大规模多语言多模态摘要数据集 t

Yash Verma*加尔各答
印度科学教育与研究学院
yashv7523@gmail.com

Anubhav Jangra*印度理工学院巴特那分校
anubhav0603@gmail.com

Raghvendra Kumar
印度理工学院巴特那分校
raghvendra.kumar1004@gmail.com

Sriparna Saha
印度理工学院巴特那分校
sriparna@iitp.ac.in

arXiv:2302.06560v1 [cs.CL] 13 Feb 2023

摘要

编码器-解码器模型等技术的显著进展，使我们能够表示包含多种模态的信息。这些信息还能进一步提升信息检索和自然语言处理领域中的许多下游任务；然而，多模态技术及其性能评估的改进需要大规模且足够多样化的多模态数据。对于多模态摘要、文本生成和翻译等各类任务，多语言建模会利用从高质量多语言标注数据中获得的信息。在这项工作中，我们推出了目前规模最大的多语言多模态摘要数据集（M3LS），其中包含超过一百万个文档-图像对实例，并为每一对提供了经专业标注的多模态摘要。该数据集源自英国广播公司（BBC）十年来发布的新闻文章，涵盖20种语言，旨在覆盖五种语系的多样性；它也是涵盖13种语言的最大摘要数据集，并包含2种语言的跨语言摘要数据。我们利用该数据集正式定义了多语言多模态摘要任务，并报告了多种最先进摘要技术在多语言环境下的基线分数。我们还将其与许多类似数据集进行比较，以分析M3LS的独特性和难度。 [1]

表1：所提出的数据集与现有大规模多语言多模态数据集的比较。

多语言摘要数据集			
数据集名称	数据集规模	#语言数量	领域
XL-Sum (Hasan 等人,2021)	1M	44	News
MLSUM(Scialom 等, 2020年)	1.5M	5	News
WikiLingua (Ladhak 等, 2020年)	770K	18	教程
MLGSum(Wang 等人，2021)	1.1M	12	News
M3LS（我们的方法）	1.1M	20	News
多模态摘要数据集			
数据集名称	**数据集规模**	**模态**	**领域**
MSMO (Zhu 等人，2018年)	314K	文本 +图像	News
E-DailyMail(Chen和Zhuge，2018年)	219K	文本 +图像	News
How2 (Sanabria 等人，2018年)	190K	文本 +视频 + 音频	多个领域
MMSS (Li 等人，2018年)	66K	文本 +图像	News
VMSMO(Li 等人，2020年)	185K	文本 + 视频 +音频	社交网络
M3LS（我们的方法）	1.1M	文本 +图像	News

1 引言

如今我们生活的世界非常多元，全球使用的语言数量超过 7,000+ 种[2]。这些语言各有不同的特点，使用者社群的规模也各不相同，取决于这种语言的普及程度。例如，普通话由超过 50,000 个汉字 （字符）组成，使用者超过 11.17 亿人[3],；而像 Rotokas语 这样的语言则不同，它是巴布亚新几内亚布干维尔岛上约 4,320 人使用的土著语言，仅由 12 个字母组成[4]。

尽管这些语言至关重要，但它们限制了人们只能与使用同一种语言的人交流思想。然而，视觉是地球上的每个人都普遍拥有的能力，无论其文化、族裔或所使用的语言是什么。通过这项工作，我们旨在利用多种语言和视觉模态的信息，推动改进现有自动摘要系统的研究。

过去的各种研究表明，与单语言摘要框架相比，跨多种语言的统一摘要框架能够提升摘要质量（Wang 等人，2021）。同样，也有研究表明，与文本摘要系统相比，多模态输入有助于提升多模态摘要的质量（Jangra 等人，2020a，b；Chen和Zhuge，2018年；Mukherjee 等，

*这些作者对本研究的贡献相同。[1]本研究使用的数据集和代码可在以下地址获取：https://github.com/anubhav-jangra/M3LS。[2] https://www.ethnologue.com/guides/how-many-languages

[3]https://www.berlitz.com/en-uy/blog/most-spoken-languages-world[4] https://en.wikipedia.org/wiki/Rotokas_语言

2022）。此外，在摘要输出中包含多种模态有助于提升用户的整体满意度（Zhu 等人，2018年；Jangra 等人，2021b）。多种模态也可以弥补单一模态无法表达摘要各个方面的不足。例如，“自由”“重力”等抽象概念很难通过图像表达，而用文本表达则很方便。类似地，要向一个从未见过穿山甲的人描述穿山甲，也非常困难。

因此，在这项工作中，我们提出了多模态多语言摘要（M3LS）任务，并发布了 M3LS 数据集[5]，以推动这一方向的研究。该数据集包含 110 万篇新闻文章，涵盖20种语言，包括英语、中文、西班牙语、俄语、法语、乌克兰语、葡萄牙语、日语、泰米尔语、印地语、马拉地语、古吉拉特语、孟加拉语、僧伽罗语、乌尔都语、普什图语、印度尼西亚语、泰卢固语、旁遮普语和尼泊尔语；它是目前覆盖语言范围最广的摘要数据集。据我们所知，所提出的数据集是以下13种语言的最大摘要数据集（俄语、乌克兰语、泰米尔语、印地语、马拉地语、古吉拉特语、孟加拉语、僧伽罗语、乌尔都语、普什图语、泰卢固语、旁遮普语和尼泊尔语）。

我们希望，所提出的任务和数据集能够激发并推动针对较少探索语言的多模态和多语言研究，以解决各种任务，包括但不限于自动摘要（Nallapati 等人，2016；See等人，2017）、文章标题生成（Jin et al., 2020; Gavrilov et al., 2019; Zhang 等人，2018）、关键词提取（Showrov and Sobhan, 2019; Lee and Kim, 2008; Yao et al., 2019）、图像描述生成（Xu et al., 2015; Bai and An, 2018）、多模态嵌入生成（Sun et al., 2019; Lu et al., 2019; Li 等人，2019; Zhou etal., 2020）、大规模语言建模（Raffel 等人，2020；Devlin等人，2019）等。

本研究的主要贡献如下：*1)* 我们提出了多模态多语言摘要（M3LS）任务。*2)* 我们发布了涵盖*20*种语言的最大规模多模态摘要数据集。*3)* 所提出的数据集是面向*13*种语言的最大规模文本摘要数据集。*4)* 据我们所知，我们首次推出了多模态跨语言数据集（包含日译英和英译日）。*5)* 我们提供了该数据集的多模态摘要基线结果，并对数据集进行了详细分析。

2 相关工作

文本摘要领域已有五十多年的历史（Edmundson，1969），近年来取得了长足发展。在序列到序列框架（Sutskever等人，2014）取得进展之前，人们主要关注抽取式摘要技术，旨在通过提取词语、短语或句子来生成摘要（Mihalcea 和 Tarau，2004；Saini等人，2019；Alguliev等人，2010）。See等人（2017）提出了指针生成网络，这是一种基于注意力循环神经网络的框架（Bahdanau等人，2015）。近年来，借助基于 Transformer 的模型（Zhang 等，2020；Devlin等人，2019）和注意力机制（Vaswani等人，2017），自动摘要研究取得了重大进展。

本节将介绍展示多模态数据集和多语言数据集的相关研究。表中详细比较了这些数据集与M3LS 的规模，见表1。

2.1 多模态摘要数据集

多模态摘要是对包含两种或更多输入模态的内容进行摘要的任务。根据任务不同，输出可以是单模态，也可以是多模态。本节将讨论学界提出的现有大规模多模态摘要数据集。关于全面的综述，请参阅Jangra 等人（2021a）。

MSMO：Zhu 等人（2018年）提出了一个由文本和图像构成的多模态摘要数据集。该数据集取自DailyMail[6] 网站，包含 314,581 个英语实例。然而，Hasan 等人（2021）指出，DailyMail 新闻摘要缺乏新颖n元语法。Fabbri等人（2021）也指出，CNN/DailyMail 数据集中的部分参考摘要质量不一致（Nallapati 等人，2016年）。

E-Dailymail Chen和Zhuge (2018年) 提出了 E-Dailymail 数据集，其中包含从DailyMail 网站提取的文本和图像。该数据集包含 219,100 个英语实例，

[5]我们的数据集样例可在此处获取：https://github.com/zenquiorra/M3LS，完整数据集将在本文的终稿中发布

[6]https://www.dailymail.co.uk/home/index.html

其中包含输入文档、文章标题、图像和图像说明文字。

How2： Sanabria 等人(2018年）提出了一个包含文本、视频和音频模态的多模态摘要数据集，其中包含超过 2000 小时的视频，以及相应的音频和语音转录文本。

MMSS：Li 等人(2018年）提出了一个由文本和图像组成的多模态摘要数据集，旨在提出一种图像辅助句子摘要框架。该数据集包含 6.6 万个英语实例，这些实例是从 Gigaword 语料库[7]。

VMSMO：据我们所知，Li 等人(2020年）提出了首个大规模异步文本—音频—视频摘要数据集。该数据集取材于知名微博平台新浪微博[8]，，包含 184,920 个中文实例。

在问答（Singh 等人，2021）、翻译（Elliott 和 Kádár，2017）、情感分析（Soleymani 等人，2017）、词汇语义分类（Jha 等人，2022）、关键词提取（Verma 等人，2022）等任务中，也可以观察到在语言任务中融合多种模态的类似趋势。

2.2 多语言文本摘要数据集

过去几年，研究不同语言摘要的益处以提升摘要质量越来越受到关注。双语设置下已有大量研究工作；然而，为了简洁起见，本文仅讨论多语言摘要数据集。

MLSUM： Scialom 等（2020年）提出了 MLSUM 数据集，该数据集包含从DailyMail/CNN 网站获取的 150 万篇新闻文章。该数据集涵盖五种语言——法语、德语、西班牙语、俄语和土耳其语。

XL-Sum： Hasan 等人 (2021) 提出了 XL-Sum 数据集，其中包含从 BBC 新闻获取的 44种语言的 135 万篇文章，使其成为迄今为止语言种类最多的摘要数据集。然而，这 44 种语言中有 25 种甚至没有 10,000 个实例，因此不适合用于训练任何语言模型。

WikiLingua： Ladhak 等 (2020年) 提出了Wikilingua 数据集，这是迄今为止最大的平行多语言摘要数据集。该数据集包含 77 万个英语实例，并扩展到另外 17 种语言，每种语言所涵盖的英语文章数量各不相同。

MLGSum： Wang 等人 （2021）提出了 MLGSum数据集，该数据集包含来自各种新闻提供商的文章，例如 英国广播公司、 France 24 和 Süddeutsche Zeitung。该数据集包含五种高资源语言和七种低资源语言，共有 110 万个实例；其中德语有 50 万个实例，是文本摘要的丰富资源。

我们发现，多模态摘要和多语言摘要领域中的多个热门数据集（见表1）对技术评估和技术改进都很有用。然而，多语言多模态摘要这一交叉领域仍基本未被探索，这可归因于缺乏专门的高质量数据集，也缺乏将其形式化为问题陈述的工作。因此，我们正式定义了 M3LS 任务，并进一步讨论用于解决该问题的数据集。

3 M3LS任务

对于每种语言 $l_k \in L$，其中 L 是所有语言的集合，我们有数据 $M^{l_k} =< T^{l_k},\ I^{l_k} >$，其中 $T^{l_k} = \left\{t_1^{l_k},\ t_2^{l_k},\ .....,\ t_{|T|}^{l_k}\right\}$ 是一组文档，$I^{l_k} = \{I^{t_1^{l_k}}, I^{t_2^{l_k}}, \ldots, I^{t_{|T|}^{l_k}}\}$ 是一组图像，其中 $I^{t_j^{l_k}} = \left\{i_1,\ i_2,\ .....,\ i_{|I|}\right\}^{t_j^{l_k}}$ 表示属于文档 $t_j^{l_k} \in T^{k_k}$ 的图像集合，而 $|.|$ 表示集合的基数。

任务是获得一个函数 F，该函数将语言 l_{k_1} 的文档 $t_j^{l_{k_1}} \in T^{l_{k_1}}$ 及其对应的图像 $I^{t_j^{l_{k_1}}} \in I^{l_{k_1}}$ 映射为目标语言 l_{k_2} 的一组多模态摘要，其中包括文本摘要（记为 $O^{l_{k_2}}$）以及来自输入的图像（记为 $I^{l_{k_1}}$）。

$$F :< T^{l_{k_1}}, I^{l_{k_1}} >\rightarrow< O^{l_{k_2}}, I^{l_{k_1}} > \quad (1)$$

当 $k_1 \neq k_2$时，我们进行多模态跨语言摘要；否则，该任务就是多模态单语言摘要。该任务的图示见图1。

4 M3LS 数据集

通过 M3LS 任务，我们通过研究以下内容，说明了构建多模态多语言数据集的必要性：

[7] https://github.com/harvardnlp/sent-summary
[8] http://ir.weibo.com/



图1：提出的 M3LS任务。

摘要技术的发展，例如使用图像进行二次增强并生成多模态输出（Zhu 等人，2018年）、基于视频的多模态摘要（Li 等人，2020年）以及采用多目标优化（Jangra 等人，2020b）。另一方面，诸如 Xue 等人 （2020年）提出的多语言 Transformer 模型也已公开提供检查点，这些检查点针对多种语言建模任务进行了微调，其中包括多语言摘要。

开发此类模型需要高质量的异构数据，并改进各种模型，以利用多模态共享注意力层，为特定语言任务的标注数据提供图文对。为解决这些问题，我们提出了 M3LS，并在本节中讨论其构建过程中涉及的各个步骤。

4.1 数据集构建

我们研究新闻领域，因为它是内容最丰富、最容易获取的领域之一，涵盖多个主题的文章，既报道事件，又不带有极端偏见。我们分析了文章的结构，并调查了多家新闻提供商，最终选定 BBC 新闻。它以统一的结构化格式提供多种语言的完整句子摘要。这些摘要由作者专业撰写，从而确保了数据质量。我们将说明创建 M3LS 数据集所涉及的步骤，并讨论数据的各个方面。

BBC 新闻：BBC 新闻[9] 是英国广播公司旗下负责收集和播报时事新闻的部门。每篇 BBC 新闻文章都附有由完整句子组成的文本摘要，使用现在时态，避免表达观点和渲染煽情内容。我们涵盖 20 种不同语言，并以相应语言撰写摘要。我们从多个部分提取数据



图 2：M3LS 开发过程中使用的网页快照格式，以及抓取过程中提取的各项特征

如图2所示的网页

获取文章：我们从各 BBC 语言新闻数据集对应的 Twitter[10] 页面获取文章链接。为扩充数据集，我们会抓取[11] 从各语言已解析文章中获取的有效链接[12] 。

最终链接集合使用 Scrapy[13] 分别进行抓取，以获取最终数据集。由于这些链接展示在相应的 Twitter 页面上，因此能够确保文章涵盖感兴趣且广受欢迎的主题，

[9]https://www.bbc.com/新闻

[10]https://twitter.com/bbc[11]数据的收集遵循网站上提及的条款和条件[12]如果链接包含对应领域的 BBC 文章摘要，则该链接有效。[13]https://scrapy.org

并通过从网页中的推荐内容和超链接递归提取链接来扩展数据集。

数据结构化：我们从网页中提取了如图（2）所示的各种特征，并将其整理为JSON 格式；我们还提供了专用解析器、使用说明和教程，方便从任意实例中获取特征。数据可根据 BBC 新闻的条款和条件免费使用，相关细节已在上传数据集的同一链接页面中说明。

文本验证：为了确保来源文本的高质量，我们从收集到的链接中手动阅读了每种语言的 10 个实例 [14]，以核实文章是否具有描述性，以及是否由完整句子构成。

摘要验证：我们从数据集中抽取了 4 种语言[15]的各 100 篇文章，手动检查摘要质量，并验证所给摘要是否涵盖文本中呈现的信息。对于每篇文章，在仔细阅读文本后，我们会给黄金摘要打 1 至 5 分，5 分代表最佳摘要，即最充分地涵盖文章中的重要信息，反之亦然；评分时还会考虑摘要长度和文章长度等参数。我们的分析发现，所评估的各语言中，>70 篇文章的得分为$>4/5$。鉴于 BBC 在多个领域发布的文章具有一致性，我们认为这一事实适用于数据集中的每种语言。

最终数据集：最终数据集中，每篇新闻文章都包含文本文件、配有相应说明文字的图像、关键词、相关新闻文章的链接，以及由若干句子和一张图像组成的多模态摘要。

跨语言数据集：我们的跨语言数据集包含最终数据集中的所有特征，以及由另一种语言的文本构成的多模态摘要。该数据集根据作者在日语文章中提供的链接，找到对应的英语文章后构建。我们使用谷歌翻译手动核查了两篇文章中的信息[16]并抽取 100 个实例，以验证内容和所提供摘要的相似性。

训练集-测试集-验证集划分：该数据集包含 120 万篇新闻文章。对于拥有 $\leq 50{,}000$ 篇文章的语言，我们将其按 80% 作为训练集、10% 作为测试集和 10% 作为验证集进行划分；对于其他语言，则选取 90% 的数据作为训练集、5% 作为测试集、5% 作为验证集。

5 数据集分析

5.1 概述

该M3LS 数据集包含 111 万个+涵盖 20 种语言的多语言多模态实例，以及超过 9,000 个英语-日语语言对的跨语言多模态实例。该数据集可分为 8 种高资源语言和 12 种低资源语言[17]（更多详情请参阅附录B）。所选语言来自全球不同地区，属于 5 个不同的语系：印欧语系、南岛语系、日本语系、达罗毗荼语系和汉藏语系。

M3LS 数据集具有很强的多样性，其中文章数量最少的是 僧伽罗语 （10,148），文章数量最多的是 英语 （376,367）。不同语言的输入文档大小各不相同，使该数据集变得更加复杂且更具挑战性，其中˜文档大小从330个词元到超过 2800+ 词元不等。数据集中的文章涵盖了较长的时间跨度，时间范围从2009年至2021年（详见附录A ）。

我们希望 M3LS 数据集能够推动并激发对较少研究语言的研究，因为该数据集涵盖的20种语言中，有14种属于世界上使用人数最多的前20种语言[18]；这种多样性有助于为研究充分和较少研究的语言开展建模任务。

5.2 数据集比较

为了研究我们数据集的规模和覆盖范围，我们将 M3LS 与从 BBC 新闻领域提取的其他摘要数据集进行了比较。我们发现，XSum 包含了我们数据集中 53% 的词元，而 XL-Sum 在 M3LS 涵盖的所有语言中包含了我们数据集中 58% 的词元。然而，它们本质上都是单模态的，而 XSum 则是单语言的。我们观察到，与 XSum 相比，M3LS 的规模大了几个数量级；与 XL-Sum 相比，M3LS 在几乎所有单独的语言类别中都超出 2 至 3 倍。

[14]对于作者不熟悉的语言，我们使用谷歌翻译 https://translate.google.com将内容翻译成英语[15]由于本文作者对相关语言的理解能力有限，我们仅使用 4 种语言（英语、印地语、孟加拉语和马拉地语）[16] https://google.com/translate

[17]该分类基于5万条数据实例的阈值。[18] https://lingua.edu/the%2D20%2Dmost%2Dspoken%2Dlanguages%2Din%2Dthe%2Dworld%2Din%2D2022/

这两个数据集都用于训练和微调多种最先进的摘要模型，例如 Pegasus。因此，我们认为，凭借更大的规模和更多样的内容，M3LS 能为其中涵盖的语言提供更广泛、更优质的语言建模支持，同时还具备多模态这一额外优势。

6 实验

6.1 设置

根据 M3LS 中各语言的实例数量，我们采用 80:10:10 的比例划分训练集、测试集和验证集（当实例数量低于 50K 时），否则采用 90:5:5 的比例。为在多语言设置下开展实验，我们调研了适用于多种语言的公开分词器和句子分割器，并将它们整合到一个专门的软件包中用于实验。此外，我们还为该软件包中缺少外部软件包支持的语言制定了一套句子分割规则[19]。

我们使用 segtok[20] 处理印欧语系语言，使用 IndicNLP[21] 处理印度语言，使用 fugashi (McCann, 2020年）处理日语（日语），并使用 chinese[22] 处理中文。

对于停用词移除等数据预处理步骤，我们从 nltk[23]软件包以及 spaCy[24] 代码库中收集所有语言的公开停用词，并在集中式流程中用于实验。

我们使用数据集评估了各种摘要技术的性能，包括更简单的技术，例如 LEAD-3 和 RANDOM，这些技术在过去已被证明相当有用（Gha- landari 等人，2020年；Scialom 等，2020年；Sharma 等人，2019年）。我们还纳入了基于统计的方法 CENTROID （Radev 等人，2004年）以及基于图的方法TextRank （Mihalcea 和 Tarau，2004年）。

为了使用共享的专用模型，在多种语言之间进行公平比较，我们在多语言环境下，利用一个预先训练的

检查点[25] 对基于 Transformer 的MT5 （Xue 等人，2020年）模型的摘要性能进行了评估。最后，为了探索我们数据集的多模态特性，我们评估了一种基于多模态编码器-解码器的技术（Zhu 等人，2018年），该技术利用图像和文本生成多模态文本摘要。然而，公开可用的实现[26] 对MSMO的限制使我们只能针对英语进行评估。不过，为了比较这一得分，我们评估了三种最先进的基于 Transformer 的模型——Pegasus（Zhang 等人，2020年）、BART （Lewis 等人，2020年）和T5 （Xue 等人，2020年）的摘要性能，这些模型均兼容英语。

由于我们在上文描述的两个预训练模型是在 XSum 和 XL-Sum 数据集上微调的，而这两个数据集均提取自同一来源——BBC 新闻——因此，为了公平比较这些模型，我们避免对模型进行微调，并在讨论中解释相关得分。

在所有技术中，我们将生成摘要的长度阈值设为语料库中对应语言的黄金摘要平均长度。

6.2 基线

更简单的抽取式方法

LEAD-3：在这一基线方法中，提取源文本的前三个句子作为最终摘要。正如（Sharma 等人，2019年）所示，对于新闻摘要数据集而言，这是一种稳健的基线方法。

RANDOM：我们从源文本中递归地随机提取词语，直到达到摘要长度阈值。该基线方法旨在以无偏模型为参照，了解并比较其他基线方法。

统计方法

CENTROID：我们采用Radev 等人（2004年）提出的策略，根据句子中各词语的中心性得分对句子进行排序。我们使用 TF-IDF 得分衡量每个词语的相似度，并从每个排名中提取排名靠前的句子，直到达到摘要长度阈值。

基于图的方法

TEXTRANK：TextRank（Mihalcea 和 Tarau，

[19]https://github.com/zenquiorra/TokSeg[20] https://pypi.org/project/segtok/1.1.0/[21] https://github.com/anoopkunchukuttan/indic_nlp_library[22]https://pypi.org/project/chinese/[23] https://nltk.org/[24] https://github.com/explosion/spaCy/tree/master/spacy/lang

[25]https://huggingface.co/csebuetnlp/mT5_多语言_XLSum[26]我们使用作者提供的实现。该实现是一个多层软件包，鉴于软件复杂度，对其进行修改以适用于多语言环境并不可行

2004年）是一种基于源文本中句子相关性的无监督图排序技术[27] 我们将根据排序结果，在文档中处于最中心位置的句子视为生成的摘要。

基于循环神经网络的方法

MSMO：MSMO（Zhu 等人，2018年）是一种经过训练的编码器-解码器模型，用于多模态摘要。它利用多模态注意力机制，结合文本和图像生成多模态摘要。

基于 Transformer 的方法

MT5：MT5（Xue 等人，2020年）是一种基于 Transformer 的序列到序列模型，经过预训练以执行多种自然语言任务。我们使用公开可用的检查点[28]，该检查点针对XL-Sum 数据集（Hasan 等人，2021）进行了文本摘要预训练，适用于多语言环境。

PEGASUS：Pegasus（Zhang 等人，2020年）是一种基于 Transformer 的模型，通过预训练学习从输入文本中移除有意义的句子，因此适用于摘要任务。我们使用了一个检查点[29]，该检查点来自 PEGASUS 模型，并在 XSum 数据集（Narayan 等人，2018年）上进行了摘要预训练。

BART：BART（Lewis 等人，2020年）采用标准的序列到序列架构，包含一个双向编码器和一个从左到右解码器。我们在评估中使用了一个基于 DailyMail/CNN（Nal-lapati 等人，2016年）训练的预训练模型。

T5：T5（Raffel 等人，2020年）是一种编码器-解码器模型，通过混合多种自然语言任务进行训练，包括翻译和摘要；它将任何任务转换为文本到文本格式。我们使用预训练的 T5-large 模型执行摘要任务。

7 结果与讨论

我们使用ROUGE (Lin，2004年) 评估指标，将生成的摘要与参考摘要进行比较。我们报告了上述各个基线的ROUGE-1、ROUGE-2、ROUGE-L F分数（Lin，2004年）（参见表1 和3）。此外，我们还报告了BERTSCORE 在英语基线上的得分（Zhang 等人，2019年）（参见表1）。

[27]我们使用gensimhttps://radimrehurek.com/gensim_3.8.3/summarization/summariser.html 软件包提供的实现，并使用我们专用的软件包修改分词和分词器部分。[28] https://huggingface.co/csebuetnlp/mT5_多语言_XLSum[29] https://huggingface.co/google/pegasus-xsum

表2：比较使用多模态基线 MSMO 和基于单模态 Transformer 的基线生成的摘要，与英语数据集中的参考摘要之间的“ROUGE”F分数。“R-f1”表示 ROUGE-1 F分数，“R-f2”表示 ROUGE-2 F分数，“R-fL”表示 ROUGE-L F分数，“BrS”表示 BERTSCORE。

英语	R-f1	R-f2	R-fL	BrS
BART	0.195	0.031	0.131	0.863
Pegasus	**0.389**	**0.181**	**0.321**	**0.910**
T5	0.197	0.0328	0.131	0.858
MSMO	0.217	0.046	0.158	0.851

7.1 多语言基线分数

我们观察到，在我们的实验中使用的基于Transformer的技术，表现显著优于其他技术。然而，对于“MT5”这一列，我们观察到分数很高，但也出现了极低分数的峰值，如表3所示；这种情况可能由两个因素造成：

- 相对较高的分数可归因于使用了“MT5”检查点；该检查点针对摘要任务进行了微调，所用数据集（XL-Sum）与我们使用的数据集来自同一来源。

- 某些语言的分数非常低，可归因于“ROUGE”评估指标依赖词元重叠[30]。许多此类语言，尤其是源自达罗毗荼语系和印欧语系的语言，其词语会根据在文本中的位置及出现语境而发生显著变化，因此，如果不考虑词语的词根形式，简单的词元重叠指标就会显示较低的分数。

我们观察到，对于基于 Transformer 的基线表现不佳的语言，LEAD-3 的表现更好；这可归因于两个因素：

- 正如Sharma 等人(2019年)所指出的，LEAD-3在新闻领域的摘要任务中表现出色，这表明排名靠前的句子能够涵盖新闻文章中的大量信息。

- LEAD-3从文本中选取排名前三的句子，不同于抽象式摘要，新的

[30]由于我们用于实验的各种软件均不支持多语言词干提取方法，为了与受支持的语言进行公平比较，我们在评估过程中不对词元进行词干提取

表3：针对 M3LS 数据集中各语言参考摘要的各种摘要技术性能。“Lang”指根据 ISO639-1 标准表示语言的语言代码，“R-f1”指 ROUGE-1 F分数，“R-f2”指 ROUGE-2 F分数，“R-fL”指 ROUGE-L F分数

Base	RANDOM			LEAD-3			TextRank			CENTROID			MT5		
Lang	R-f1	R-f2	R-fL	R-f1	R-f2	R-fL	R-f1	R-f2	R-fL	R-f1	R-f2	R-fL	R-f1	R-f2	R-fL
bn	0.003	0.000	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.004**	0.001	0.003
mr	0.013	0.000	0.012	0.041	0.005	0.040	0.025	0.002	0.025	0.006	0.001	0.006	**0.044**	0.005	**0.044**
gu	0.014	0.001	0.014	**0.039**	0.005	0.038	0.014	0.001	0.014	0.016	0.002	0.016	0.036	0.005	0.036
ps	0.002	0.000	0.001	0.000	0.000	0.000	0.002	0.000	0.001	0.000	0.000	0.000	**0.003**	0.000	0.001
uk	0.030	0.002	0.029	0.062	0.016	0.061	0.043	0.010	0.042	0.032	0.006	0.032	**0.094**	0.025	**0.094**
pt	0.179	0.009	0.114	0.204	0.033	0.124	0.199	0.030	0.128	0.089	0.008	0.075	**0.276**	0.085	0.193
id	0.118	0.001	0.083	0.172	0.037	0.117	0.144	0.030	0.104	0.104	0.014	0.080	**0.289**	0.115	0.233
ne	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
pa	0.012	0.000	0.012	**0.038**	0.004	**0.038**	0.014	0.001	0.014	0.010	0.002	0.010	0.026	0.000	0.026
si	0.014	0.000	0.014	0.032	0.004	0.031	0.019	0.002	0.019	0.007	0.001	0.007	**0.039**	0.018	**0.039**
ur	0.006	0.000	0.006	0.023	0.001	0.023	0.006	0.000	0.005	0.024	0.001	0.023	**0.044**	0.000	**0.044**
fr	0.168	0.007	0.107	0.206	0.043	0.126	0.177	0.033	0.115	0.164	0.024	0.110	**0.209**	0.041	0.141
ru	0.032	0.001	0.032	0.071	0.017	0.069	0.041	0.012	0.040	0.036	0.008	0.036	**0.081**	0.011	**0.081**
ja	0.069	0.001	0.068	0.126	0.012	0.120	0.084	0.007	0.081	0.063	0.004	0.062	**0.306**	0.081	0.291
te	0.010	0.000	0.009	0.023	0.001	0.023	0.011	0.000	0.011	0.008	0.001	0.008	**0.026**	0.000	**0.026**
ta	0.014	0.001	0.014	**0.034**	0.005	**0.034**	0.023	0.003	0.022	0.012	0.001	0.012	0.026	0.000	0.026
zh	0.022	0.001	0.022	**0.053**	0.008	0.051	0.042	0.005	0.041	0.025	0.003	0.025	0.125	0.042	0.118
es	0.177	0.008	0.117	0.180	0.033	0.117	0.110	0.018	0.073	0.081	0.008	0.067	**0.280**	0.084	0.202
hi	0.010	0.000	0.010	**0.018**	0.002	0.018	0.013	0.001	0.013	0.005	0.000	0.005	0.002	0.000	0.001
en	0.146	0.002	0.102	**0.175**	0.026	0.114	0.100	0.014	0.071	0.140	0.016	0.102	**0.427**	0.182	0.345

词元或现有词元的新形式不会出现在给定的文章中。由于这是一种抽取式技术，词元重叠的可能性更高，因此“F分数”也会更好。

7.2 多模态基线分数

由于数据集中的大多数语言在多模态设置下缺乏预训练框架，我们只能在英语数据集上评估多模态技术。在比较各种单模态技术与多模态技术的“F分数”时，我们注意到，基于 Transformer 的模型 Pegasus 优于其他技术。这主要归因于这样一个事实：我们用于通过 Pegasus 模型评估摘要的预训练检查点是在 XSum 数据集上微调的，而该数据集所收集的数据与我们使用的数据来自同一来源。我们观察到，对于其他未在与我们数据集来自同一来源的数据集上进行微调的模型，多模态技术 MSMO 能够优于其他技术。

7.3 所提出数据集的抽象性

我们提出了一个抽象式摘要数据集，其中的目标摘要由人工撰写。M3LS 数据集需要采用抽象式方法，因为新

数据集中的一元组占比相当高（请参阅附录B中的“abs.gold”列）。这一点也体现在基线技术的结果中。例如，MT5在多种语言上的表现始终更优，如表3所示；抽象式基线的ROUGE分数是抽取式基线的三倍。

8 结论

在这项工作中，我们发布了一个大型多模态多语言摘要数据集，其中包含超过 1.1M+ 篇新闻文章，涵盖 20种语言，并利用 M3LS 提出了多模态多语言摘要这一问题。据我们所知，这是首个涵盖多种语言的多模态摘要数据集。对于 20 种语言中的 13 种，所提出的数据集是规模最大的摘要数据集。我们评估了各种基线系统的表现，以验证所提出的数据集在多模态和多语言设置下的质量。我们希望通过这项工作，推动社区对各种研究较少的语言开展研究，探索包括但不限于摘要、标题生成、关键词提取、图像图注生成、多模态嵌入生成等研究问题。未来，我们计划利用该数据集开发可解决 M3LS任务的共享模型。

局限性

我们的工作有几点需要注意。**首先**，该数据集目前采用多模态输入，生成文本摘要。不过，未来的工作可以通过为图像添加标注，以多模态输出丰富数据集。**其次**，M3LS数据集中各语言的数据分布不均衡，这是由于英国广播公司发布的文章在不同语言间数量不平衡，以及某些语言的虚拟印刷媒体起步较晚所致（如附录A所示）。**第三**，当前数据集采用独立同分布划分来构建训练集和测试集，但未来也可以探索对抗划分和似然划分等更先进的方法。**第四**，尽管当前论文稿并未同时从多模态和多语言两个方面评估该数据集，但我们相信，该数据集有望为未来此类系统的开发作出贡献。

致谢

本出版物是印度政府电子与信息技术部维斯维斯瓦拉亚博士奖学金计划下项目所开展的研发工作的成果，该计划由印度数字公司（前身为Media LabAsia）实施。

参考文献

Rasim Alguliev、Ramiz Aliguliyev 和 Makrufa Hajirahimova。2010。基于整数线性规划的多文档摘要模型。智能控制与自动化，1(02)：105。

Dzmitry Bahdanau、Kyung Hyun Cho 和 Yoshua Bengio。2015。通过联合学习对齐与翻译实现神经机器翻译。载于第三届国际学习表征会议，*ICLR 2015*。

Shuang Bai 和 Shan An。2018年。自动图像图注生成综述。《神经计算》，311:291–304。

Jingqiang Chen 和 Hai Zhuge。2018年。使用多模态注意力层次循环神经网络进行抽象式文本图像摘要。在*2018*年自然语言处理经验方法会议论文集中，第 4046–4056 页。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向 Transformer 预训练。载于*2019*年会议论文集计算语言学协会北美分会人类语言技术会议论文集：第 *1* 卷（长篇与短篇论文），第 4171–4186 页，明尼阿波利斯，明尼苏达州。计算语言学协会。H. P. Edmundson。1969年。自动抽取的新方法。《*ACM*期刊》，16：264–285。Desmond Elliott 和 Ákos Kádár。2017年。想象力提升多模态翻译。载于第八届自然语言处理国际联合会议论文集（第 *1* 卷：长篇论文），第 130–141 页。

A. R. Fabbri、Wojciech Kryscinski、Bryan McCann、R. Socher 和 Dragomir Radev。2021。Summeval：重新评估摘要评估。计算语言学协会会刊，9:391–409。

Daniil Gavrilov、Pavel Kalaidin 和 Valentin Malykh。2019年。用于标题生成的自注意力模型。在欧洲信息检索会议中，第 87–93 页。施普林格。

Demian Gholipour Ghalandari、Chris Hokamp、John Glover、Georgiana Ifrim 等。2020年。一种基于维基百科当前事件门户的大规模多文档摘要数据集。载于第*58*届计算语言学协会年会论文集，第1302–1308页。

Tahmid Hasan、Abhik Bhattacharjee、Md Saiful Islam、Kazi Samin、Yuan-Fang Li、Yong-Bin Kang、M So-hel Rahman 和 Rifat Shahriyar。2021。XL-Sum：面向44种语言的大规模多语言抽象式摘要。*arXiv*预印本 *arXiv:2106.13822*。

Anubhav Jangra、Adam Jatowt、Mohammad Hasanuz-zaman 和 Sriparna Saha。2020a。使用联合整数线性规划生成文本-图像-视频摘要。载于信息检索进展：第*42*届欧洲信息检索研究会议，*ECIR 2020*，里斯本，葡萄牙，*2020*年*4*月*14*日至*17*日，论文集，第二部分，第190–198页。

Anubhav Jangra、AdamJatowt、SriparnaSaha 和 Mohammad Hasanuzzaman。2021a。多模态摘要综述。*arXiv* 预印本 *arXiv:2109.05199*。

Anubhav Jangra、Sriparna Saha、Adam Jatowt 和 Mohammed Hasanuzzaman。2020b。使用多目标优化生成多模态摘要。载于第*43*届*ACM SIGIR*信息检索研究与发展国际会议论文集，第1745–1748页。Anubhav Jangra、Sriparna Saha、Adam Jatowt 和 Mohammed Hasanuzzaman。2021b。使用多目标优化进行多模态补充-互补摘要。载于第*44*届*ACM SIGIR*信息检索研究与发展国际会议论文集，第818–828页。

Prince Jha、Gaël Dias、Alexis Lechervy、Jose G Moreno、Anubhav Jangra、Sebastião Pais 和 Sriparna Saha。2022。结合视觉与语言表征，通过基于片段的识别来识别词汇语义关系。载于第*30*届*ACM*国际多媒体会议论文集，第4406–4415页。

Di Jin、Zhijing Jin、Joey Tianyi Zhou、Lisa Orii 和 Peter Szolovits。2020年。标题中的钩子：学习以受控风格生成标题。载于计算语言学协会第*58*届年会论文集，第5082–5093页。

Faisal Ladhak、Esin Durmus、Claire Cardie 和 Kathleen McKeown。2020年。WikiLingua：一个用于多语言抽象式摘要的新基准数据集。载于*2020*年自然语言处理实证方法会议：研究发现论文集，第4034–4048页。

Sungjick Lee 和 Han-joon Kim。2008年。用于主题追踪的新闻关键词提取。载于*2008*年第四届网络计算与高级信息管理国际会议，第2卷，第554–559页。IEEE。

Mike Lewis、Yinhan Liu、Naman Goyal、Mar- jan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020年。BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练。收录于 计算语言学协会第*58*届年会论文集，第7871–7880 页。Haoran Li、Junnan Zhu、Tianshang Liu、Jiajun Zhang 和 Chengqing Zong。2018年。结合模态注意力和图像筛选的多模态句子摘要。收录于 第二十七届国际人工智能联合会议论文集，*IJCAI-18*，第 4152–4158 页。国际人工智能联合会议组织。Liunian Harold Li、Mark Yatskar、Da Yin、Cho-Jui Hsieh 和 Kai-Wei Chang。2019年。VisualBERT：一种简单且高效的视觉与语言基线。*arXiv* 预印本 *arXiv:1908.03557*。Mingzhe Li、Xiuying Chen、Shen Gao、Zhang- ming Chan、Dongyan Zhao 和 Rui Yan。2020年。V MSMO：学习为视频新闻文章生成多模态摘要。*arXiv* 预印本 *arXiv:2010.05406*。Chin-Yew Lin。2004年。ROUGE：用于自动评估摘要的软件包。收录于文本摘要拓展，第 74–81 页。Jiasen Lu、Dhruv Batra、Devi Parikh 和 Stefan Lee。2019年。ViLBERT：面向视觉与语言任务的任务无关视觉语言表征预训练。收录于神经信息处理系统进展，第 13–23 页。

Paul McCann。2020年。fugashi：一种用于在 Python 中对日语进行分词的工具。收录于 第二届 *NLP* 开源软件研讨会论文集（*NLP-OSS*），第44–51 页，线上举行。计算语言学协会。

Rada Mihalcea 和 Paul Tarau。2004年。TextRank：为文本引入秩序。载于*2004*年自然语言处理经验方法会议论文集，第404–411页。

Sourajit Mukherjee、Anubhav Jangra、Sriparna Saha 和 Adam Jatowt。2022。主题感知多模态摘要。载于计算语言学协会论文集：*AACL-IJCNLP 2022*，第387–398页，仅在线发布。计算语言学协会。

Ramesh Nallapati、Bowen Zhou、C. D. Santos、Çaglar Gülçehre 和 Bing Xiang。2016年。使用序列到序列循环神经网络及其他方法进行抽象式文本摘要。载于*CoNLL*。

Shashi Narayan、Shay B Cohen 和 Mirella Lapata。2018年。别告诉我细节，只要摘要！面向极端摘要的主题感知卷积神经网络。载于*2018*年自然语言处理经验方法会议论文集，第1797–1807页。

Dragomir R Radev、Hongyan Jing、Małgorzata Sty´s 和 Daniel Tam。2004年。基于质心的多文档摘要。信息处理与管理，40(6):919–938。Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J Liu。2020年。探索统一文本到文本 Transformer 的迁移学习极限。机器学习研究期刊，21:1–67。Naveen Saini、Sriparna Saha、Anubhav Jangra 和 Pushpak Bhattacharyya。2019年。使用多目标优化进行抽取式单文档摘要：探索自组织差分进化、灰狼优化器和水循环算法。知识型系统，164:45–67。Ramon Sanabria、Ozan Caglayan、Shruti Palaskar、Desmond Elliott、Loïc Barrault、Lucia Specia 和 Florian Metze。2018年。How2：一个用于多模态语言理解的大规模数据集。发表于*NeurIPS*。Thomas Scialom、Paul-Alexis Dray、Sylvain Lamprier、Benjamin Piwowarski 和 Jacopo Staiano。2020年。MLSUM：多语言摘要语料库。*arXiv* 预印本 *arXiv:2004.14900*。Abigail See、Peter J. Liu 和 Christopher D. Manning。2017。直奔主题：使用指针生成网络进行摘要。发表于计算语言学协会第*55*届年会论文集（第*1*卷：长篇论文），第1073–1083页，加拿大温哥华。计算语言学协会。

Eva Sharma、Chen Li 和 Lu Wang。2019年。Bigpatent：一个用于抽象式和连贯摘要的大规模数据集。发表于计算语言学协会第*57*届年会论文集，第2204–2213页。

Md Imran Hossain Showrov 和 Masrur Sobhan。2019年。孟加拉语新闻的关键词提取。载于*2019*年第五届电气工程进展国际会议（*ICAEE*），第658–662页。IEEE。

Hrituraj Singh、Anshul Nasery、Denil Mehta、Aish- warya Agarwal、Jatin Lamba 和 Balaji Vasan Srini- vasan。2021年。Mimoqa：多模态输入、多模态输出问答。载于计算语言学协会北美分会*2021*年人类语言技术会议论文集，第5317–5332页。

Mohammad Soleymani、David Garcia、Brendan Jou、Björn Schuller、Shih-Fu Chang 和 Maja Pantic。2017年。多模态情感分析综述。图像与视觉计算，65：3–14。

Chen Sun、Austin Myers、Carl Vondrick、Kevin Mur- phy 和 Cordelia Schmid。2019年。VideoBERT：用于视频和语言表征学习的联合模型。载于*IEEE*国际计算机视觉会议论文集，第7464–7473页。

Ilya Sutskever、Oriol Vinyals 和 Quoc V Le。2014。基于神经网络的序列到序列学习。载于神经信息处理系统进展，第 3104–3112 页。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017。注意力就是你所需要的一切。载于神经信息处理系统进展，第 5998–6008 页。

Yash Verma、Anubhav Jangra、Sriparna Saha、Adam Jatowt 和 Dwaipayan Roy。2022。MAKED：多语言自动关键词提取数据集。载于第十三届语言资源与评估会议论文集，第 6170–6179 页。

Danqing Wang、Jiaze Chen、Hao Zhou、Xipeng Qiu 和 Lei Li。2021。用于多语言摘要的对比对齐联合学习。载于计算语言学协会论文集：*ACL-IJCNLP 2021* 研究成果，第 2739–2750 页。

Kelvin Xu、Jimmy Ba、Ryan Kiros、Kyunghyun Cho、Aaron Courville、Ruslan Salakhudinov、Rich Zemel 和 Yoshua Bengio。2015。《展示、关注与讲述：利用视觉注意力进行神经图像描述生成》。收录于国际机器学习会议，第 2048–2057 页。PMLR。

Linting Xue、Noah Constant、Adam Roberts、Mi- hir Kale、Rami Al-Rfou、Aditya Siddhant、Aditya Barua 和 Colin Raffel。2020年。mt5：一种大规模多语言预训练文本到文本转换器。*arXiv*预印本 *arXiv:2010.11934*。

Lu Yao、Zhang Pengzhou 和 Zhang Chi。2019年。基于 TF-IDF 和 TextRank 的新闻关键词提取技术研究。收录于 *2019*年 *IEEE/ACIS* 第 *18* 届国际计算机与信息科学会议（*ICIS*），第 452–455 页。IEEE。

Jingqing Zhang、Yao Zhao、Mohammad Saleh 和 Pe- ter Liu。2020年。《Pegasus：利用提取的间隙句进行预训练，以实现抽象式摘要》。收录于国际机器学习会议，第 11328–11339 页。PMLR。

Ruqing Zhang、JiafengGuo、Yixing Fan、Yanyan Lan、Jun Xu、Huanhuan Cao 和 Xueqi Cheng。2018年。《面向新闻文章的问题式标题生成》。收录于第 *27* 届 *ACM* 信息与知识管理国际会议论文集，第 617–626 页。

Tianyi Zhang、Varsha Kishore、Felix Wu、Kilian Q Weinberger 和 Yoav Artzi。2019年。BERTSCORE：使用 BERT 评估文本生成。*arXiv* 预印本 *arXiv:1904.09675*。

Luowei Zhou、Hamid Palangi、Lei Zhang、Houdong Hu、Jason J Corso 和 Jianfeng Gao。2020年。用于图像图注生成和视觉问答的统一视觉-语言预训练。发表于*AAAI*，第 13041–13049 页。

Junnan Zhu、Haoran Li、Tianshang Liu、Yu Zhou、Ji- ajun Zhang 和 Chengqing Zong。2018年。Msmo：具有多模态输出的多模态摘要。载于*2018*年自然语言处理实证方法会议论文集，第 4154–4164 页。

A M3LS 数据集中给定年份的文章数量频率

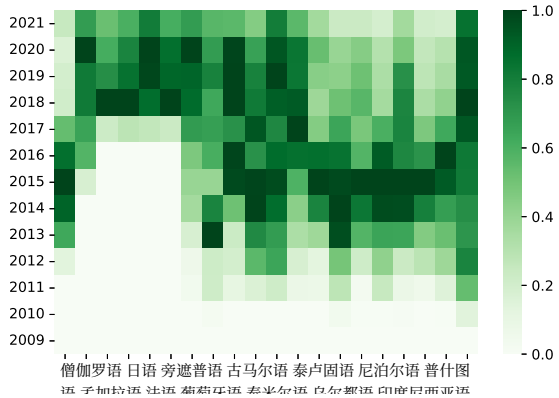


图3：M3LS 数据集中各语言的时间跨度。颜色越深，得分越接近 1.0，反之亦然；这表明该语言在相应年份发表的文章数量频率越高。

B M3LS 数据集的详细统计数据

精心整理的M3LS 数据集的详细统计数据见表4。

表4：“Lang.”表示依据 ISO 639-1 标准的语言代码，**文章**表示语料库中每种语言的文章数量，**toks.** 表示语料库中该语言每篇文章的平均词元数，**tok.uni.** 表示该语言任意一篇文章中唯一词元的平均数量，**sum.tok.** 表示该语言每篇文章摘要中的平均词元数，**sum.uni** 表示该语言每篇文章摘要中唯一词元的平均数量，**sent.** 表示该语言每篇文章中的平均句子数，**abs.gold** 表示该语言摘要中未出现在文章里的词元平均占比，**图像**表示该语言每篇文章中的平均图像数量，**i.c.r** 表示附有图注的图像数量与未附图注的图像数量之比的平均值。

语言	文章	tok.	词元数（唯一）	摘要词元数	摘要词元数（唯一）	句子数	abs.gold	图像	i.c.r
bn	25283	464.78	254.73	23.12	21.97	28.42	43.18	2.44	0.58
mr	16161	871.36	404.82	27.88	25.02	63.32	48.33	0.00	0.00
gu	12175	868.75	404.84	25.87	23.63	50.11	40.87	4.94	0.30
ps	23205	523.35	207.95	32.85	27.52	18.29	33.32	2.10	0.49
uk	90846	471.90	216.57	24.59	22.78	23.49	54.07	1.56	0.22
pt	39454	2855.68	424.14	38.57	33.01	114.22	35.23	3.34	0.64
id	56108	587.88	225.12	24.59	22.89	28.54	37.07	2.47	0.57
ne	18953	402.17	229.71	21.50	20.95	23.79	45.00	2.07	0.24
pa	11600	843.74	319.75	30.96	27.35	38.87	30.42	4.57	0.37
si	10148	331.55	186.26	24.30	23.43	15.53	51.33	1.67	0.34
ur	55107	690.47	264.57	35.95	31.00	1.07	27.94	2.35	0.43
fr	25923	413.45	179.22	31.22	27.26	14.43	40.75	1.64	0.47
ru	95345	668.61	302.63	28.38	25.74	26.68	52.80	1.96	0.38
ja	11023	1052.94	282.58	48.59	36.97	33.92	29.86	2.90	0.56
te	15511	626.24	353.70	23.71	21.39	51.72	52.19	4.41	0.27
ta	38523	354.35	209.30	21.18	19.91	23.39	56.34	2.55	0.27
zh	60830	787.68	309.11	34.38	29.80	29.87	37.21	2.00	0.48
es	66816	2649.57	345.35	30.64	26.28	83.22	38.43	4.15	0.66
hi	61852	776.39	265.10	29.45	25.70	37.21	32.00	1.24	0.0
en	376367	657.95	268.60	25.27	23.40	22.28	32.86	1.39	0.35