6980, 2014.

[20] LIN C Y, HOVY E. Automatic evaluation of summaries using n-gram co-occurrence statistics[C]//Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. 2003: 150-157.

(责任编辑: 李　艺)