

文本主题分割任务中的深度学习研究综述

周日飞, 文 勇, 黄文旭

(广西民族大学 人工智能学院, 广西 南宁 530000)

摘 要: 深入探讨文本主题分割任务, 系统梳理并提炼现有文献, 开展针对基于深度学习的文本主题分割技术的多维度分析。首先, 回顾了文本主题分割技术的发展历程, 分析了基于语篇结构理论、统计学习和深度学习的分割方法, 并考察文本主题分割任务常用的数据集及评价指标; 其次, 进一步比较了卷积神经网络、循环神经网络和Transformer网络等深度学习技术在文本主题分割任务中的应用和优缺点分析, 并对比采用各模型方法在相同数据集上的实验结果; 最后, 指出了当前文本主题分割任务面临的挑战和潜在发展方向, 为文本主题分割任务的进一步研究提供了参考。

关键词: 文本主题分割; 深度学习; 自然语言处理; 数据集; 评价指标

DOI: 10.11907/rjdk.241967

中图分类号: TP391.1

文献标识码: A

文章编号: 1672-7800(2026)001-0202-11

扫描二维码阅读全文:



Deep Learning Research Review on Text Topic Segmentation Task

ZHOU Rifei, WEN Yong, HUANG Wenxu

(School of Artificial Intelligence, Guangxi Minzu University, Nanning 530000, China)

Abstract: This study aims to provide an in-depth exploration of text topic segmentation tasks by systematically reviewing and synthesizing existing literature, conducting a multidimensional analysis of deep learning-based text topic segmentation techniques. The research begins with a review of the evolution of text topic segmentation technologies, analyzing methods based on discourse structure theory, statistical learning, and deep learning, and examining the datasets and evaluation metrics commonly used in text topic segmentation tasks. Then, It further compares the applications and advantages and disadvantages of deep learning technologies such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer models in text topic segmentation tasks, and contrasts the experimental results of various model methods on the same dataset. Ultimately, the study identifies the challenges currently faced in text topic segmentation tasks and potential directions for future development, offering a reference for further research in text topic segmentation.

Key Words: text topic segmentation; deep learning; natural language processing; datasets; evaluation metrics

0 引言

在自然语言处理(Natural Language Processing, NLP)领域, 文本主题分割任务(Topic Segmentation, TS)的发展根植于对文本语篇结构的细致建模与研究, 它不仅是文本结构分析的基础, 也是文本构造的起点。然而, 随着互联网技术的突飞猛进, 信息爆炸时代已然到来, 文本数据的来源和数量呈现出前所未有的增长。在这一背景下, 准确识别文本中的主题单元变得尤为关键, 它不仅关系到信息的

组织和检索效率, 也影响着知识的提取和传播。首先, 在语义理解和分析任务中, 文本主题分割具有重要作用。例如, 在文本摘要系统的研究中, 主题分割算法用于对长文本内容按语义内容执行分割, 以便于减少无关信息的计算^[1]; 在对话系统的研究中, 它通过识别对话中的不同主题段落, 帮助系统理解对话的脉络和结构, 从而更准确地捕捉用户意图和上下文信息^[2]。其次, 在信息组织和检索工作中其也发挥着重要作用。例如, 在话题检测与跟踪(Topic Detection and Tracking, TDT)任务中, 主题分割技术通过识别文本中的话题边界和特征, 助力话题检测与跟踪

收稿日期: 2024-11-15

基金项目: 广西重点研发计划项目(AB22035034); 广西壮族自治区科学技术厅广西科技基地和人才专项(AD22035200)

作者简介: 周日飞(1999-), 男, CCF会员, 广西民族大学人工智能学院硕士研究生, 研究方向为文本主题分割; 文勇(1969-), 男, 硕士, 广西民族大学人工智能学院副教授、高级工程师, 研究方向为大数据处理、自然语言处理、知识图谱; 黄文旭(2000-), 男, 广西民族大学人工智能学院硕士研究生, 研究方向为自然语言处理。本文通讯作者: 文勇。

任务,提升信息检索的准确性和个性化内容推荐的效率^[3];在警情系统中,通过主题分割技术对犯罪人员的对话记录按主题分割,可以帮助警务人员快速筛选关键信息,以掌握犯罪事实^[4]。随着技术的不断进步,文本主题分割任务有望在更多领域发挥关键作用,为现代社会的信息处理和知识管理带来革命性的变化。因此,系统地分析和研究主题分割算法具有重要意义。

文本主题分割任务现有的研究方法可以归纳为三大主流类别:基于语篇结构理论的方法、基于统计学习的方法以及基于深度学习的方法。基于语篇结构理论的方法主要依赖于研究者的目的、直觉和语篇结构理论,如修辞结构理论^[5]、Grosz&Sider的三分理论^[6]。这些方法通常产生粒度较细的分割结果,即所谓的细粒度分割(fine-grained segmentation)。基于统计学习的方法主要以算法学习识别文本内容的主题变化作为依据来实现文本主题分割。这一方法的理论依据建立于Halliday等^[7]所著的《英语的凝聚力》对于文本的属性的阐述。文本内聚关系(Cohesion,连贯性)是体现文本片段主题内容差异的重要考量。基于语篇结构理论和统计学习的文本主题分割方法,大多采用线性分割策略。这些方法在识别文本片段边界时,往往忽视了片段的内在结构,导致在数据集规模受限的情况下,性能提升遭遇瓶颈。

深度学习网络(Deep Learning)是机器学习(Machine Learning)领域一个重要研究方向,旨在通过学习样本数据的内在规律和表示层次,使机器能够像人一样具有分析学习能力,识别文字、图像和声音等数据,其核心思想是通过多层神经网络来提取数据的特征,逐渐将低层特征转化为高层特征,最终用简单的模型完成复杂的分类等学习任务。这种方法被称为“特征学习”或“表示学习”。近年来,深度学习在计算机视觉、图像处理以及自然语言处理等多个领域引发了重大变革,并取得了令人瞩目的成就。在文本主题分割这一自然语言处理的细分领域,深度学习同样展现了其强大的潜力和影响力。随着深度学习技术的兴起,文本主题分割的研究范式经历了显著转变。当前的研究趋势更加注重文本片段的内容分析,不仅关注单词层面的变化,还深入探讨了语义信息、上下文语句的动态变化等多维度语言特征,以期获得更强的文本主题分割性能。

近年来,主题分割(TS)任务的研究领域经历了显著的变化,无论是在研究范式、内容还是方法上。尽管已有研究如PAK等^[8]对文本分割在多种分析方法中的应用进行了分析,揭示了不同语言中文本分割的使用情况,以及石晶^[9]对早期TS任务进行了系统性总结,包括其原理和评测手段,但这些研究并未全面覆盖最新的深度学习方法,而且发布时间较早,对当前TS任务的参考价值有限。此外,这些研究普遍对于数据集、模型性能等方面的说明以及方法优缺点分析讨论深度不足。因此,本文旨在提供一个全面的系统性分析和报告,内容涵盖研究方法、数据集说明、

模型性能对比分析等,以填补这一空白。

1 文本主题分割任务概述

1.1 文本主题分割定义

文本主题分割任务作为文本分割领域的一个重要分支,在近年来已成为该领域的主要研究方向。该任务专注于主题层面的分割粒度,旨在识别包含人为主观特征的文本,如会议记录、网络文章或对话中的主题转换边界。

文本主题分割是指在文本处理中识别并划分出不同主题段落的过程。从形式上而言,对于给定的文本序列 $S = \langle U_1, U_2, U_3, \dots, U_n \rangle$,经过文本主题分割算法分割后会得到以每个文本序列单元之间间隔成为分割点的可能概率。其中,高概率的间隔将被判定为分割点,并给予分割标记。如图1所示,经过文本主题分割器后, U_1 和 U_2 以及 U_4 和 U_5 之间的间隔被预测成为分割点,从而划分得到了 $\{U_1\}$ 、 $\{U_2, U_3, U_4\}$ 、 $\{U_5, U_6\}$ 3个片段内主题一致的段落。

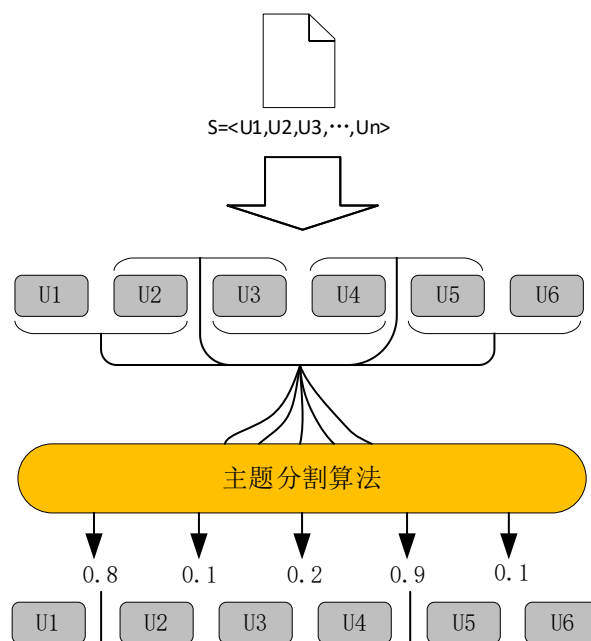


Fig. 1 Text topic segmentation task instance

图1 文本主题分割任务示例

1.2 数据集

早期的主题分割(TS)研究长期受到带注释训练语料库稀缺的限制,这一挑战促使传统统计学习和机器学习算法主要依赖于无监督学习方法来寻找可能的解决方案。然而,随着近些年互联网的迅猛发展,主题分割领域迎来了转机,涌现出了一系列高质量的数据集,如WiKi-727K、Wiki-Section等。这些数据集的共同特点是规模庞大且覆盖广泛的主题领域,为研究者提供了丰富的资源来训练和评估模型。这些数据集的出现极大地推动了主题分割技术的进步,使得研究者能够更深入地探索和改进算法,从而在自然语言处理的多个子领域中实现更准确的文本分割。因此,

本文对数据集按照发布时间顺序进行汇总整理,旨在为相关研究提供全面的资源概览,具体如表1所示。

Table 1 Summary of common datasets for topic segmentation tasks

表1 主题分割常用数据集汇总

数据集	语言	年份	语料来源	文档总数	地址
Choi ^[10]	英文	2000	Brown 语料库	920	—
Clinical ^[11]	英文	2008	医学著作节选	224	https://onlinebooks.library.upenn.edu
Wikicities ^[12]	英文	2009	维基百科	100	https://manifestoproject.wzb.eu/
Wikielements ^[12]	英文	2009	维基百科	118	—
Fiction ^[13]	英文	2011	虚构小说集	85	—
WIKI-727k ^[14]	英文	2018	维基百科	727 746	https://github.com/koomri/text-segmentation
WikiSection ^[15]	多语言	2019	维基百科	38 000	https://github.com/sebastianarnold/WikiSection
Doc2Dial ^[16]	英语	2020	人机对话记录	4 130	https://github.com/doc2dial/sharedtask-dialdoc2021
DialSeg711 ^[2]	英语	2021	人机对话记录	711	https://github.com/lxing532/Dialogue-Topic-Segmenter/tree/main/data/eval
SuperDialseg ^[17]	英语	2023	人机对话记录	9 478	https://drive.google.com/drive/folders/19YiHVfel_M4HivrErLi9bghvUvsw9-Ws?usp=sharing

本文将重点对比各模型在 Wikicities、Wikielements、Choi、Clinical、WIKI-727k、WikiSection 等数据集上的性能。相关数据集介绍如下:

(1)Choi 数据集发布于2000年,由700个文档组成,每个文档包含10个不同主题的段落。这些段落从Brown语料库中随机选取组合而成,特点鲜明。但由于数据量有限且段落不连续,通常用于TS模型算法性能的测试而非训练。这限制了其在实际应用中的广泛性。

(2)WikiCities 和 Wikielements 数据集均基于维基百科内容,覆盖了城市描述和化学元素两大领域。Wikicities 包含100个关于主要城市的描述文档,而 Wikielements 则包含118个化学元素的描述文档。这2个数据集的优势在于其广泛的主题覆盖和结构化的知识元素信息,但数据集规模并不大。因此,在TS任务中,通常作为TS模型算法的测试数据集所使用。

(3)Clinical 数据集源自专业医学书籍,其包含227章内容、1 136个章节,平均每章节包含5个句子。该数据集结构化良好,章节划分明确,为医学文本分析提供了便利。由于样本量较少,其通常作为测试数据集使用,用于评估TS模型算法的性能和准确性,但可能存在知识更新不及时的风险。

(4)WIKI-727K 数据集发布于2018年,标志着主题分割领域的一个重要进展。在深度学习技术不断推进的背景下,Koshorek 等^[8]提出了WIKI-727K数据集。这是一个经过精心标注的、面向文档的文本分割数据集,极大地推动了有监督文本分割技术的发展。WIKI-727K数据集由727 746个英语维基百科文档组成,其庞大的数据规模和广泛的主题覆盖,使其成为评估TS模型算法性能的理想基准数据集。

(5)WikiSection 数据集发布于2019年,包含来自英语和德语维基百科约38 000篇文章,涵盖了242 000个章节。每篇文章都配有原始标题和标准化的主题标签,覆盖了多达30个不同的主题领域。WikiSection数据集为有监督的TS任务提供了宝贵的资源,推动了TS模型算法的发展和应用。

(6)Doc2Dial 数据集由IBM在2021年发布,是一个富有创新性的人机对话数据集。该数据集涵盖了4个不同领域,包含480篇文档和4 800个多轮对话,每个对话平均包含14次交互。这一数据集的先进之处在于,它模拟了对话系统根据一本手册或文档来回复人类问题的场景,确保所有问题的答案都能在文档中找到。Doc2Dial数据集涉及2个核心任务:一是从文档中检索与问题相关的句子(information-seeking);二是根据检索结果生成合理的答复(response generation)。该数据集为研究者在探索如何从大量文档中高效检索信息以及理解用户查询的复杂性时提供了新的思路。但该数据集的缺点在于,对话内容必须严格基于文档内容,这可能会限制对话的自由度和创造性,因此可能不太适合作为TS任务的训练数据集使用。

(7)DialSeg711 数据集是一个专注于对话分割(Dialogue Segmentation)任务的数据集。该数据集包含了711个对话实例,每个实例都由多个对话轮次组成,涉及不同的主题和语境。数据集中的对话都已被人工标注,即对话从一个主题转换到另一个主题的边界标记。这些标注为训练和评估对话分割模型提供了重要的参考,适用于自然语言处理中的对话分析、对话系统开发以及任何需要理解和处理对话结构的场景。该数据集的出现填补了TS任务中带标注对话文本数据集的空白,为研究者提供了宝贵的资源。

通过上述系统性分析,本小节旨在揭示不同数据集的特点、优势和局限性,以及它们在主题分割任务中的应用价值。这将为未来的研究提供指导,并帮助研究者选择合适的数据集进行TS模型算法训练和评估。

1.3 相关评估指标

TS任务核心在于识别文本中主题的转换点,属于文本分析领域的关键问题。在此任务中,部分算法专门用于检测主题变化,并通过预测文本序列中的变化点来确定话语间隔是否标志着新主题的开始。在评估这些算法性能时,通常需要将模型的预测结果与真实值进行比较,并基于此采用多种评估指标来衡量其性能。其中,F1得分、召回率、全类平均精确率(Mean Average Precision, mAP)、 P_k 、Win-

dowDiff(以下简称WD)是5种广泛应用的指标。本文将重点介绍 P_k 和WD这2个专用于文本主题分割任务的评估指标,因为后续的模型性能讨论将主要基于这2个指标进行比较。

1.3.1 P_k 指标

Beeferman等^[18]于1997年提出 P_k 评估指标。该指标使得算法和模型的分割能力能够得到更丰富、多层次的体现,如公式(1)所示。

$$P_k(\text{ref}, \text{hyp}) = \frac{\sum_{i=1}^{N-k} \delta_{\text{ref}}(i, i+k) \oplus \delta_{\text{hyp}}(i, i+k)}{N-k} \quad (1)$$

其中, $\delta_s(i, j)$ 代表了阶跃函数,而ref和hyp则分别表示在 n 个句子上的真实分割标签和预测分割标签; N 表示测试文本中的话语个数, k 表示窗口大小; $\oplus$ 表示异或运算符号,式子左右两边数值相同则为0,不同为1。

P_k 的取值范围在 $[0, 1]$ 之内,当预测的分割标签与真实标签完全一致时,其值为0。该指标将预测分割标签的位置与真实分割标签的位置之间的距离进行关联,反映预测误差的程度。具体而言,预测分割位置与真实分割位置之间的偏差越大,即错误程度越高;相应的,预测分割位置与真实分割位置越远,即错误的程度越高, P_k 值也就越大,从而施加更大的惩罚。

1.3.2 WindowDiff(WD)指标

在Pevzner等^[19]研究中指出, P_k 指标背后的基础假设前提存在瑕疵,如假负比假正更容易受到惩罚,对文本块的大小较为敏感等。因此,在使用 P_k 指标衡量模型分割能力时仍存在问题。于是,针对 P_k 指标进行改进,他们提出了更完善的指标WindowDiff(以下简称WD),如公式(2)所示。

$$\text{WindowDiff}(\text{ref}, \text{hyp}) = \frac{\sum_{i=1}^{N-k} |b(\text{ref}_i, \text{ref}_{i+k}) - b(\text{hyp}_i, \text{hyp}_{i+k})|}{N-k} \quad (2)$$

其中, $b(\text{ref}_i, \text{ref}_{i+k})$ 和 $b(\text{hyp}_i, \text{hyp}_{i+k})$ 分别代表了真实分割文本和预测分割文本中 i 和 $i+k$ 位置之间的边界数量,其他数字含义同公式(1)所示。该式在对文本主题分割算法的结果中,使用一个固定大小为 k 的窗口在文本上从左向右移动,将预测的分割结果(hyp)与真实的分割结果(ref)进行比较。具体而言,是计算在固定大小为 k 的窗口内,真实分割(ref)和预测分割(hyp)之间的边界数量差异。该差异通过绝对值来衡量,即计算真实分割和预测分割在每个窗口位置的边界数量之差的绝对值,随后对所有窗口位置的差异求和并取平均,得到最终的WD指标得分。得分越低,表示预测分割与真实分割之间的差异越小,模型的分割能力越强。

2 基于深度学习的TS方法

深入分析深度学习技术在文本主题分割(TS)任务中

的应用,具体包括卷积神经网络(CNN)、循环神经网络(RNN)以及预训练模型在内的深度学习模型。通过在大量预定义任务样本上的学习,展现出了超越传统方法的性能。这些模型能够有效理解文本的复杂语义、感知上下文,并捕捉文本的语义、结构和语法等潜在特征,从而显著提高了文本主题分割算法的准确性。

2.1 基于深度学习的TS流程

基于深度学习的文本主题分割任务通常遵循一个层次化结构模型的处理流程,主要包括3个层级:嵌入表示层、序列编码层和标签解码层,如图2所示。这一流程的设计旨在通过主题分割算法来逐步提炼和转化文本数据。原始文本经过算法的主题识别后,得到以文本语义为分割依据的主题片段结果。

(1)嵌入表示层。负责将文本序列转换为高维空间中的向量表示,捕捉词汇的语义信息。这一步骤借助预训练的词嵌入技术,如Word2vec和Glove等静态词向量模型实现。

(2)序列编码层。进一步对这些向量进行深度加工,提取文本的上下文和语义特征,建立上下文联系。同时,利用CNN、LSTM、BiLSTM以及BERT等特征提取器,进一步提取和整合文本的深层特征。

(3)标签解码层。将编码后的文本特征通过TextTiling、softMax等手段映射到具体的分割标签上,完成文本主题的自动识别和分割。

2.2 嵌入表示层

在基于深度学习的文本主题分割任务中,嵌入表示层扮演着将原始文本转换为机器可理解的数值向量的关键角色。这一层次常用的技术主要包括Word2Vec和GloVe向量模型。其中,Word2Vec这一经典的词向量训练模型由Mikolov等^[20]于2013年提出,其通过预测上下文以及分析词与词之间的共现信息来学习词的向量表示。Word2Vec方法提供了连续词袋(CBOW)和跳字模型(Skip-gram)2种训练方法,从而更好地捕捉局部上下文信息,生成固定长度的词向量,但其缺点在于无法有效处理词汇的上下文变化。Pennington等^[21]针对Word2Vec的局限性,即主要依赖有限窗口内的上下文信息,提出了GloVe模型。GloVe结合全局矩阵分解的思想,能够捕捉更广泛的上下文关系,从而提供一种更全面的方法来学习词向量。在主题分割任务中,Word2Vec和GloVe因其在捕捉词项语义和上下文关系方面的优势,成为了主题分割方法中嵌入表示层的主要词向量表示方法。

2.3 文本编码层

在探讨基于深度学习的文本主题分割(TS)模型时,文本编码层的设计选择对模型性能有着决定性的影响。当前,TS模型算法中文本编码层的方法主要集中在3种主流的深度学习架构:卷积神经网络(CNN)、循环神经网络(RNN)以及Transformer网络。这些方法各自具有独特的

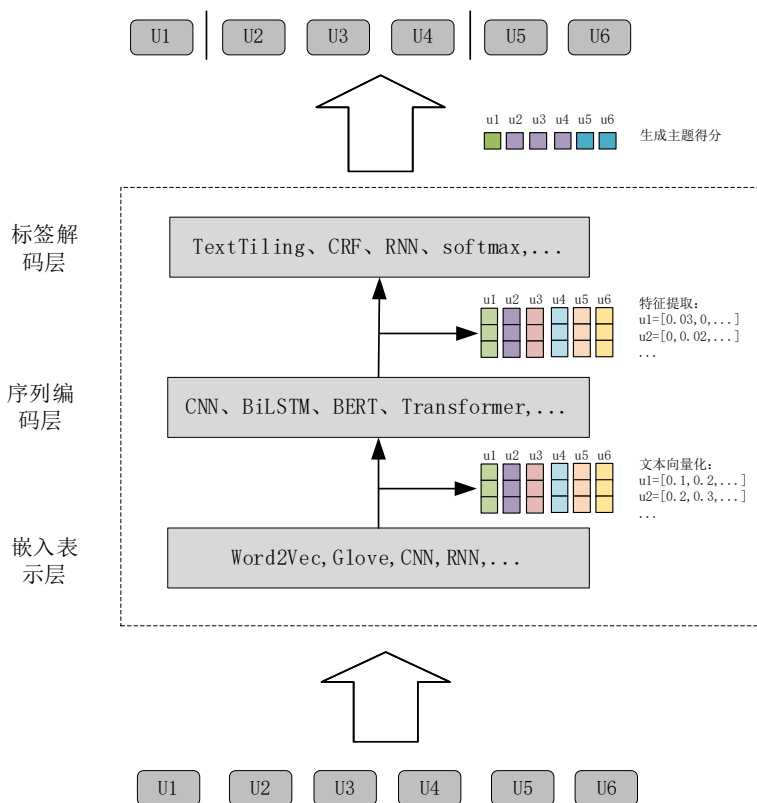


Fig. 2 The process of deep learning-based topic segmentation (TS) tasks

图 2 基于深度学习的 TS 任务流程

优势和适用场景,汇总如表 2 所示。它们在处理文本数据时展现出不同的特征提取能力和对序列依赖的处理方式。

2.3.1 卷积神经网络(CNN)

卷积神经网络(CNN)作为新一轮人工智能研究浪潮

Table 2 Comparison of text encoding techniques

表 2 文本编码层方法比较

深度学习技术	方法描述	优点	缺点
卷积神经网络	设计对称卷积操作,提取局部文本结构向量特征,通过结构向量特征判断文本的连贯性	适合处理具有明显局部结构的文本数据,如句子或短语级别的特征提取,适用于短文本场景	在处理长距离依赖和序列数据的整体上下文方面存在局限,因其缺乏循环结构,难以有效捕捉跨长段落的语义联系
循环神经网络及其融合模型	对文本主题内容建模,辅以学习主题分割任务中语言学特征如韵律、语法结构等以识别主题转换点	适用于长文本处理,并且可以根据任务组合,灵活性较强,适用于主题单一的分割场景	跨域泛化能力不足,需要特征工程辅助;可能遇到梯度消失问题,且在大规模数据集上易过拟合
Transformer	按辅助任务设计微调 Transformer 预训练模型,增强对数据的理解,融合其他网络模型或算法输出分割预测	跨域泛化能力强,可以实现并行计算,提高计算效率;可以处理超长文本,适用于主题转变快且主题丰富的分割场景	对数据集质量要求高,模型资源消耗随模型参数增长而呈现指数型增长,同时模型收敛速度慢

的开端,不仅在图像领域屡创佳绩,同时在循环神经网络尚未流行的时候,引领了早期自然语言处理的发展。

CNN 通过局部感受野捕捉文本中的局部模式,对于识别局部上下文和短语级别的特征非常有效。它们在图像处理中的表现优异,而在文本处理中,一维 CNN 能够识别局部 n -gram 特征,这对于文本主题分割中的单词、短语和语句结构等识别尤为重要。如 Wang 等^[22]首次引入卷积神经网络组成语义连贯性神经网络(Semantic Coherence Neural Network, SCNN)。SCNN 模型采用对称结构的卷积网络,其构建理念基于无监督学习的核心直觉:同一主题的相邻文本片段比不同主题的片段更具连贯性。Wang 等^[23]最早探索了利用 CNN 对段落文本进行特征提取,结合 BiLSTM 提出了有监督学习的 BiLSTM-CNN 模型,证明了 CNN 在段落文本上学习表示的有效性。Badjatiya 等^[24]提出了一种基于注意力的双向 LSTM 模型。其中,使用 CNN 学习句子嵌入,并根据上下文信息,通过一层 Softmax 操作取得主题边界标记,相较于基准模型在 Fiction 和 Clinical 数据集的 WD 指标上取得了接近 7% 的提升。杜雨奇等^[25]采用图论思想针对现有文本分割模型提取文本段落结构信息、语义相关性及上下文交互等细粒度特征的不足,提出了一种基于图卷积网络(GCN)的文本分割模型 TS-GCN。卷积神经网络(CNN)在主题分割任务中在捕捉局部特征和模式方面具有强大能力,尤其适合处理具有明显局部结构的文本数据。但 CNN 并不擅长处理长距离依赖问题,一旦涉及处理长文本类型的数据(如对话主题分割场景中的跨多

轮对话的主题变化),其性能会显著下降。

2.3.2 循环神经网络(RNN)

循环神经网络(Recurrent Neural Network, RNN)是专为序列数据建模设计的一类神经网络模型^[26]。这种网络模型融合了存储系统的功能,使其能够记忆先前处理的序列信息,这是RNN区别于卷积神经网络等其他模型的关键特性。在自然语言处理(NLP)领域,RNN因其能够处理序列数据而受到青睐。在基于RNN网络开展主题分割任务的研究中,大多研究方法首选以RNN的变体——长短期记忆网络(Long Short-Term Memory, LSTM)和双向长短期记忆网络(BiLSTM)为主^[27-28]。因为,LSTM和BiLSTM相较于传统RNN网络更有优势,如LSTM在RNN的基础上引入了遗忘门、更新门和输出门机制,有效解决了RNN在训练过程中常见的梯度消失和梯度爆炸问题,从而提升了模型在处理长序列数据时的性能和稳定性。而BiLSTM通过结合前向和后向LSTM,能够同时捕捉序列数据的正向和反向依赖关系,提高了模型对序列数据的理解能力。

基于上述特性,循环神经网络(RNN)在主题分割领域受到广泛关注。首先,RNN能够自然地处理任意长度的序列数据,无需依赖固定窗口大小。在转录文本、专业领域文本等场景中,这一特点对于识别文档中不同位置出现的主题边界至关重要。如Xia等^[29]提出了采用BiLSTM模型实现并行计算的PEN-NS模型。该模型融合了循环神经网络以及前馈神经网络的分割模型,提出了一种领域平滑(Neighbor Smoothing)方法,以增强分段边缘句子的向量表示。其次,RNN模型的灵活性允许其根据不同任务需求进行调整,例如通过改变隐藏层的数量或引入长短期记忆网络(LSTM)和门控循环单元(GRU)等门控机制来增强模型的表达能力,能够适应数量级大小不同的主题分割数据集。如Li等^[30]提出了一种名为SEGBOT的端到端的文本分割模型,使用BiLSTM来编码输入文本序列,并使用另一个RNN网络连同指针网络^[31](Pointer Network)来确定输入序列中的文本边界。王炳浩^[32]提出一种基于LSTM的对话主题分割模型,使用多层LSTM来对句子和句子序列进行建模,并结合了说话人信息的表示学习以提升对话主题分割效果。再次,RNN能够累积和整合前后的上下文信息,这对于主题细微差异以及语法结构复杂的对话场景中具有重要意义。如Sheikh等^[33]利用双向长短期记忆网络模型捕获过去以及下一组单词所在文本的上下文,在文本向量表示中融入词汇特征变化,随后对比这些过去和未来的上下文中单词相似性以识别主题转换点。最后,RNN的通用性不仅可以作为文本向量化方法,还能够用于实现多任务场景应用如生成序列标签等,这对于某些需要生成主题标签的主题分割任务极具价值。Barrow等^[34]则是提出S-LSTM分割模型,首次尝试引入了分段预测和池化的结合,可以动态预测和标记片段(Segment),为预测的分段(Segment)生成单个向量,同时,在标签预测阶段引入对齐

(Alignment)技术,确保预测标签与真实标签对齐,保证模型分段的准确性,取得了可观的成绩。

尽管RNN模型具备多方面优势,但也面临着一些挑战。例如,在长序列数据场景中,容易出现梯度消失或梯度爆炸的问题,这导致模型难以学习长期依赖关系。此外,由于RNN模型的纠错能力以及动态特征捕获能力有限,在面对这类语法结构不明确、语言错误频发且主题内容宽泛、主题转变迅速的对话文本场景时,其表现并不是特别突出。

2.3.3 Transformer网络

在自然语言处理领域,Transformer^[35]网络以其自注意力机制(Self-Attention)而著称,标志着深度学习模型的一个重要里程碑。该机制的独特之处在于,其赋予模型在处理序列数据时能够捕捉序列中任意2个位置之间的依赖关系,从而显著增强了模型对长距离依赖的捕捉能力。此外,Transformer网络的稳定架构为其提供了显著的可扩展性,即随着编码器和解码器层数的增加,模型的性能也随之提升。这种正相关的增长趋势表明,通过增加模型的深度和宽度,可以有效提高模型的处理能力和准确性。因此,Transformer网络不仅在理论上具有创新性,而且在实际应用中展现出卓越的性能。

在主题分割任务中,Transformer网络由于其自注意力机制,能够并行处理序列中的所有元素,这在处理长文本时比循环神经网络(RNN)更高效。因此,对于Transformer网络的一种使用方式是延续循环神经网络模型的方法,将主题分割视为序列标记问题,但是在嵌入表示层以及特征提取层时使用Transformer网络替代RNN模型以完成更全面的文本向量表示和特征提取。如Glavaš等^[36]使用Transformer网络结构中的编码器模块完全替换TextSeg分割模型中的循环神经网络模块,提出了TLT-TS模型,在WIKI-727K、WIKI-50等数据集上的性能表现超过TextSeg模型;Lukasik等^[37]提出的BERT-BiLSTM模型在底层向量编码部分采用以Transformer编码器模块组成的BERT模型;Xing等^[38]提出添加语义相关的辅助任务和限制性自我注意,增强基于分层注意力BiLSTM网络的分割器,以更好地建模上下文,在CHOI、WikiSection数据集上的训练和测试取得了当时的SOTA成绩。

对于Transformer网络的另一种使用方式则是为模型的编码器或分类器额外添加辅助任务的训练,实现向量编码的对齐或增强模型特定能力。如Lo等^[39]在对模型的编码器部分的训练任务中加入关于句子的语义、连贯性、情感等方面的内容,以期句子不同方面的先验知识的加持下,生成更为准确、泛化性强的句子编码表示;Glavaš等^[36]和Xing等^[40]各自为模型额外增加了一个基于相邻话语的连贯性假设的语义一致性建模任务,确保分类考虑句子之间的连贯性,增加分类的准确性;Xu等^[41]在电力电网风险文本信息处理中采用BERT-BiLSTM模型结构,在BERT模

型完成向量编码后,添加了辅助任务 BIO 标签任务,以便融入 BIO 分段标签知识,从而提高分段标记的性能,相较于 RNN 及其融合模型平均提高 3~5 个百分点;Wang 等^[42]对字符展开了一种基于 Transformer 的多级文本分割方法,这是一种细粒度分割方法,将字符看作是由笔画等更小单位组成的部分,以增强在不同粒度上提取特征。在对话文本领域,Xing 等^[40]基于相邻话语的连贯性假设设计了相邻话语对的一致性评分弱监督任务,用以微调下句预测任务的 BERT 模型,即 CSM 模型;Jiang 等^[17]首次尝试利用大语言模型 InstructGPT^[43]和 ChatGPT^[44]实现 TS 任务,InstructGPT 和 ChatGPT 分别给定一个简单的提示模板,使得 InstructGPT 和 ChatGPT 可以进行对话主题分割;Gao 等^[45]提出一种名为 DialSTART 的对话主题分割模型,其在正负样本对比学习方法的基础上,通过邻近话语匹配和伪分割辅助任务的设计训练模型,从无对话主题分割标签的对话数据中学习主题感知的话语表示以进行对话主题分割。

在对话文本主题分割领域,尽管 Transformer 网络因其

卓越的性能而占据主导地位,但也面临一些挑战。首先,在 Xing 等^[38]提出的 BERT-BiLSTM 模型中,尽管 Transformer 在无监督和半监督任务中表现出色,但在训练基于句子对连贯性的辅助任务时,仍然需要大量高质量的标注数据。其次,对话场景中的模型可解释性是一个重要问题。因为,现有数据集中的对话往往是虚拟重构的,这与自然随意的对话存在明显差异^[61],意味着在实际应用中,模型可能无法准确捕捉到自然对话中的复杂性和多样性。最后,尽管基于 Transformer 的主题分割模型性能普遍较强,但它们也面临着资源消耗大和模型难收敛的问题^[53,62,63]。

综上所述,本文对文本编码层中的几种主流方法进行了深入的比较分析,以评估它们在主题分割任务中的性能和适用性。此外,为了进一步展示这些模型的实际应用效果,本文提供了它们在同一数据集上的具体表现,相关详细数据和分析结果分别如表 3~表 5 所示。研究结果表明,本文揭示了不同模型在处理文本主题分割时的优势与局限性,为未来的研究和实践提供了有价值的参考和指导。

Table 3 P_k score comparison on datasets such as Wikicities, Wikielements and WIKI-727K

表 3 Wikicities、Wikielements、WIKI-727K 等数据集上的 P_k 得分对比

(%)

模型方法		Wikicities	Wikielements	WIKI-727K	Fiction	WIKI-50	CHOI	Clinical
		P_k						
人类	-	-	-	-	-	14.97	-	-
随机划分(Random)	Random ^[14]	47.14	50.08	53.09	47.5	52.65	49.4	44.1
	TextTiling ^[46]	-	-	-	-	-	45.25	-
	C99 ^[10]	-	-	-	42.1	-	10.50	-
基于语篇结构方法	BayesSeg ^[47]	36.2	35.6	-	-	49.2	20.8	34.0
	GraphSeg ^[48]	39.95	49.12	-	-	63.56	6.64	64.6
	TopicTiling ^[49]	-	-	-	-	-	0.88	-
	TSM ^[50]	-	-	-	32.5	-	-	30.6
基于统计学习方法	PLDA ^[51]	-	-	-	36.1	-	-	32.4
	TopSeg ^[51]	-	-	-	-	-	8.22	-
	SECTOR ^[15]	33.4	42.8	-	-	28.6	-	36.9
	CNN+Attn-BiLSTM ^[24]	-	-	-	37.8	-	-	31.8
RNN 及其融合模型	TextSeg ^[14]	19.68	41.63	-	-	18.2	1.0	30.8
	S-LSTM ^[34]	21.2	30.2	-	-	22.7	-	36.1
	LSTM-LSTM-CRF ^[52]	16.1	39.4	-	-	26.8	0.54	30.5
	SEGBOT ^[30]	-	-	-	-	-	0.11	-
	TLT-TS ^[36]	19.21	20.33	19.41	-	17.47	23.26	-
Transformer 网络	CATS ^[37]	16.85	18.41	15.95	-	16.53	18.50	-
	CSB ^[37]	-	-	-	-	-	0.07 0.04	-
	GAP+HSCM+	-	-	17.92	-	15.49	22.03	26.31
	SWAGRUSeg ^[53]	-	-	-	-	-	-	-
	TS-GCN ^[25]	19.13	18.03	-	-	-	-	-
图神经网络	MHG-TS ^[54]	32.55	39.55	-	-	20.41	34.4	-

由表 3~表 5 可以看出,基于语篇结构理论、基于统计学习方法以及基于深度学习的分割方法在大多数数据集上的性能表现存在明显差异。如表 4 所示,在数据集 Wiki-Section 上,不同方法的对比结果显示,基于语篇结构理论和统计学习方法划分文本片段的准确率显著低于基于深度学习的主题分割算法。具体而言,C99、TopicTiling 算法分别在 en_city 上取得了 36.8 和 30.5 的 P_k 值,而 RNN 及其融合模型和 Transformer 网络类的方法普遍取得低于 20,基

至 10 以内的 P_k 值。在表 3 的若干数据集上,也展示出了类似的趋势。此外,表 5 显示,随着数据集规模的扩大,在动态文本的主题分割场景下,基于语篇结构和统计学习的方法与 Transformer 网络方法之间的性能差距逐渐缩小。值得注意的是,Doc2Dial 和 DialSeg711 都是针对对话场景的数据集,其中 Doc2Dial 的规模约为 DialSeg711 的 6 倍。

基于图卷积网络(Graph Convolutional Network, GCN)的 TS-GCN 模型是图卷积神经网络模型(GCN)在主题分割

Table 4 P_k score comparison of different types of models on the WikiSection dataset

基础		WikiSection			
		en_dis-	de_dis-	en_ci	de_ci
		ease	ease	ty	ty
		P_k			
基于语篇结构方法	C99 ^[10]	37.4	42.7	36.8	38.3
基于统计学习方法	TopicTiling ^[49]	43.4	45.4	30.5	41.3
	TextSeg ^[14]	24.3	35.7	19.3	27.5
	DS ^[54]	33.5	-	8.2	-
	SECTOR ^[15]	26.3	27.5	15.5	16.2
RNN 及其融合模型 算法	LSTM-LSTM-CRF ^[52]	23.9	23.6	9.7	10.2
	S-LSTM ^[34]	20.0	18.8	9.1	9.5
	PEN-NS ^[29]	19.13	17.7	8.0	8.8
	Dot-BiLSTM ^[56]	20.69	-	8.68	-
	SS ^[54]	24.8	-	9.2	-
	CSB ^[37]	21.1	28.0	9.3	11.3
Transformer 网络	Transformer ^{2[57]}	18.8	16.0	9.1	7.3
	Seq2Seq-MP ^[58]	15.1	15.7	6.5	7.0
	STP+TC+NSP ^[59]	13.7	-	4.6	-

Table 5 P_k, WD score comparison on the DialSeg711 and Doc2Dial dialogue datasets

模型名称		DialSeg711		Doc2Dial	
		P_k	WD	P_k	WD
随机划分(Random)	-	52.92	70.04	55.60	65.29
	TextTiling (TeT) ^[46]	40.44	44.63	52.02	57.42
基于语篇结构和统计 学习的方法	BayesSeg ^[47]	30.97	35.60	46.65	62.13
	GraphSeg ^[48]	43.74	44.76	51.54	51.59
	GreedySeg ^[2]	50.95	53.85	50.66	51.56
	TeT + Embedding ^[60]	39.37	41.27	53.72	55.73
	TeT + NSP ^[40]	46.84	48.50	50.79	54.86
	TeT + CLS ^[2]	40.49	43.14	54.34	57.92
	TADAM ^[2]	-	52.34	-	-
	CSM ^[40]	24.30	26.35	45.30	49.84
Transformer 网络	DialSTART ^[45]	17.86	19.80	38.11	40.72

任务上的初次尝试。该模型在 Wikielements 数据集上取得的成绩甚至超出了 CATS 和 TLT-TS 主题分割模型。其突出表现的原因主要体现在 2 个方面:①Wikielements 作为小规模数据集仅包含 118 篇文档,而基于 Transformer 网络开发的 CATS 和 TLT-TS 等模型对于数据集规模要求较高;②TS-GCN 作为一种融合模型,通过引入图的拓扑结构,对文本段落语义信息进行全局信息的融合表达,同时,结合注意力机制对邻居节点的段落信息进行有侧重性的聚合,从而有效提升了文本数据的空间性和逻辑性。此外,TS-GCN 还补充利用了文本段落之间的主题差异信息。综上,TS-GCN 模型在 Wikielements 数据集上的性能表现更好。

同时,值得关注的是部分基于 RNN 模型开发的融合模型也取得了令人满意的成绩。如表 3 所示, LSTM-LSTM-CRF 模型在 WikiCities 数据集上取得了当前最优(SOTA)

的成绩, 优于基于 Transformer 网络的 TS 模型算法;而表 4 中, S-LSTM 模型和 PEN-NS 模型也同样展现了这种潜力。这些模型的共同特点在于,它们都在一定程度上融合了其他模型方法。例如, LSTM-LSTM-CRF 模型结合了 RNN 和传统的条件随机场(Conditional Random Field, CRF)方法, 而 PEN-NS 模型则在编码层面融合了 BiLSTM、注意力机制以及前馈神经网络(FFN)。这些模型的成功表明,融合模型在主题分割任务中具有巨大的潜力和应用价值。

此外, 基于 Transformer 网络的模型在 WikiSection 数据集上展现了显著的性能, 并且多任务辅助方式成为常用手段。如 Seq2Seq-MP 模型, 其借鉴了 SEGBOT 模型的设计理念, 将句子向量的编码/解码过程与分段点的标记任务解耦, 并在分段点标记任务中集成了多个辅助任务, 强化了主题与分割点之间的关联, 从而在 WikiSection 数据集上实现了优异的主题分割性能;如表 5 所示, DialSTART 模型也在序列编码层和标签标记层, 分别设计了主题感知建模和相邻话语匹配 2 个辅助任务。

如表 5 所示, 在对话文本场景的主题分割任务性能评估中, 基于 Transformer 网络开发的模型表现出了明显的优势。这种对话具有动态交互的特性, 同时可能包含多个不同主题。相比之下, CNN 和 RNN 在处理这类对话时面临较大挑战, 难以有效完成主题分割任务。因此, 目前在对话领域的主题分割模型算法研究范式通常是基于 Transformer 网络的预训练模型, 以设计多辅助任务微调为主。

综上, 在深度学习技术的加持下, 文本主题分割任务的性能不断突破, 其对数据集的要求也从小规模逐步扩展至大规模, 且所覆盖的数据内容范围也在不断扩大。尤其是在以 Transformer 网络为基础开发的分割模型中, 这一变化尤为明显。

3 研究趋势

在主题分割任务早期研究中, 基于统计学习及传统机器学习方法的文本分割技术在软件及硬件的限制条件下取得了一定成就, 但是在开放领域的文档处理以及细分领域上仍然面临严重挑战。而随着深度学习技术的发展, 如卷积神经网络、循环神经网络和 Transformer 网络等新技术冲击下, 主题分割任务发生了一些新的变化, 也出现了新的问题。具体如下:

(1) 文本片段的主题语义相似性成为现代深度学习技术中主题分割任务的重要依据。过去的研究人员在文本分割任务的研究中, 常常倾向于根据文体和任务要求将字、词或句子等可具象化的单位作为划分粒度的工具。而随着深度学习技术的崛起, 更多情况下, 即便是不同文体和不同任务中, 优先也会以文本片段的主题语义这一抽象概念作为划分文本的依据。如以相邻话语的连贯性假设构建的 SCNN 模型、层次化 TextSeg 模型以及 Transformer 网

络架构下的分割模型 CATS、TLTTS 等,均按照文本主题语义的不同分割文本片段,并且对于模型的训练方法往往也以文本主题语义的向量比较为主。

(2)主题分割模型的泛化性成为重要考量。在以统计学习和传统机器学习方法作为主流方法的背景下,大多数研究人员在研究过程中仅仅考虑单一领域或特定文体类型的文本分割算法,如新闻文本分割、转录文本分割和议论文论据挖掘等。这些算法普遍只能解决特定领域的问题。深度学习技术是一种基于数据驱动的算法,其构建的神经网络模型通过训练数据调整后,能够具备跨领域的任务感知能力,因此泛化性始终是该技术的关键评价指标。随着深度学习技术被引入主题分割任务领域后,泛化性也随之成为了衡量主题分割模型的重要指标。

(3)数据集匮乏问题仍然没有改善,甚至更为严重。过往的传统算法如 TextTiling 算法、C99 算法、贝叶斯分割算法等受制于标注数据集昂贵且匮乏的情况,普遍选择了无监督方式来避免这一问题。虽然随着深度学习技术的普及,这一情况有所改善,但是带标注数据集匮乏仍然是一个无法避免的问题,并且由于深度学习技术对于数据集的要求比过去更高,使得这一问题在现阶段显得更为严峻。在某些领域特定的场景中,如针对聊天对话文本的主题分割任务,数据获取面临着不少挑战。这些领域的数据不仅稀缺,而且现有数据集的质量普遍存在问题。许多数据集是通过人机对话系统生成的,如斯坦福数据集、Multi-WOZ 语料库和 Doc2Dial 等,这些数据集与实际应用场景存在一定偏差。

(4)多模型融合已成为文本主题分割领域的一种重要发展趋势。在单一模型与多模型融合的分割性能对比中,多模型融合通常展现出显著的优势,有效提升了分割任务的性能。目前,预训练模型在文本主题分割任务中占据主导地位。在此基础上,将传统相似性判别算法与不同架构模型的优势相结合,不仅能够充分发挥各模型的特性,还能进一步提升整体的分割性能。然而,多模型融合策略和权重分配的设计需要精细化考量,以适应不同类型文本数据集的特点和特定任务的需求。若不考虑这些因素而盲目追求模型融合,可能会产生负面效果,甚至降低分割性能。

(5)对跨语言的文本主题分割探索。跨语言的文本主题分割对于数据和模型的设计要求较高,它可能涉及复杂的上下文和语义理解问题。为了有效处理跨语言文本中的主题分割任务,模型必须具备深入理解复杂上下文和语义的能力,而不同语言的上下文线索和语义表达方式差异显著,这要求模型不仅要能够识别和解析语言的表面结构,还要能够捕捉语言背后深层的文化内涵和语境含义。因此,模型需要在训练过程中学习捕捉各种语言中的主题切换信号,例如特定的词汇、短语或语法结构,这些信号在不同语言中呈现出不同的形式。此外,在处理语言中的歧

义现象时,模型还需要能够理解并利用上下文中的隐含信息。

4 结语

本文深入探讨了文本主题分割任务的研究背景、最新研究进展,以及深度学习技术在该领域的应用现状。研究成果不仅为文本主题分割领域提供了一个全面的概述,而且指出了未来研究可以进一步探索多模态数据的融合,结合文本、图像和其他媒体类型,以增强模型对主题变化的理解和识别能力。此外,开发更加有效的多模型融合方法、特征提取方法以及多任务设计方法,同时整合领域知识和先验信息,也将有助于进一步提升主题分割任务的性能。

值得期待的是,深度学习驱动的主题分割技术在多个领域展现出巨大潜力,包括信息检索、内容推荐和人机交互等。这些技术的应用不仅能显著提升信息处理的效率和准确性,还能极大地优化用户体验。它们在商业智能、健康医疗和教育等关键领域的重要作用不容忽视。同时,随着这些技术的不断发展,新的隐私保护和数据安全挑战仍然存在,这对社会的影响将是深远和复杂的。因此,在推动技术进步的同时,需要认真考虑并妥善解决由此产生的社会问题。

综上所述,文本主题分割技术的发展将为人工智能领域带来新的机遇和挑战,为社会带来更多的创新和价值。

参考文献:

- [1] MICULICICH L, HAN B. Document summarization with text segmentation [DB/OL]. <https://arxiv.org/pdf/2301.08817>.
- [2] XU Y, ZHAO H, ZHANG Z. Topic-aware multi-turn dialogue modeling [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 14176-14184.
- [3] BU Q, PARK S, KHANG M, et al. SRFormer: text detection transformer with incorporated segmentation and regression [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2024: 855-863.
- [4] GU C W. Research on segmentation and recognition of criminal suspects' dialogue topics based on deep learning [D]. Nanning: Guangxi Minzu University, 2023.
古长汶. 基于深度学习的犯罪嫌疑人对话主题分割与识别方法研究 [D]. 南宁: 广西民族大学, 2023.
- [5] MANN W C, THOMPSON S A. Rhetorical structure theory: toward a functional theory of text organization [J]. Text-interdisciplinary Journal for the Study of Discourse, 1988, 8(3): 243-281.
- [6] GROSZ B J, SIDNER C L. Attention, intentions, and the structure of discourse [J]. Computational Linguistics, 1986, 12(3): 175-204.
- [7] HALLIDAY M A K, HASAN R. Cohesion in English [M]. London: Routledge, 2014.
- [8] PAK I, TEH P L. Text segmentation techniques: a critical review [M]. Berlin: Springer, 2018: 167-181.
- [9] SHI J. Text segmentation overview [J]. Computer Engineering and Applications, 2006(35): 155-159, 171.

- 石晶. 文本分割综述[J]. 计算机工程与应用, 2006(35): 155-159, 171.
- [10] CHOI F Y Y. Advances in domain independent linear text segmentation [DB/OL]. <https://arxiv.org/pdf/cs/0003083>.
- [11] CLARK V L, KRUSE J A. Clinical methods: the history, physical, and laboratory examinations[J]. *Jama*, 1990, 264(21): 2808-2809.
- [12] CHEN H, BRANAVAN S R K, BARZILAY R, et al. Global models of document structure using latent permutations[C]//Association for Computational Linguistics, 2009: 371-379.
- [13] KAZANTSEVA A, SZPAKOWICZ S. Linear text segmentation using affinity propagation[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011: 284-293.
- [14] KOSHOREK O, COHEN A, MOR N, et al. Text segmentation as a supervised learning task[DB/OL]. <https://arxiv.org/pdf/1803.09337>.
- [15] ARNOLD S, SCHNEIDER R, CUDRÉ-MAUROUX P, et al. SECTOR: a neural model for coherent topic segmentation and classification[J]. *Transactions of the Association for Computational Linguistics*, 2019, 7: 169-184.
- [16] FENG S, WAN H, GUNASEKARA C, et al. Doc2dial: a goal-oriented document-grounded dialogue dataset [DB/OL]. <https://arxiv.org/pdf/2011.06623>.
- [17] JIANG J, DONG C, KUROHASHI S, et al. SuperDialseg: a large-scale dataset for supervised dialogue segmentation[DB/OL]. <https://arxiv.org/pdf/2305.08371>.
- [18] BEEFERMAN D, BERGER A, LAFFERTY J. Text segmentation using exponential models[DB/OL]. <https://arxiv.org/pdf/cmp-lg/9706016>.
- [19] PEVZNER L, HEARST M A. A critique and improvement of an evaluation metric for text segmentation[J]. *Computational Linguistics*, 2002, 28(1): 19-36.
- [20] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [DB/OL]. <https://arxiv.org/pdf/1301.3781>.
- [21] PENNINGTON J, SOCHER R, MANNING C. GloVe: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1532-1543.
- [22] WANG L, LI S, LÜ Y, et al. Learning to rank semantic coherence for topic segmentation[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 1340-1344.
- [23] WANG L, LI S, XIAO X, et al. Topic segmentation of web documents with automatic cue phrase identification and BLSTM-CNN[C]//Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, 2016: 177-188.
- [24] BADJATIYA P, KURISINKEL L J, GUPTA M, et al. Attention-based neural text segmentation [C]//European Conference on Information Retrieval, 2018: 180-193.
- [25] DU Y Q, ZHENG J, WANG Y, et al. Text segmentation model based on graph convolutional network [J]. *Journal of Computer Applications*, 2022, 42(12): 3692-3699.
- 杜雨奇, 郑津, 王杨, 等. 基于图卷积网络的文本分割模型[J]. 计算机应用, 2022, 42(12): 3692-3699.
- [26] HOPFIELD J J. Neural networks and physical systems with emergent collective computational abilities[J]. *Proceedings of the National Academy of Sciences*, 1982, 79(8): 2554-2558.
- [27] CHENG J. Long short-term memory-networks for machine reading [DB/OL]. <https://arxiv.org/pdf/1601.06733>.
- [28] CHO K. Learning phrase representations using RNN encoder-decoder for statistical machine translation [DB/OL]. <https://arxiv.org/pdf/1406.1078>.
- [29] XIA J, LIU C, CHEN J, et al. Dialogue topic segmentation via parallel extraction network with neighbor smoothing [C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022: 2126-2131.
- [30] LI J, SUN A, JOTY S R. SegBot: a generic neural text segmentation model with pointer network [C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018: 4166-4172.
- [31] VINYS O, FORTUNATO M, JAITLEY N. Pointer networks [J]. *Advances in Neural Information Processing Systems*, 2015, 28: 2692-2700.
- [32] WANG B H. Research on topic segmentation techniques in dialogue text [D]. Harbin: Harbin Institute of Technology, 2016.
- 王炳浩. 面向对话文本的主题分割技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2016.
- [33] SEHIKH I, FOHR D, ILLINA I. Topic segmentation in ASR transcripts using bidirectional RNNs for change detection [C]//2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2017: 512-518.
- [34] BARROW J, JAIN R, MORARIU V, et al. A joint model for document segmentation and segment labeling [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 313-322.
- [35] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. *Advances in Neural Information Processing Systems*, 2017, 30: 5998-6008.
- [36] GLAVAŠ G, SOMASUNDARAN S. Two-level transformer and auxiliary coherence modeling for improved text segmentation [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 7797-7804.
- [37] LUKASIK M, DADACHEV B, SIMOES G, et al. Text segmentation by cross segment attention [DB/OL]. <https://arxiv.org/pdf/2004.14535>.
- [38] XING L, HACKINEN B, CARENINI G, et al. Improving context modeling in neural topic segmentation [DB/OL]. <https://arxiv.org/pdf/2010.03138>.
- [39] LO K, JIN Y, TAN W, et al. Transformer over pre-trained transformer for neural text segmentation with enhanced topic coherence [DB/OL]. <https://arxiv.org/pdf/2110.07160>.
- [40] XING L, CARENINI G. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring [DB/OL]. <https://arxiv.org/pdf/2106.06719>.
- [41] XU S, LIU H. Power grid risk text segmentation method based on BERT neural network [C]//Ninth International Conference on Energy Materials and Electrical Engineering, 2024: 1558-1565.
- [42] WANG X, WU C, YU H, et al. Textformer: component-aware text segmentation with transformer [C]//2023 IEEE International Conference on Multimedia and Expo, 2023: 1877-1882.
- [43] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback [J]. *Advances in Neural Information Processing Systems*, 2022, 35: 27730-27744.
- [44] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report [DB/OL]. <https://arxiv.org/pdf/2303.08774>.
- [45] GAO H, WANG R, LIN T E, et al. Unsupervised dialogue topic segmentation with topic-aware utterance representation [DB/OL]. <https://arxiv.org/pdf/2305.02747>.
- [46] HEARST M A. Text tiling: segmenting text into multi-paragraph subtop-

- ic passages[J]. Computational linguistics, 1997, 23(1): 33–64.
- [47] EISENSTEIN J, BARZILAY R. Bayesian unsupervised topic segmentation[C]//Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008: 334–343.
- [48] GLAVAS G, NANNI F, PONZETTO S P. Unsupervised text segmentation using semantic relatedness graphs[C]//Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics, 2016: 125–130.
- [49] RIEDL M, BIEMANN C. TopicTiling: a text segmentation algorithm based on LDA[C]//Proceedings of ACL 2012 Student Research Workshop, 2012: 37–42.
- [50] DU L, BUNTINE W, JOHNSON M. Topic segmentation with a structured topic model[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics, 2013: 190–200.
- [51] PURVER M, KÖRDING K P, GRIFFITHS T L, et al. Unsupervised topic modelling for multi-party spoken discourse[C]//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, 2006: 17–24.
- [52] LAMPLE G. Neural architectures for named entity recognition[DB/OL]. <https://arxiv.org/pdf/1603.01360>.
- [53] ZHAO B. Research on topic segmentation based on deep learning. [D]. Qinhuangdao: Yanshan University, 2023.
赵博. 基于深度学习的主题分割研究[D]. 秦皇岛:燕山大学, 2023.
- [54] ZHANG Y N. Research on application of topic segmentation technology based on pre-trained language model in social short text[D]. Urumqi: Xinjiang Agricultural University, 2023.
张洋宁. 基于预训练语言模型的话题分割技术在社交短文本的应用研究[D]. 乌鲁木齐:新疆农业大学, 2023.
- [55] INAN H, RUNGTA R, MEHDAD Y. Structured summarization: unified text segmentation and segment labeling as a generation task[DB/OL]. <https://arxiv.org/pdf/2209.13759>.
- [56] GHINASSI I, WANG L, NEWELL C, et al. Lessons learnt from linear text segmentation: a fair comparison of architectural and sentence encoding strategies for successful segmentation[C]//Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, 2023: 408–418.
- [57] LO K, JIN Y, TAN W, et al. Transformer over pre-trained transformer for neural text segmentation with enhanced topic coherence[DB/OL]. <https://arxiv.org/pdf/2110.07160>.
- [58] XIA J, WANG H. A sequence-to-sequence approach with mixed pointers to topic segmentation and segment labeling[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023: 2683–2693.
- [59] LEE J, HAN J, BAEK S, et al. Topic segmentation model focusing on local context[DB/OL]. <https://arxiv.org/pdf/2301.01935>.
- [60] SONG Y, MOU L, YAN R, et al. Dialogue session segmentation by embedding-enhanced texttiling[DB/OL]. <https://arxiv.org/pdf/1610.03955>.
- [61] DECKER A, AMBLARD M. With a little help from my (linguistic) friends: topic segmentation of multi-party casual conversations[DB/OL]. <https://arxiv.org/pdf/2402.02837>.
- [62] BRANTS T, CHEN F, TSOCHANTARIDIS I. Topic-based document segmentation with probabilistic latent semantic analysis[C]//Proceedings of the Eleventh International Conference on Information and Knowledge Management, 2002: 211–218.
- [63] SOLBIATI A, HEFFERNAN K, DAMASKINOS G, et al. Unsupervised topic segmentation of meetings with BERT embeddings[DB/OL]. <https://arxiv.org/pdf/2106.12978>.

(责任编辑:陈维捷)