高技术通讯 2023 年 第33 卷 第8 期:836-848

doi:10.3772/j.issn.1002-0470.2023.08.006

SOM-NCSCM +:抽取式神经网络中文标题生成方法研究①

资康莉②*** 王 石③* 曹存根*

(*中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

(**中国科学院大学 北京 100049)

摘 要 标题生成作为文本摘要任务的一个分支,能够帮助人们高效获取信息。本文针对中文标题生成任务面临的大规模、高质量中文标注数据缺乏的问题,利用标题往往可由原文中的词语来构成的特点,从将无监督学习模型与有监督的序列标注模型结合的角度出发,提出了融合聚类模型和主题模型的抽取式深度神经网络中文标题生成方法和模型。在缺乏人工分类标注信息的中文新闻数据集上,该模型可利用聚类和主题模型自动挖掘数据内部潜在的特征信息,获得不同的数据簇及各簇内的主题词来辅助中文新闻标题生成,使模型在具有潜在主题类别特征的、标题质量参差的中文新闻数据集上都具有较好的适用性。本文提出的中文标题生成模型在互联网上公开的中文新闻标题数据集上的实验结果也表明其在微观 *F*1、BLEU、ROUGE、压缩率等评价指标上都取得了较基准模型更好的效果。

关键词 中文标题生成;神经网络模型;主题模型;聚类模型;序列标注

0 引 言

随着海量文本数据在新闻网站、社交网络等网络空间极速涌现,文本摘要作为一种关键技术,广泛用于对海量内容进行提炼总结,方便更多用户快速浏览和了解大量文档。其中,标题生成作为文本摘要的一个重要应用场景,其主要任务是针对给定的篇章或者短文本,生成能够概括或评论其主要内容的一段或者一句话作为标题。

标题生成技术已被应用于搜索结果展示、文章摘要生成、新闻标题生成等众多领域。根据需要处理的数据篇幅的不同,可将其分为单文档标题生成[1]和多文档标题生成[2];根据实现方式的不同可分为抽取式标题生成[3]和生成式标题生成[4];而根据使用的技术手段,可分为传统标题生成方法[5]和基于深度学习的标题生成方法[6]。

本文主要关注中文领域基于神经网络的抽取式标题生成方法,该类方法目前仍面临一个重大挑战:缺乏大规模、高质量的中文标注数据。而产生该挑战的原因有:(1)标题生成研究工作多在公开的外文数据集上开展,中文领域的部分研究工作未公开完整数据集,使得后续研究者无法在其已有工作基础上继续探索并进行研究成果间的比较。(2)现有中文标题标注数据多收集于各类新闻网站、社交网站,数据繁杂,缺乏统一的分类体系,或数据集未提供原分类信息,并且原网站中标题的质量难以保证,有时甚至会使用夸大或缺乏与原内容相关的词句来构造标题。

因此,针对上述挑战和问题,考虑到标题往往可直接从原文中抽取词汇或者句子来构造这一特点,本文采用了基于抽取式的标题生成技术,提出将基于无监督学习的聚类模型和主题模型融入到基于有监督学习的深度神经网络模型中的方法,使得整个

① 国家重点研发计划(2022YFC3302300)和国家 242 信息安全计划(2022A056)资助项目。

② 女,1994 年生,博士生;研究方向:自然语言处理,智能问答系统;E-mail:zikangli19b@ict.ac.cn。

③ 通信作者,E-mail:wangshi@ict.ac.cn。

(收稿日期:2022-07-29)

模型在具有潜在主题类别特征的、新闻标题质量参差不齐的数据上也能取得较好效果。本文主要的贡献点如下。

(1) 将抽取式新闻标题生成问题转化为序列标注问题,并通过在注意力机制中融入聚类特征和主题词信息等多种特征,增强对新闻内容的上下文表示。

(2) 采用了基于自组织映射(self-organizing map,SOM)的聚类模型[7]和隐含狄利克雷分布(latent Dirichlet allocation,LDA)主题模型[8],能将表达相同或相似主题的新闻内容进行聚类,并进一步从各数据簇中自动挖掘相关的主题词集合。

(3) 本文从现有公开的中文新闻数据集中抽取了部分数据,并进行了分词、错别字纠错、词性标注、命名实体信息标注等预处理,再通过人工与半自动核对等策略,得到了一个可用于抽取式中文标题生成的数据集。最后,在该数据集上进行的实验表明,本文设计的模型在微观 $F1$、BLEU、ROUGE、压缩率等评价指标上都取得了较基准模型更好的效果。

1 相关工作

1.1 标题生成任务

标题生成任务作为文本摘要的一个分支,要求生成精炼且优质的标题,使得标题包含不会过分夸大实际内容的具体事实,能够对原内容信息进行有效地传递,并能吸引更多的用户,提高用户的阅读效率,提升阅读体验[9]。因此,在生成标题时,该任务要求去掉原文中的冗杂信息,只保留原文中涉及的关键信息,得到长度短于原文的、更加简洁的、可由原文中部分句子组成的集合或者仅为原文中关键信息组合成的标题句。标题生成技术有着巨大的应用价值和广泛的应用场景,例如邮件内容的自动生成、搜索结果展示、文章摘要生成、新闻标题生成、移动设备信息推送、社区问答等。

根据标题生成所需处理的数据篇幅可以将其分为单文档标题生成和多文档标题生成。其中,单文档标题生成关注的是对短文本或者单文档进行标题生成[1],多文档标题生成则是从一组主题或者内容相关的文档中总结生成标题[10]。而根据标题生成方法的实现方式或产生输出结果的类型,可以分为抽取式标题生成[3]和生成式标题生成[4]。其中,抽取式标题生成是从原文档内容中抽取关键词或关键句进行组合来生成标题,也即需要判断原文档中各个词语、语句的重要程度,使得最终生成的标题中的词或句均来自原文档;而生成式标题生成则是在充分理解原文档内容的基础上,允许模型使用除原文档内容以外的新词语、新语句来组成能够概括原文档内容的标题。

基于生成式以及基于抽取式的标题生成方法各有其优缺点。首先,这 2 类方法都要求输出的标题能够尽可能全面地包含原文档内容的关键信息。基于生成式的标题生成方法相比于抽取式而言在用词方面更加灵活,能够生成多样化的标题表述,来满足许多应用领域对于多样化、个性化的信息展示的需求。而基于抽取式的标题生成方法是抽取原文中的一部分内容(词或句)作为输出,它产生的标题的表述会受限于原文。但是,这 2 种方法及其对应的模型在实现时都会面临从互联网中获取到的数据及其原标题质量难以保证的问题。在这种现状下,使用基于生成式的标题生成方法难免会受到数据集质量的约束,而采用基于抽取式的标题生成方法,虽然损失了一定的泛化能力,但是因为其并不能“自主”生成不存在于原内容的词汇或语句,使得其在面对“噪声”数据时能够具有较好的鲁棒性,甚至能够用于发现相关数据中的“噪声”或“异常”。其次,近年来快速发展的深度神经网络技术因其强大的表征能力,给予了这 2 类方法更多的可能性,使得标题生成的效果被不断提升。但是,尤其在面对长文本或者多文档标题生成时,基于生成式的标题生成方法会因缺少对关键信息的控制与定位,而需要额外控制最终模型输出的标题与原文的相关性(例如:保持原内容的主题信息等)[6],避免出现无法处理未登录词、标题与原内容关键信息关联度不高、词语重复生成等问题。而基于抽取式的标题生成方法,虽然能更好地控制与原文档内容的相关性,但是也需要设计较好的衡量原内容中关键词或句的重要程度的方法,避免抽取得到的标题中具有较多冗余信息。

此外,根据使用的技术手段来划分,传统的标题生成方法多基于统计概率与人工特征工程,且多为抽取式标题生成,通过计算得到已有数据集中的特征信息(例如句子长度、句子位置、词序、词频、逆文档频率、最大公共子串、关键词表、类簇信息等),来判断并抽取原文中具有较多信息量的词语和句子组成标题[5,11]。而基于神经网络的标题生成技术多采用"端到端"的神经网络标题生成框架[12],既可以进行抽取式标题生成(将标题生成任务转化为序列标注任务或者对句法依存树采取剪枝的任务[13]),也可以直接生成多样化表达的标题[14]。并且,基于神经网络的标题生成方法在减少人工特征工程的同时,还能够通过神经网络模型更好地学习到数据中潜在的深层信息(例如使用现有流行的大规模预训练语言模型获取句子的语义表示[15]),这些都在一定程度上解决了因为数据不均衡导致的统计信息计算不正确、获取句子表示受到相关领域数据量的限制以及难以跨领域复用等问题。

1.2 文本聚类与主题模型

在现实互联网中,大多数获取到的数据是缺乏人工标注的分类信息的,或者一些新兴的事物是没有历史类别信息的,而文本聚类是对文本数据进行聚类分析以解决样本分类问题的一种方法。它作为一种无监督机器学习方法,具有一定的灵活性和自动处理能力,可以通过已有数据内部自身的特征,探索性地将相似数据进行归类,来得到数据中潜在的自然分组情况,而不依赖预先定义的类别标记。

因此,文本聚类方法可作为一个独立工具,对数据进行类似预处理的操作,来获得数据的基本分类情况。目前,传统的文本聚类算法有 K-means[16]、BIRCH (balanced iterative reducing and clustering using hierarchies)[17] 以及高斯混合模型(Gaussian mixture model,GMM)[18]等,这些算法和技术已被应用于自动文摘、信息检索、推荐系统等领域中。

而本文将采用一种基于神经网络的聚类方法——基于 SOM 的聚类方法。该聚类方法由 Kohoen[7]提出,对应的网络模型一般只包含输入层和输出层,不包含隐藏层。其中,输入层用于接收高维的输入向量,输出层则由一系列有序节点构成(例如输出层神经元之间的结构为二维网格,它们存在横向连接),输入层与输出层之间通过权重向量连接。

与传统聚类算法以及其他基于神经网络的聚类方法不同的是,SOM 聚类方法不需要预先设置聚类数目,或者仅在传统聚类算法中融入神经网络训练得到的词或句的表示,完全采用神经网络结构,能够直接应用在输入数据的高维词向量上,并能够在输出层根据不同的输入数据激活相应的神经元的同时,将数据进行降维且保留数据的拓扑结构。此外,现有神经网络模型一般采用的都是误差修正学习方式(例如误差反向传播算法)来进行模型的训练和学习,而基于 SOM 的聚类模型采用的学习方式为竞争学习。具体地,在模型训练和学习过程中,各输出神经元会有选择地适应具有潜在类别的输入数据,使得最终输入某一类的数据时,能找到与之距离最短的一个输出层神经元并激活,也即各个输出神经元代表了不同的簇,在输入某一类的数据到模型中时,该类对应的输出神经元会被激活,使得该数据划入该簇中。与此同时,这种竞争学习机制在模型训练过程中,除了会对被激活的"获胜"输出神经元与输入层之间的权重向量进行更新外,还会对设定的邻近区域内的其他输出神经元与对应的输入层之间的权重向量也进行一定程度的权值更新,这使得输出神经元之间能够保持输入层向量的拓扑特征。

另外,因为各相似数据簇中的新闻数据往往会表达相似或者相同主题,而各个主题又是以文本中所有字词为支撑集的概率分布,所以可以在文本聚类结果上进一步获取各数据簇中的主题词信息。本文采用的是被广泛应用的 LDA 主题模型[8]来进一步分析各簇中的文本数据,并获取各簇中与该簇主题关联性高的、有较大出现概率的主题词集合。

具体地,LDA 主题模型作为一种文档生成模型,也是一种无监督学习技术。它采用词袋方法,将每篇文档视为一个词频向量,在生成文档时,认为一篇文档可以有多个主题,每个主题又对应不同的主题词。在采用 LDA 模型进行文档生成的过程中,首先以一定概率选择某个主题,然后在该主题下再以一定概率选择一个词,之后不断重复这个过程,直到

整篇文档生成结束。相应地,在本文中利用 LDA 主题模型从聚类模型得到的各簇新闻内容数据中获取主题词的过程是上述文档生成过程的逆过程,即根据数据集中的新闻内容的文本集合,找到各簇数据的主题以及每一个主题对应的高频词集合。

2 抽取式神经网络中文标题生成方法

2.1 问题定义

如表 1 所示,使用形式化语言对基于抽取式方法的中文新闻标题生成问题进行描述和定义。每一条新闻内容 d 由句子序列 $\{s_1, s_2, \cdots, s_n\}$ 组成,其中 s_i 对应新闻内容中的第 i 条句子。而 s_i 是一条进行了分词的词语序列 $\{w_{i,1}, w_{i,2}, \cdots, w_{i,m_i}\}$,其中 $w_{i,j}$ 对应该文档 d 中第 i 条句子中的第 j 个词语。则对一条新闻内容 d 进行抽取式标题生成,是判断各条句子中哪些词语应该被保留,并最终产生一条包含了 k 个词语的词语序列(即标题) $c = \{c_1, c_2, \cdots, c_k\}$,其中任意一个词语 c_l 均来自新闻内容 d。

表 1 中文新闻标题生成问题的形式化定义

单条新闻-标题数据:
新闻内容 $d = \{s_1, s_2, \cdots, s_n\}$
d 中的句子 $s_i = \{w_{i,1}, w_{i,2}, \cdots, w_{i,m_i}\}$
d 对应的标题 $c = \{c_1, c_2, \cdots, c_k\}$
d 对应的标签序列 $y = \{y_{1,1}, y_{1,2}, \cdots, y_{1,m_1}, y_{2,1}, y_{2,2}, \cdots, y_{2,m_2}, \cdots\}$
包含 N 条数据的新闻-标题数据集:
数据集 $D = \{(d^t, c^t)\}_{t=1}^{N}$
D 对应的 0/1 标签序列集合 $C = \{(d^t, y^t)\}_{t=1}^{N}$

由此,基于抽取式的中文新闻标题生成问题也可转化成一个序列标注问题,即根据新闻内容 d 中各条句子的词序列以及其标题对应的词序列,可以使用一条使用“0/1”作为标记的标签序列 $y = \{y_{1,1}, y_{1,2}, \cdots, y_{1,m_1}, y_{2,1}, y_{2,2}, \cdots, y_{2,m_2}, \cdots\}$ 来表示对各句子中的词语进行的二元操作,其中 $y_{i,m_i} \in \{0,1\}$。当 $y_{i,m_i} = 0$ 时,表示在生成的新闻标题中不保留对应 d 中第 i 句话中的词语 w_{i,m_i};而当 $y_{i,m_i} = 1$ 时,表示保留对应的词语 w_{i,m_i}。该新闻内容 d 的词序列中被标为“1”的词语的总数等于新闻标题 c 的词总数 k。

而对于包含了 N 条新闻数据的数据集来说,将数据集形式化表示为 $D = \{(d^t, c^t)\}_{t=1}^{N}$,对应的标签序列集合表示为 $C = \{(d^t, y^t)\}_{t=1}^{N}$,则本文基于抽取式方法的神经网络中文标题生成模型的训练目标是使用 C 进行模型训练,得到神经网络序列标注模型。之后对于任意一条用于测试的中文新闻数据 d_{test},模型可以预测并输出对应的标签序列 y_{test},再根据标签序列中标“1”的标签可以找到新闻内容中相应位置的、构成新闻标题的具体词汇。

2.2 基于 SOM 的聚类模型

本文采用的 SOM 聚类模型是一个一维前向网络结构,输入层神经元与输出层神经元以及输出层神经元之间都是全连接的结构,具体的模型结构如图 1 左下角所示。

为训练 SOM 模型,本文将各条新闻内容中各个分词对应的词向量 $\boldsymbol{e}^w = (\boldsymbol{e}^{w_{1,1}}, \boldsymbol{e}^{w_{1,2}}, \cdots, \boldsymbol{e}^{w_{1,m_1}}, \boldsymbol{e}^{w_{2,1}}, \boldsymbol{e}^{w_{2,2}}, \cdots, \boldsymbol{e}^{w_{2,m_2}}, \cdots)$ 作为其输入,其中 $\boldsymbol{e}^{w_{i,m_i}} \in R^{d_w}$,$d_w$ 为词向量的维度,$\boldsymbol{e}^{w_{i,m_i}}$ 对应分词 w_{i,m_i} 的词向量。然后,SOM 模型的输出结果为一条输入数据在输出层激活了的神经元所在的坐标。其输出表示为

$$\boldsymbol{z}_s = som(\boldsymbol{e}^w; \theta_s) \tag{1}$$

其中,$som(\cdot)$ 表示 SOM 模型对各条新闻数据的计算过程,θ_s 表示 SOM 模型中的权重参数。整个 SOM 模型在设置的迭代轮次结束后停止训练。则针对输入的单条新闻内容,其对应激活的输出神经元坐标 $\boldsymbol{z}_s$ 可以被转化为索引表示,也就对应聚类结果中该条新闻内容所归属的簇编号。

最后,利用预训练好的 SOM 模型,可以获取并赋予每一条输入的新闻内容所对应的簇编号。

2.3 抽取式神经网络中文标题生成模型

2.3.1 基准模型

本文采用了常用的一种神经网络序列标注框架作为基准模型,它由一个双向的长短期记忆网络与一个条件随机场模型组成,且其输入使用了多种特征信息,包括词向量、命名实体信息、词性信息等。

具体地,对每一条新闻内容 d,双向长短期记忆网络会将其对应的词向量和词汇特征(命名实体特

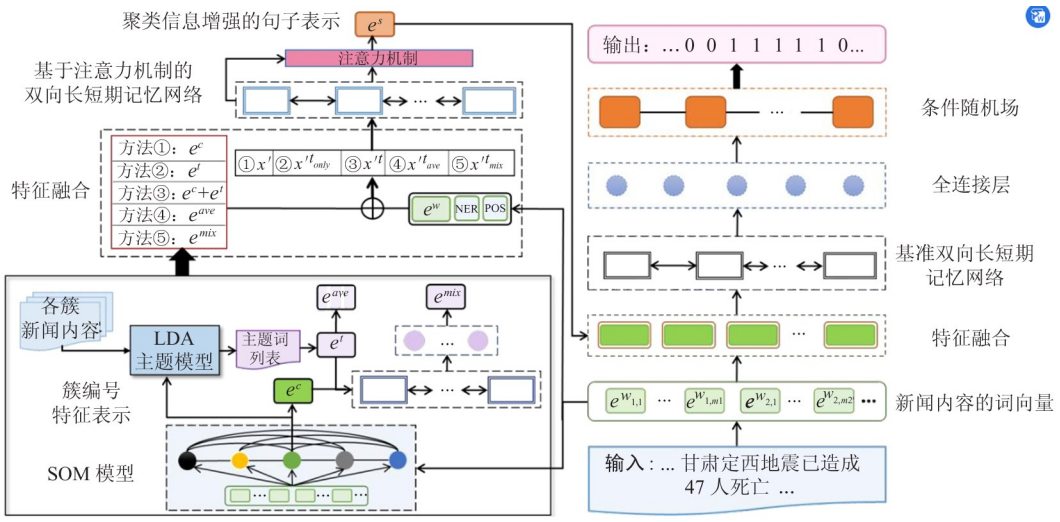

图 1　辅以聚类方法的抽取式神经网络中文标题生成模型的整体框架

征和词性特征）进行联合作为输入 $\boldsymbol{x}=(\boldsymbol{e}_{1,1},\boldsymbol{e}_{1,2},\cdots,\boldsymbol{e}_{1,m_1},\boldsymbol{e}_{2,1},\boldsymbol{e}_{2,2},\cdots,\boldsymbol{e}_{2,m_2},\cdots)$，其中 $\boldsymbol{e}_{i,m_i}\in R^{d_w+d_n+d_p}$，$d_w$、$d_n$、$d_p$ 分别为词向量、命名实体特征向量、词性特征向量的维度。之后，双向长短期记忆网络的输出为一条隐层状态序列 $\boldsymbol{h}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_{T_x})$，其中 T_x 为输入向量 $\boldsymbol{x}$ 的长度，每一个 $\boldsymbol{h}_i$ 是向前和向后的长短期记忆网络结构的输出表示的连接：

$$\begin{aligned}\boldsymbol{h}_i&=[\overrightarrow{\boldsymbol{h}_i};\overleftarrow{\boldsymbol{h}_i}],\\ \overrightarrow{\boldsymbol{h}_i}&=\mathrm{LSTM}(\boldsymbol{e}_{i,m_i},\overrightarrow{\boldsymbol{h}_{i-1}}),\\ \overleftarrow{\boldsymbol{h}_i}&=\mathrm{LSTM}(\boldsymbol{e}_{i,m_i},\overleftarrow{\boldsymbol{h}_{i+1}})\end{aligned}\tag{2}$$

其中，$\overrightarrow{\boldsymbol{h}_i}$ 和 $\overleftarrow{\boldsymbol{h}_i}$ 为 d_h 维的向量。

随后，将双向长短期记忆网络的输出 $\boldsymbol{h}$ 输入到一个全连接层进行一定程度上的数据降维，再将得到的输出输入到条件随机场模型中：

$$\hat{\boldsymbol{h}}=\tanh(\boldsymbol{W}_d\boldsymbol{h}+\boldsymbol{b}_d)\tag{3}$$

其中，$\boldsymbol{W}_d$ 和 $\boldsymbol{b}_d$ 为全连接层的权重和偏置向量。

则根据条件随机场模型计算得到的对应当前新闻内容 d 的一条标签序列 y 的得分为

$$s(x,y)=\sum_{i=1}^{T_x}(\boldsymbol{W}^{\mathrm{T}}_{c(y_{i-1},y_i)}\hat{\boldsymbol{h}}+\boldsymbol{b}_{c(y_{i-1},y_i)})\tag{4}$$

另外，因条件随机场模型的计算会考虑所有可能预测出的标签序列，所以 y 出现的概率最终可以定义为

$$p(y\mid\hat{\boldsymbol{h}};\boldsymbol{W}_c,\boldsymbol{b}_c)=\frac{\prod^{T_x}\boldsymbol{e}^{s(x,y)}}{\sum_{y'\in\gamma(\hat{h})}\prod^{T_x}\boldsymbol{e}^{s(x,y')}}\tag{5}$$

其中，$\boldsymbol{W}_c$ 和 $\boldsymbol{b}_c$ 为模型处理 (y_{i-1},y_i) 标签对时的权重和偏置向量。

在基准神经网络中文标题生成模型的训练过程中，整个模型的目标是使正确的标签序列所对应的对数概率最大化。因而本文采用维特比算法来训练条件随机场模型，并使用得分最高的标签序列 y^* 作为模型预测并输出的最优标签序列结果。

$$\begin{aligned}L(\boldsymbol{W}_c,\boldsymbol{b}_c)&=\sum_{j=1}^{N}\log p(y^j\mid\hat{\boldsymbol{h}}^j;\boldsymbol{W}_c,\boldsymbol{b}_c),\\ y^*&=\mathrm{argmax}_{y'\in\gamma(\hat{h})}p(y\mid\hat{\boldsymbol{h}};\boldsymbol{W}_c,\boldsymbol{b}_c)\end{aligned}\tag{6}$$

2.3.2　融入聚类模型的中文标题生成模型

为更好地探索新闻数据内部的隐含类别特征信息，本文采用了文献[19]设计的一种神经网络框架——辅以聚类的神经网络中文句子压缩模型（SOM-enhanced neural Chinese sentence compression model，SOM-NCSCM），如图 1 中的方法①所示，并将其用于中文新闻标题生成任务。

具体地，为更丰富地表示新闻内容，本文将 2.2 节预训练好的 SOM 模型得到的簇编号特征 $\boldsymbol{e}^c$ 与新闻内容各个分词对应的词向量以及词汇特征（命名实体特征和词性特征）进行连接，得到特征集合 $\boldsymbol{x}'=(\boldsymbol{e}'_{1,1},\boldsymbol{e}'_{1,2},\cdots,\boldsymbol{e}'_{1,m_1},\boldsymbol{e}'_{2,1},\boldsymbol{e}'_{2,2},\cdots,\boldsymbol{e}'_{2,m_2},\cdots)$，其中 $\boldsymbol{e}'_{i,m_i}\in R^{d_w+d_n+d_p+d_c}$，$d_w$、$d_n$、$d_p$ 分别为与基准模型相同的词向量、命名实体特征向量、词性特征向量的维度，而 d_c 为随机初始化的簇编号特征的维度。在

得到特征集合之后,将其作为文献[19]设计和采用的基于注意力机制的双向长短期记忆网络模型[20]的输入,用于将簇编号特征融入对新闻内容的上下文表示中。相应地,针对当前某条新闻内容的输出,其计算方式如下。

$$\begin{aligned} \boldsymbol{e}_t^s &= \sum_{i=1}^{T_x} \alpha_{it} \boldsymbol{h}_i' \\ \alpha_{it} &= \mathrm{softmax}(\mu_{it}) \\ u_{it} &= \vartheta^{\mathrm{T}} \tanh(\boldsymbol{W}_h^{\mathrm{T}} \boldsymbol{h}_i' + \boldsymbol{W}_s^{\mathrm{T}} \boldsymbol{h}_t') \end{aligned} \tag{7}$$

其中,$t \in [1, T_x]$,$\boldsymbol{W}_h$、$\boldsymbol{W}_s$ 和 ϑ 都是模型中可训练的参数,而 $\boldsymbol{h}_i'$ 和 $\boldsymbol{h}_t'$ 同样是向前和向后的长短期记忆网络结构的输出的连接,使用同式(2)的计算方式。

由此,通过这一额外的神经网络模型,可以得到聚类结果增强的新闻内容句子表示 $\boldsymbol{e}^s$。最后,再将新闻内容各个分词对应的词向量 $\boldsymbol{e}^w$ 与整个新闻内容的句子表示 $\boldsymbol{e}^s$ 进行连接,作为基准模型的输入,以此来改进基准模型:

$$\hat{\boldsymbol{x}} = [\boldsymbol{e}^w; \boldsymbol{e}^s] \tag{8}$$

之后整个模型的训练过程与 2.3.1 节介绍的基准模型的训练过程类似。

2.3.3 融入聚类模型和 LDA 模型的中文标题生成模型

本文进一步对文献[19]提出的 NCSCM 框架进行了改进,设计了 4 种将聚类模型得到的聚类结果和 LDA 主题模型获取的主题词特征信息进行融合来加强对新闻内容的上下文表示的方法和模型。

首先,利用 LDA 主题模型对聚类结果中每个簇内的所有新闻内容进行分析,获取与该簇主题相关的、出现概率最高的前 k 个主题关键词。之后,将每一个簇的 k 个主题词转化为主题词特征向量,并构建了 4 种在模型中融合主题词信息与簇编号信息的方法。各方法对应的模型如图 1 中的方法②~⑤所示,具体的实现方法如下所述。

(1) 方法②:模型称为 SOM-NCSCM _ word,将 2.3.1 节 $\boldsymbol{x}'$ 中的簇编号特征 $\boldsymbol{e}^c$ 替换为主题词特征 $\boldsymbol{e}^t$,作为基于注意力机制的双向长短期记忆网络模型的输入:$\boldsymbol{x}'^{t_{\text{only}}} = (\boldsymbol{e}_{1,1}'^{t_{\text{only}}}, \cdots, \boldsymbol{e}_{1,m_1}'^{t_{\text{only}}}, \boldsymbol{e}_{2,1}'^{t_{\text{only}}}, \boldsymbol{e}_{2,2}'^{t_{\text{only}}}, \cdots, \boldsymbol{e}_{2,m_2}'^{t_{\text{only}}}, \cdots)$。其中 $\boldsymbol{e}_{i,m_i}'^{t_{\text{only}}} \in R^{d_w+d_n+d_p+d_t}$,$d_w$、$d_n$、$d_p$、$d_t$ 分别为词向量、命名实体特征向量、词性特征向量以及主题词特征 $\boldsymbol{e}^t$ 的向量维度。

(2) 方法③:模型称为 SOM-NCSCM _ combine,将主题词特征 $\boldsymbol{e}^t$ 与 $\boldsymbol{x}'$ 进行连接,作为基于注意力机制的双向长短期记忆网络模型的输入:$\boldsymbol{x}'^t = (\boldsymbol{e}_{1,1}'^t, \boldsymbol{e}_{1,2}'^t, \cdots, \boldsymbol{e}_{1,m_1}'^t, \boldsymbol{e}_{2,1}'^t, \cdots, \boldsymbol{e}_{2,m_2}'^t, \cdots)$。其中 $\boldsymbol{e}_{i,m_i}'^t \in R^{d_w+d_n+d_p+d_c+d_t}$,$d_w$、$d_n$、$d_p$、$d_c$、$d_t$ 分别为词向量、命名实体特征向量、词性特征向量、簇编号特征以及主题词特征的向量维度。

(3) 方法④:模型称为 SOM-NCSCM _ ave,首先对主题词特征 $\boldsymbol{e}^t$ 进行向量平均,然后将主题词特征的平均词向量进行复制与填充(Padding)操作后得到 $\boldsymbol{e}^{\text{ave}}$,再将其与 $\boldsymbol{x}'$ 进行连接,作为基于注意力机制的双向长短期记忆网络模型的输入:$\boldsymbol{x}'^{t_{\text{ave}}} = (\boldsymbol{e}_{1,1}'^{t_{\text{ave}}}, \cdots, \boldsymbol{e}_{1,m_1}'^{t_{\text{ave}}}, \boldsymbol{e}_{2,1}'^{t_{\text{ave}}}, \cdots, \boldsymbol{e}_{2,m_2}'^{t_{\text{ave}}}, \cdots)$。其中,$\boldsymbol{e}_{i,m_i}'^{t_{\text{ave}}} \in R^{d_w+d_n+d_p+d_c+d_{t\_\text{ave}}}$,$d_w$、$d_n$、$d_p$、$d_c$、$d_t$ 分别为与词向量、命名实体特征向量、词性特征向量、簇编号特征以及主题词特征的平均向量进行填充后的维度。

(4) 方法⑤:模型称为 SOM-NCSCM _ mix,将簇编号特征 $\boldsymbol{e}^c$ 与主题词特征 $\boldsymbol{e}^t$ 进行连接后,输入到额外的一个双向长短期记忆网络后,再经过一个全连接层后得到聚类融合特征 $\boldsymbol{e}^{\text{mix}}$。再将 $\boldsymbol{e}^{\text{mix}}$ 与词向量、命名实体特征和词性特征进行连接,作为基于注意力机制的双向长短期记忆网络模型的输入:$\boldsymbol{x}'^{t_{\text{mix}}} = (\boldsymbol{e}_{1,1}'^{t_{\text{mix}}}, \cdots, \boldsymbol{e}_{1,m_1}'^{t_{\text{mix}}}, \boldsymbol{e}_{2,1}'^{t_{\text{mix}}}, \cdots, \boldsymbol{e}_{2,m_2}'^{t_{\text{mix}}}, \cdots)$ 其中 $\boldsymbol{e}_{i,m_i}'^{t_{\text{mix}}} \in R^{d_w+d_n+d_p+d_{t\_\text{mix}}}$,$d_w$、$d_n$、$d_p$ 分别为词向量、命名实体特征向量、词性特征向量的维度,而 d_{t_mix} 为聚类融合特征 $\boldsymbol{e}^{\text{mix}}$ 的维度。

以上 4 种模型在后续的训练过程,与 2.3.1 节介绍的基准模型的训练过程类似,这里不再赘述。

3 数据与实验

3.1 数据与预处理

本文在实验中采用的数据集是哈尔滨工业大学整理的大规模中文短文摘要数据集(large-scale Chinese short text summarization dataset, LCSTS)[1]。其中的摘要数据来源于新闻媒体在中国社交平台新浪微博上发布的新闻内容,每条数据包含一个中文短文本和一条对应标题。另外,根据抽取式标题生成任务需求,本文对该数据集中的数据进行了一定的

预处理,具体操作包括:

(1) 从原始数据集中获取新闻正文内容和对应标题。

(2) 因数据来自网络,需去掉文本数据中的特殊符号。之后再使用 jieba 库对新闻内容和标题进行分词。

(3) 根据新闻标题中的分词,预先在新闻正文内容中依次、自动标注出标题中各词语出现过的位置。

(4) 人工核对,在新闻正文内容中选择语义更加连贯的词语片段,将对应的标题词语所在位置的标签标为"1",得到标签序列。之后再经过一遍自动核对,判断新闻内容中标为"1"的那些词语是否与标题中各个词语一致,也即无漏标、多标等情况。

(5) 随机抽取标注好的数据用于模型训练、验证与测试,并使用斯坦福大学提供的自然语言处理工具包 CoreNLP 对抽取的数据进行命名实体与词性标注。

最后,本文预处理后随机抽取 LCSTS 得到的用于本文实验的数据集合的统计信息如表 2 所示。在实验时,将其拆分成 8000 条训练数据,1000 条验证数据以及 1064 条测试数据。

表 2 LCSTS 数据集中抽取的新闻数据

总数	10 064 对(新闻内容,新闻标题)
压缩率	$\frac{\text{新闻标题总词数}}{\text{新闻内容总词数}}=0.136$
总词数	新闻内容:49 154 个词 新闻标题:18 287 个词
句子长度	新闻内容:76 ~ 151 个字、36 ~ 93 个词 新闻标题:3 ~ 30 个字、1 ~ 22 个词

3.2 实验设置

本文实验了 3 种初始化表示中文新闻数据以及主题词信息的方法,包括以字为基本单位的中文来自变换器的双向编码器表征量(bidirectional encoder representation from transformers, BERT)预训练模型[21]以及以词为基本单位的 2 种中文预训练模型(300 维度的 Word2Vector 预训练中文词向量[22]和中文 WoBERT 预训练模型),并采用 MiniSom 库来构建 SOM 模型,且将其输出层神经元结构设置为大小是 10 的一维线型结构,其他参数保持库中提供的模型默认值。另外,将命名实体特征、词性特征以及簇信息特征在模型训练阶段都分别初始化为 32 维的向量,双向长短期记忆网络的隐层维度都设置为 128 维,全连接层维度为 64 维。而为防止过拟合,在基准模型的双向长短期记忆网络与全连接层接受输入之前使用比例为 0.5 的 dropout 操作。整个模型在训练时的批大小为 64,并使用学习率为 0.001 的 Adam 算法来进行模型参数的优化和学习。

模型对应的预测结果都是在测试集上进行,且每个模型至少重复训练过 5 次,并最终选择效果趋于平均值的模型进行效果展示与比较。此外,为评估各新闻标题生成模型的效果,本文采用的主要评价指标为微观 $F1$ 值(micro $F1$)和压缩率(compression ratio, CR)[23],以及辅助评价指标 BLEU 值[24]和 ROUGE 值[25]。这 4 种评价指标的计算方式分别为

(1) 微观 $F1$ 值:

$$F1_{\text{micro}}=2\cdot\frac{Precision_{\text{micro}}\cdot Recall_{\text{micro}}}{Precision_{\text{micro}}+Recall_{\text{micro}}}$$

$$Precision_{\text{micro}}=\frac{\text{预测结果中标 0/1 正确的总数}}{\text{预测结果的 0/1 标签序列总长度}}$$

$$Recall_{\text{micro}}=\frac{\text{预测结果中标 0/1 正确的总数}}{\text{原新闻内容的 0/1 标签序列总长度}} \tag{9}$$

(2) 压缩率:

$$CR=\frac{\text{新闻标题的总词数}}{\text{新闻内容的总词数}} \tag{10}$$

(3) BLEU 值:

$$BLEU_n=\frac{\text{预测结果中 } n\text{-gram 出现在原标题中的总数}}{\text{预测结果中 } n\text{-gram 的总数}} \tag{11}$$

(4) ROUGE 指标:

$$ROUGE\text{-}N=\frac{\text{预测结果中 } n\text{-gram 出现在原标题中的总数}}{\text{原标题中 } n\text{-gram 的总数}}$$

$$ROUGE\text{-}L=\frac{(1+\beta^2)R_LP_L}{R_L+\beta^2P_L},(\beta\text{ 在本文中取 }1)$$

$$R_L=\frac{\text{预测结果与原生标题之间的最长公共子串的长度}}{\text{原数据的总长度}}$$

$$P_L = \frac{\text{预测结果与原标题的最长公共子串的长度}}{\text{预测结果的总长度}} \tag{12}$$

3.3 模 型

本文进行实验和效果评估与对比的各模型如下,在各模型中也分别实验了 3 种初始化词向量的方法。其中,使用以字为基本单位的预训练模型时,会对按照分词进行“0/1”标注的数据进一步结合 BIO(begin inside outside)标注法进行处理后再训练和测试模型。

(1) 基准模型:本文 2.3.1 节中介绍的模型。

(2) 融入传统聚类算法的中文标题生成模型:为了比较 SOM 模型的聚类效果,本文在 NCSCM 框架中实验了2种传统的聚类算法(K-means算法和 GMM 模型),并将这 2 种算法需要提前设置的聚类数目设置为与取得较好效果的 SOM 模型得到的聚类数目相同的数值(聚类数目设置为 10)

(3) SOM-NCSCM:本文 2.3.2 节中的方法①。

(4) 融入 SOM 模型和 LDA 模型的中文标题生成模型:本文 2.3.3 节中的方法② ~ ⑤,也即对 NCSCM 框架的 4 种改进模型。

3.4 主要实验结果

表 3 列出了各模型在抽取的 LCSTS 的测试数据上取得的效果。实验结果中,加粗字体对应的结果是当前评价指标下的最优值,下划线对应的结果是当前评价指标下的次优值,加星号对应的结果是当前预训练模型下各评价指标中的最优值。表 4 列出了在取得较好实验结果的聚类结果中,各簇经过

表 3 所有模型在 LCSTS 的测试集上的实验结果

	*F*1	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L	CR
Char-BERT											
基准模型	87.593	31.311	27.706	24.755	22.344	42.110	31.698	24.994	19.831	42.708	0.094
K-means + NCSCM	87.521	33.643	29.573	26.647	24.095	43.171	32.526	25.763	20.710	43.838	0.096
GMM + NCSCM	87.763	21.156	18.360	16.335	14.654	36.059	26.713	20.920	16.532	38.856	0.075
SOM-NCSCM	87.591	31.869	28.145	25.456	23.179	42.425	31.608	25.040	20.099	43.742	0.093
SOM-NCSCM _ word	87.654	35.874	31.006	27.777	25.144	44.648	33.752	26.851	22.028	43.608	0.102
SOM-NCSCM _ combine	87.769	32.019	27.906	24.997	22.656	42.524	32.245	25.764	20.914	42.145	0.095
SOM-NCSCM _ ave	87.794*	**37.687***	**33.030***	**29.639***	**26.818***	**46.131***	**35.407***	**28.621***	**23.551***	44.714*	**0.105***
SOM-NCSCM _ mix	88.076	31.618	27.494	24.715	22.496	43.064	31.923	25.439	20.718	44.642	0.094
Word2Vector											
基准模型	87.557	23.397	20.486	18.246	16.502	33.960	24.754	19.293	15.376	34.579	0.073
K-means + NCSCM	87.240	30.959	27.679	24.668	22.182	39.339	26.797	19.823	15.001	42.699	0.085
GMM + NCSCM	87.407	33.783	29.842	26.546	23.950	40.969	27.750	20.438	15.382	44.062*	0.087
SOM-NCSCM	87.443	31.294	28.065	24.908	22.193	39.357	26.162	18.920	14.166	42.770	0.085
SOM-NCSCM _ word	87.754	34.180	30.605	27.662	25.172	42.385	31.011*	23.989*	19.317*	42.934	0.090
SOM-NCSCM _ combine	87.764*	31.335	27.817	24.725	22.164	40.181	27.889	20.803	15.986	42.799	0.083
SOM-NCSCM _ ave	87.598	33.259	29.410	26.582	23.985	41.685	30.168	23.243	18.210	42.318	0.089
SOM-NCSCM _ mix	87.352	36.215*	32.032*	28.640*	25.784*	42.490*	29.428	21.797	16.869	43.932	0.093*
Word-BERT											
基准模型	88.638	33.735	29.804	26.514	24.011	43.498	33.018	26.219	20.855	43.966	0.088
K-means + NCSCM	88.388	32.310	28.642	25.706	23.052	41.806	29.960	22.855	17.810	44.243	0.084
GMM + NCSCM	88.620	32.763	29.976	27.040	24.468	42.924	32.095	25.301	20.284	**45.510***	0.085
SOM-NCSCM	88.426	34.439	30.677*	27.274	24.580	43.630	31.992	24.606	19.224	45.489	0.093*
SOM-NCSCM _ word	88.533	35.062*	30.675	27.514*	24.854*	43.888*	33.669*	26.983*	21.787*	43.440	0.090
SOM-NCSCM _ combine	88.680	31.657	27.814	24.991	22.813	41.847	32.585	26.322	21.438	41.445	0.084
SOM-NCSCM _ ave	**88.933***	32.140	28.079	25.007	22.607	42.805	33.168	26.663	21.763	42.670	0.084
SOM-NCSCM _ mix	88.848	32.026	28.880	25.888	23.385	42.516	31.654	24.954	19.560	45.493	0.082

表 4 LDA 主题模型从聚类结果的各簇数据中抽取的取得较高概率的前个 10 主题词

主题编号	主题	主题词
0	灾害灾难	禽流感、H7N9、人员伤亡、嫌疑人、遇难者、工作人员、直升机、枪击案、公交车、抢救无效
1	社会新闻	消费者、毕业生、大学生、负责人、互联网、教育部、公务员、出租车、工作人员、有限公司
2	信息科技	互联网、网络营销、O2O、智能手机、关键词、程序员、友情链接、二维码、阿里巴巴、科学家
3	国际新闻	奥巴马、钓鱼岛、发言人、委员会、领导人、金正恩、克里米亚、习近平、希拉里、联合国
4	社会活动	李克强、博览会、国务院、习近平、开幕式、李纪恒、第二届、互联网、丝绸之旅、有限公司
5	企业金融	亿美元、有限公司、净流入、证监会、投资者、人民币、董事长、上市公司、IPO、阿里巴巴
6	文娱新闻	好莱坞、世界杯、奥斯卡、万美元、科学家、小伙伴、奥巴马、洛杉矶、艺术家、第一次
7	金融新闻	人民币、百分点、中间价、亿美元、上半年、万亿元、美元汇率、交易日、去年同期、制造业
8	时事政治	国务院、习近平、群众路线、总书记、人大常委会、中央纪委、李克强、深化改革、领导小组、监督部
9	出行生活	高速公路、试运营、天然气、全线贯通、轨道交通、万人次、交通运输、博物馆、发改委、万平方米

LDA 主题模型从训练数据中获取的具有较高概率的前 10 个主题词。从各个模型在测试集上取得的效果中可以看到：

（1）整体上看，使用以词为基本单位的 WoBERT 预训练模型来初始化词向量的各模型在主要评价指标 *F*1 值上取得的效果均较使用另外 2 种预训练模型的效果好，而使用以字为基本单位的 BERT 预训练模型的各模型效果能够取得最优的压缩率、BLEU 和 ROUGE 值，这也就体现了 BERT 预训练模型能够输出语境信息更丰富的词向量的能力。

（2）相较于基准模型，融入了聚类结果以及主题词特征的各模型，其效果都明显提升。这说明数据中相似数据之间存在的特征信息（主题类别以及主题词信息）能够辅助相似新闻数据进行新闻标题词的选取。

（3）采用传统聚类算法的标题生成模型在 *F*1 值、压缩率、BLEU 和 ROUGE 指标上都较基准模型的效果有所提升，说明融入聚类特征能够让模型保留更多的新闻标题词，确保没有过度压缩新闻数据。

（4）结合表 4 展示的主题词信息可以看到，采用 SOM 进行神经网络聚类方法得到的模型较基于传统聚类算法的模型能够取得更好效果，且在仅采用聚类编号特征信息的方法①的基础上，在方法②～⑤中融入主题词特征信息后，可以进一步提升新闻标题生成的效果。

（5）此外，本文提出的方法②～⑤实验了融入主题词信息的不同方式。从实验结果中可看到，直接融入主题词特征信息来增强新闻句子表示的方式（方法②和③）能在各个指标上达到较优或者最优的效果，而通过更为深入和复杂的融合方式得到的模型（方法④和⑤），能够取得更佳的 *F*1 值、BLEU 和 ROUGE 值，但在其训练过程中，所需训练时间更长，并因模型参数更多，更易出现过拟合。

3.5 聚类模型消融实验

为更好地观察不同 SOM 聚类模型的神经元结构大小对设计的标题生成模型效果的影响，本文还进行了针对 SOM 聚类模型的消融实验：

（1）在方法④对应的 SOM-NCSCM _ ave 模型上融入不同 SOM 神经元结构大小的聚类结果和相应 LDA 主题模型获取的主题词信息，实验结果如表 5所示（使用中文 WoBERT 预训练模型初始化词向量），“SOM = X”对应着设置的不同 SOM 神经元结构大小，也即聚类结果中簇的数量。。

（2）从传统聚类方法（K-means 算法和 GMM 模型）以及不同 SOM 神经元结构大小的 SOM 聚类模型得到的各簇中分别都抽取了 200 条数据，计算不同聚类模型取得的轮廓系数（silhouette coefficient），结果如表 6 所示。

当神经元结构较小时，聚类得到的簇较少，而随着神经元结构增大，聚类得到的簇数量也逐渐增多。相应地，簇较少时，新闻数据难以得到充分聚类，而簇增多时，聚类到各簇的新闻数据量就会减少，导致有些相似新闻数据被过度细分。因此，本文采用了聚类效果最佳的、神经元结构大小为 10 的 SOM 模型进行各模型的实验和效果对比。

表 5　SOM-NCSCM _ ave 模型在不同 SOM 神经元结构大小下的实验结果

SOM 大小	*SOM* = 7	*SOM* = 8	*SOM* = 9	*SOM* = 10	*SOM* = 11	*SOM* = 12	*SOM* = 13	*SOM* = 14
*F*1	88.515	88.666	88.836	**88.933**	88.641	88.793	88.790	88.760
BLEU1	27.006	29.251	29.648	**32.140**	31.042	29.730	31.171	28.242
BLEU2	23.975	26.026	26.225	**28.079**	27.177	26.221	27.149	24.641
BLEU3	21.436	23.312	23.533	**25.007**	24.421	23.625	24.106	22.102
BLEU4	19.278	21.182	21.301	**22.607**	22.083	21.391	21.824	19.969
ROUGE-1	38.918	40.475	41.040	**42.805**	41.403	41.082	41.793	39.689
ROUGE-2	28.543	31.659	32.075	**33.168**	31.948	31.880	32.104	30.681
ROUGE-3	21.890	25.538	25.987	**26.663**	25.799	25.728	25.980	24.752
ROUGE-4	17.319	20.698	21.252	**21.763**	21.116	20.966	20.985	20.235
ROUGE-L	41.657	40.902	41.615	**42.670**	41.039	41.333	41.304	40.133
CR	0.076	0.081	0.081	0.084	**0.084**	0.080	0.083	0.078

表 6　不同聚类模型的轮廓系数

聚类模型	轮廓系数	聚类模型	轮廓系数
GMM	-0.0853	*SOM* = 10	**0.0493**
K-means	0.0467	*SOM* = 11	0.0471
SOM = 7	0.0456	*SOM* = 12	0.0439
SOM = 8	0.0486	*SOM* = 13	0.0455
SOM = 9	0.0480	*SOM* = 14	0.0483

3.6　举例分析

表 7 举例展示了 3 条测试集中的新闻数据以及 3 种在测试集上取得较好效果的模型的预测输出,从表中可以直观地看到融入 SOM 聚类模型与 LDA 主题模型对新闻标题生成效果的影响。其中,分词之间使用"/"分隔。

通过分析各模型的预测结果,可以总结出在处理该数据集上的新闻标题生成任务时现有模型的优势和存在的问题。

(1) 原标题与预测标题的质量:文献[1]也说明了其收集的新闻数据中原标题质量有好有差的情况。部分原标题包含了充分的新闻信息且语言更简练,而另一部分原标题相对更抽象、未能概括新闻内容的完整信息。如表 6 中的例 1 就是原标题缺少地点关键词("甘肃/定西")以及更充分的信息量("296/人/重伤"),而文本设计的各模型能预测并补全其原标题中缺乏的这些关键信息;例 3 则是原标题较抽象、缺乏事实相关信息的例子。此外,在依据本文实验需求进行数据标注时,也会存在少量的分词错误或分词不一致问题,如例 1 中的分词错误"中寨至",例 3 中的分词不一致"冷鲜/鸡"与"冷/鲜/鸡"。而从各模型的预测结果中可以看到,相较于其原标题,本文设计的模型能生成更可读的、与新闻内容关联更大的、包含更充分的信息量的新闻标题,且在模型中更充分地融入聚类和主题词信息,能够对分词问题导致的影响具有一定的鲁棒性。

(2) 新闻数据压缩程度:即新闻标题需要对新闻内容更加精炼的表达,同时不能丢失新闻内容中的关键信息。从表 6 的例子中可以看到,各模型对新闻标题词的选取,有时会保留更多的、不存在于原标题中的词语,但预测的标题在一定程度上也是可读且合理的。结合表 4 的实验结果,从压缩率指标上来看,各模型在预测时,总体上仍会倾向于保留较少的词语,这导致一些关键词被遗漏,例如表 6 例 2 中的"南海/网",以及"在/海南"在新闻内容出现的顺序偏后,没有得到模型更多的关注而被漏标。

4　结 论

针对中文新闻标题生成任务面临的大规模且高质量中文标注数据缺乏的问题,本文利用标题往往由原文中的词汇构成这一特点,将中文抽取式标题生成问题转化为序列标注问题,并提出了多种在深度神经网络中文标题生成模型中融入聚类和主题模型的方法。利用基于无监督学习的 SOM 聚类模型和 LDA 主题模型自动挖掘出表达相同或相似主题

表 7 3 条新闻内容以及 3 种模型的预测结果

例 1	截止/12:30/,/甘肃/定西/地震/已造成 47/人/死亡/、/296/人/重伤/。/其中/:/岷县/死亡/45/人/、/漳县/1/人/,/陇南市/礼县/死亡/1/人/。/漳县/13/个/乡镇/多数/房屋/出现/裂缝/、/严重/损毁/房屋/5600/户/21000/间/,/倒塌/380/户/1203/间/;/岷县/中寨至/小寨/二级/光缆/中断/,/移动/通信/信号/中断/。		
原标题	已/造成/47/人/死亡		
模型	Char-BERT	Word2Vector	Word-BERT
SOM _ NCSCM _ word	甘肃/定西/地震/已/造成/47/人/死亡/296/人/重伤	甘肃/定西/地震/已/造成/47/人/死亡/296/人/重伤	甘肃/定西/地震/已/造成/47/人/死亡/296/人/重伤
SOM _ NCSCM _ ave	甘肃/定西/地震/已/造成/47/人/死亡/296/人/重伤	甘肃/定西/地震/已/造成/47/人/死亡/、/296/人/重伤	甘肃/定西/地震/已/造成/47/人/死亡/296/人/重伤
SOM-NCSCM _ mix	甘肃/定西/地震/已/造成/47/人/死亡/296/人/重伤	甘肃/定西/地震/已/造成/47/人/死亡/296/人/重伤	甘肃/定西/地震/已/造成/47/人/死亡/296/人/重伤
例 2	6/日/下午/,/南海/网/对/博鳌/亚洲/论坛/期间/举行/的/海南省/新闻/发布会/进行/了/4G/手机/网络/视频/直播/。/作为/海南省/重点/新闻/网站/,/南海/网/联合/中国/移动/海南/公司/率先/成功/实现/4G/手机/视频/直播/,/这/在/海南/媒体/中/还是/首例/。		
原标题	南海/网/率先/在/海南/成功/实现/4G/手机/视频/直播		
模型	Char-BERT	Word2Vector	Word-BERT
SOM _ NCSCM _ word	4g/手机/网络/视频/直播/实现/4g/手机/视频/直播	中国/移动/海南/公司/率先/成功/实现/4G/手机/视频/直播	4G/手机/网络/视频/直播/率先/成功/实现/4G/手机/视频/直播
SOM _ NCSCM _ ave	4g/手机/网络/视频/直播/成功/实现/4g/手机/视频/直播	南海/网/联合/中国/移动/4G/手机/视频/直播	4G/手机/网络/视频/直播/4G/手机/视频/直播
SOM-NCSCM _ mix	实现/4g/手机/视频/直播	4G/手机/视频/直播/首例	率先/成功/实现/4G/手机/视频/直播
例 3	5 月/30 日/起/,/上海/开始/将/冷鲜/鸡/全面/推向/市场/。/众多/销售点/出现/销售/一空/情景/,/有的/店/6/点/半开门/,/7/点/半/就/全部/卖完/,/某/优选/店/内/每/只/冷/鲜/鸡/约/合/40/元/上下/。/据悉/,/上海/活禽/交易点/或/减半/,/你/会/去/尝试/冷鲜/鸡/吗/? /你/知道/冷鲜/鸡/和/热气/鸡肉/的/区别/吗/? /戳		
原标题	冷/鲜/鸡/,/你/会/尝试/吗/?		
模型	Char-BERT	Word2Vector	Word-BERT
SOM _ NCSCM _ word	(空)	(空)	(空)
SOM _ NCSCM _ ave	上海/活禽/交易点/或/减半	上海/开始/将/冷鲜/鸡/全面/推向/市场	上海/活禽/交易点/或/减半
SOM-NCSCM _ mix	上海/冷鲜/鸡/全面/推向/市场	上海/冷鲜/鸡/全面/推向/市场	上海/活禽/交易点/或/减半

的数据以及数据中的主题词信息,在基于监督学习的深度神经网络模型中融入这些特征,增强对新闻内容的上下文表示,从而辅助中文新闻标题生成。在互联网上公开的、缺乏人工标注分类信息的 LCSTS 中文新闻数据集上的实验表明,本文提出的模型在各评价指标上的结果较基准模型都有所提升,也提高了中文标题生成的质量。未来的工作可以从提升压缩率来避免过度压缩、减少关键信息遗漏的角度出发继续研究,也可以设计人工评价策略来更细致地评估模型的效果,允许模型生成多样化的标题。

参考文献

[1] HU B, CHEN Q, ZHU F. LCSTS: a large scale Chinese short text summarization dataset[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: EMNLP, 2015:1967-1972.

[2] CAO Z, LI W, LI S, et al. Improving multi-document summarization via text classification[C] // Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2017,31(1):3053-3059.

[3] 鹿忠磊,刘文芬,周艳芳, 等. 基于预读及简单注意力机制的句子压缩方法[J]. 计算机应用研究, 2019,36(2):371-375, 394.

[4] CHOPRA S, AULI M, RUSH A M. Abstractive sentence summarization with attentive recurrent neural networks [C] // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: NAACL HLT, 2016:93-98.

[5] HIGURASHI T, KOBAYASHI H, MASUYAMA T, et al. Extractive headline generation based on learning to rank for community question answering[C] // Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: COLING, 2018:1742-1753.

[6] AYANA, WANG Z, XU L, et al. Topic-sensitive neural headline generation[J]. Science China Information Sciences, 2020,63(8):1-16.

[7] KOHONEN T. Self-organized formation of topologically correct feature maps[J]. Biological Cybernetics, 1982, 43(1):59-69.

[8] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003,3:993-1022.

[9] 焦利颖, 郭岩, 刘悦, 等. 基于序列模型的单文档标题生成研究[J]. 中文信息学报, 2021,35(1):64-71.

[10] REN P, CHEN Z, REN Z, et al. Leveraging contextual sentence relations for extractive summarization using a neural attention model[C] // Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieva. Tokyo: Association for Computing Machinery, 2017:95-104.

[11] AFSHARIZADEH M, EBRAHIMPOUR-KOMLEH H, BAGHERI A. Query-oriented text summarization using sentence extraction technique[C] // 2018 4th International Conference on Web Research. San Francisco: ICWR, 2018:128-132.

[12] WANG S, ZHAO X, LI B, et al. Integrating extractive and abstractive models for long text summarization[C] // The 2017 IEEE International Congress on Big Data. Honolulu: IEEE, 2017:305-312.

[13] XU J, DURRETT G. Neural extractive text summarization with syntactic compression[C] // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: EMNLP, 2019:3292-3303.

[14] ZHOU Q, WEI F, ZHOU, M. At which level should we extract? An empirical study on extractive document summarization[C] // Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: COLING, 2020:5617-5628.

[15] ZHONG M, LIU P, WANG D, et al. Searching for effective neural extractive summarization: what works and what's next[C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 1049-1058.

[16] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C] // Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967:281-297.

[17] ZHANG T, RAMAKRISHNAN R, LIVNY M. BIRCH: an efficient data clustering method for very large databases [J]. ACM Sigmod Record, 1996,25(2):103-114.

[18] RASMUSSEN C. The infinite Gaussian mixture model [C] // Proceedings of the 12th International Conference on Neural Information Processing Systems. Denver: MIT Press, 1999:554-560.

[19] ZI K, WANG S, LIU Y, et al. SOM-NCSCM: an efficient neural Chinese sentence compression model enhanced with self-organizing map[C] // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana: EMNLP, 2021:403-415.

[20] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C] // The 3rd International Conference on Learning Representations. San Diego: ICLR, 2015: 1-15.

[21] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2019-05-24) [2022-07-29]. https://arxiv.org/pdf/1810.04805.pdf.

[22] LI S, ZHAO Z, HU R, et al. Analogical reasoning on Chinese morphological and semantic relations[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018: 138-143.

[23] NAPOLES C, VAN DURME B, CALLISON-BURCH C. Evaluating sentence compression: pitfalls and suggested remedies[C]//Proceedings of the Workshop on Monolingual Text-To-Text Generation. Portland: Association for Computational Linguistics, 2011: 91-97.

[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 of the Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002: 311-318.

[25] LIN C Y. ROUGE: a package for automatic evaluation of summaries[C]//Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004. Barcelonal: Association for Computational Linguistics, 2004:74-81.

SOM-NCSCM + : research on Chinese headline generation method based on extractive neural network

ZI Kangli[*,**], WANG Shi[*], CAO Cungen[*]

([*]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

([**]University of Chinese Academy of Sciences, Beijing 100049)

Abstract

As a branch of text summarization task, headline generation can help people obtain information efficiently. In this paper, aiming at the lack of large-scale and high-quality Chinese annotation data in the Chinese headline generation task, taking advantage of the feature that headlines can often be formed from words in the contents, a Chinese headline generation method and model based on extractive deep neural network is proposed. The whole model is enhanced with the clustering model and the topic model, from the perspective of combining unsupervised learning model with supervised sequence labeling model. On the Chinese news data lacking manual annotated classifications, the whole model can automatically mine potential feature information within the data, and obtain different data clusters and the topic words to assist Chinese news headline generation by applying the clustering model and topic model, which makes the whole model more adaptable on the Chinese news data of different topics and uneven annotation quality. The experimental results on a dataset of Chinese news headline generation publicly available on the Internet also show that this whole model achieves better performance on the evaluation metrics, including the micro *F*1, BLEU, ROUGE and compression ratio than the baseline models.

Key words: Chinese headline generation, neural network model, topic model, clustering model, sequence labeling