arXiv:2302.04790v1 [cs.CL] 9 Feb 2023

面向低资源印度语言的跨语言事实抽取：大规模多语言语言模型

Bhavyajeet Singh*，**Pavan Kandru***，**Anubhav Sharma** 和**VasudevaVarma**信息检索与抽取实验室，海得拉巴国际信息技术学院，印度 {bhavyajeet.singh, siri.venkata, anubhav.sharma}@research.iiit.ac.in vv@iiit.ac.in

摘要

维基数据等大型知识图谱试图捕捉关于多个实体的世界知识。近期方法致力于从文本中自动丰富这些知识图谱。然而，低资源语言的自然文本中包含的大量信息往往未被利用。跨语言信息抽取旨在从低资源印度语言文本中抽取事实信息，并以英语三元组的形式呈现。尽管这一任务潜力巨大，但与单语信息抽取相比，其进展仍然滞后。本文提出了从文本中进行跨语言事实抽取（CLFE）这一任务，并为此设计了一种端到端生成式方法，其总体 F1 分数达到 77.46。

1 引言

知识图谱是关于世界的大型结构化信息来源。近年来，人们投入了大量关注，致力于寻找自动构建或扩充大型知识图谱（KG）的方法（Gupte 等人，2021年）、（Zou，2020）。维基数据（Vrandeˇci´c 和 Krötzsch，2014）是最大的公开可用知识图谱之一，拥有超过 9900 万个实体。维基数据等知识图谱已广泛应用于多种场景，例如文本生成（Koncel-Kedziorski 等人，2019）、问答（Srivastava 等人，2021）、（Li 等人，2022）等。

知识图谱由多个相互关联的事实构成。事实通常表示为三元组，由两个实体及其间的语义关系组成。这些信息可以编码为三元组 $< h, r, t >$ 其中 h 是主语实体，r 是关系，t 表示尾实体。

事实抽取是指从自然语言文本中提取结构化事实信息的任务（Charu C.Aggarwal，2012）。此前关于单语事实抽取任务的研究已相当广泛，尤其是英语领域（Zenget al.，2019）（Liet al.，2021），然而，跨语言事实抽取任务却未受到太多关注。

本文提出一项重要任务：针对 7 种低资源印度语言和英语进行多语言及跨语言事实到文本抽取（CLFE）。该任务旨在直接从 8 种不同语言中抽取英语三元组。我们还为此任务提出了强大的基线方法和方法，其结果可与现有单语言最先进的事实抽取流程相媲美，并显著优于以往其他跨语言事实抽取尝试（Zhanget al.，2017a）。我们的工作能够利用印度语文本中包含的事实知识，从而扩大现有知识图谱的覆盖范围。这还将进一步有助于事实核查、文本生成等多种下游任务。据我们所知，这是首次尝试从低资源印度语言中进行多语言及跨语言事实抽取。图1 展示了 CLFE 任务输入和输出的多个示例。

总体而言，我们做出了以下贡献。(1) 提出了面向低资源印度语言的跨语言和多语言事实抽取问题。(2) 提出了一种端到端生成式方法，可从低资源印度语言文本中抽取以主体为中心的事实信息，其效果显著优于基于分类的流程。(3) 我们训练了多个多语言 CLFE 模型，总体 F1 分数达到 77.46。

2 相关工作

从自由文本中提取结构化信息是一项已有大量研究的问题。T-REx (Elsahar 等人，2018) 使用实体链接、共指消解和基于字符串匹配的链接

* 同等贡献

代码可在以下地址获取：https://github.com/bhavyajeet/CLFE

Multilingual Texts	Head Entity	Extracted Facts
Elon Reeve Musk (born 28 June 1971) is a South African-Canadian-American veteran businessman, investor, engineer, and inventor.	Elon Musk	< nationality, South_Africa> < nationality, Canda> < nationality, USA> < date_of_birth, 28_June_1971> < occupation, engineer> < ocuupation, entrepreneur> < occupation, inventor> < occupation, investor>
एलन मस्क (जन्म 28 जून 1971) एक दक्षिण अफ्रीकी-कनाडाई-अमेरिकी दिग्गज व्यापारी, निवेशक, इंजीनियर, और आविष्कारक हैं।	Elon Musk	< nationality, South_Africa> < nationality, Canda> < nationality, USA> < date_of_birth, 28_June_1971> < occupation, engineer> < ocuupation, entrepreneur> < occupation, inventor> < occupation, investor>
এলন মাস্ক (জন্ম 28 জুনজু 1971) দক্ষিণ আফ্রিকা-কানাডিয়ান-আমেরিকান প্রবীণ ব্যবসায়ী, বিনিয়োগকারী, প্রকৌশলী এবং উদ্ভাবক।	Elon Musk	< nationality, South_Africa> < nationality, Canda> < nationality, USA> < date_of_birth, 28_June_1971> < occupation, engineer> < ocuupation, entrepreneur> < occupation, inventor> < occupation, investor>
એલોન મસ્ક (જન્મ 28 જૂન 1971) એ દક્ષિણ આફ્રિકા-કેનેડિયન□અમેરિકન પીઢ ઉદ્યોગપતિ, રોકાણકાર, ઇજનેર અનેશોધક છે.	Elon Musk	< nationality, South_Africa> < nationality, Canda> < nationality, USA> < date_of_birth, 28_June_1971> < occupation, engineer> < ocuupation, entrepreneur> < occupation, inventor> < occupation, investor>

图 1：CLFE 任务的输入与输出示例。输入为任意语言的文本及感兴趣的实体（头实体），以抽取英语事实（关系与尾实体对）。并非所有语言中都一定会出现同一个句子。

流程，在 DBpedia (Lehmann 等人,2015) 摘要与 Wiki- data (Vrande˘ci´c 和 Krötzsch,2014) 三元组之间执行事实链接。RE- FCOG (Kolluru 等人, 2021年) 在跨语言空间中链接事实，性能优于 (Elsahar 等人, 2018) 等现有的基于流程的方法。但这些方法的应用范围有限，因为它们执行事实链接，并且需要将事实集作为输入。

开放信息抽取（Angeli 等人，2015）通过利用语言结构来解决开放域信息抽取问题。早期的开放域信息抽取系统（如 Ollie（Mausam 等人，2012））使用大量覆盖面广的模式来抽取事实，而 OpenIE 使用少量模式，这些模式适用于规范结构的句子。然而，这些开放域信息抽取器生成的事实包含冗长且过度具体的关系，无法用于构建知识图谱。

（Zhong 和 Chen，2020）；（Sui 等人，2020a）通过联合抽取实体及其关系来解决信息抽取问题，使用神经网络模型处理输入文本，无需参照任何事实库。尽管这些工作能够构建系统，从文本中抽取 WebNLG 数据集（Gardent 等人，2017）中的开放信息，但它们仅适用于单一语言，知识抽取也局限于单一语言。许多现有的高性能关系抽取模型，如（Yan 等人，2021）、（Sui 等人，2020b），部分依赖于源文本中实体的精确匹配，因此较难将其应用于 CLFE 任务。

跨语言事实抽取，即从不同语言的源文本中抽取事实，受到的关注远不如单语事实抽取。尽管（Zhang 等人，2017b）曾研究过这一任务，但仅涉及一种语言，报告的最高 F1分数为 33.67。此外，尚无人尝试从印度语言等低资源语言中抽取事实。本研究旨在通过提出面向低资源印度语言的跨语言主体中心事实抽取系统，缩小信息抽取领域的这些差距。

3 数据集

对于 CLFE 任务，我们采用 XAlign 数据集（Abhishek 等人，2022）。该数据集包含 8 种语言的 45 万对数据，其中 5402 对经过人工标注。数据集中经人工标注的部分被用作黄金测试集。XAlign 中的句子来自维基百科文章，内容涉及以印度语言撰写、属于人类类别的实体。

XAlign 数据集具有广泛的跨语言和多语言特性，非常适合所提出的任务。尽管该数据集最初是为跨语言数据到文本生成任务设计的，但也可用于 CFLE。不过，该数据集也存在一些挑战。如果将每种关系视为一个类别（用于基于分类的方法），数据集的类别分布会严重失衡。在约 367 种不同的关系（类别）中，仅出现频率最高的类别就占 27%，而出现频率最高的 20 个类别共占数据的 90%。平均而言，每个句子对应 2.02 条对齐事实。

此外，该数据集的另一个难点在于它是部分对齐的。测试集中的句子所包含的信息都能在对齐事实中找到，但训练集句子中的全部信息并未被对齐事实覆盖。数据集的这一特性，

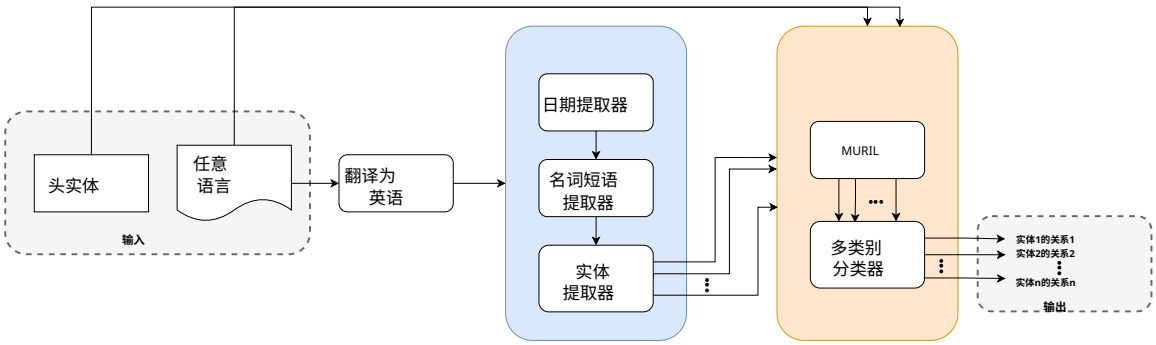


图2：CLFE 流程架构

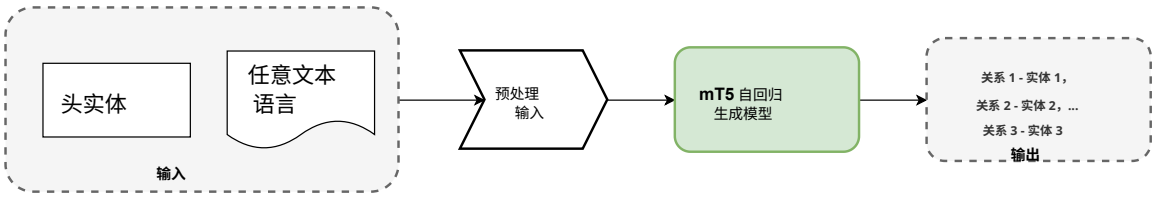


图3：CLFE 端到端架构

可能会导致模型在训练过程中即使生成了正确事实也受到惩罚，从而影响测试时的召回率得分。更多详情请参阅附录部分 A.2

4 方法

我们为 CLFE 任务提出了两种方法。第一种是基于分类的方法，先提取尾实体，再预测关系。第二种是生成式方法，一次性完成这两项任务。

4.1 尾实体提取与关系分类（TERC）

TERC 流程（图 2）包含两个步骤。第一步是从源语言文本中提取事实的尾实体。为此，我们使用 IndicTrans（Ramesh 等人，2021年）进行翻译，将输入文本转换为英语。随后，我们从文本中提取所有日期并将其规范化为统一格式。我们还会在原文中将日期替换为虚设标记，以防日期被识别为其他实体。由于每个尾实体只能是名词或专有名词，我们使用 spaCy（Honnibal 和 Montani，2018）名词短语提取器提取所有名词短语块，并按如下方式从中选取尾实体。

- 移除与头实体匹配的实体。由于在流程的这一阶段我们只关注尾实体，因此会移除所有与头实体存在较高词汇重叠的实体。
- 移除所有以代词词根为中心的名词短语块，以过滤代词。数据中的尾实体从来不是代词，因此我们会剔除所有以代词为头的已识别短语。
- 将带有形容词和专有名词词性标签的连续词元片段选作独立实体。尾实体可以由多个词构成，其范围内也可能包含形容词，因此我们利用词性标签，为每个检测到的专有名词提取尽可能长的片段。
- 如果名词短语块的中心词词性标签为名词，则将其选作一个独立实体。

下一步是为这些尾实体分别预测关系。为此，我们使用预训练的 MURIL（Khanuja 等人，2021年）来生成头实体、尾实体和源语言输入文本的联合表示。该表示作为输入传递给分类器，用于预测输入中头实体与尾实体之间的关系。分类器在训练集上进行训练，给定一个句子和一个 <头实体、尾实体> 对来预测关系，并将真实标注中的尾实体作为输入。为解决类别不平衡问题，我们在损失函数中使用 **类别分布倒数的对数**作为权重，其效果优于标准的类别分布倒数以及无权重损失。

在评估流程架构的性能时，会将从翻译后的输入文本中提取的尾实体与真实尾实体进行对齐。有关此对齐的详细信息，请参见 A.1附录

	te	bn	ta	gu	mr	en	hi	kn	所有语言		
	F1	F1	F1	F1	F1	F1	F1	F1	P	R	F1
分类 使用真实尾实体	69.19	67.50	89.44	85.74	51.38	72.87	87.10	79.74	79.04	77.93	75.37
TERC	43.66	41.96	52.19	40.30	44.59	50.80	50.46	42.57	40.45	53.71	46.15
端到端跨语言生成式模型	71.82	75.56	82.82	**72.36**	**77.79**	76.28	**86.62**	68.04	74.09	**81.15**	**77.46**
端到端生成 w 文字统一	**72.51**	75.38	**85.21**	72.04	77.19	74.56	83.44	**70.46**	78.49	76.15	77.29
双语模型	70.94	**78.01**	83.71	67.84	71.91	**76.64**	86.49	63.19	**79.79**	71.63	75.49

表 1：测试集中所有语言上应用各种方法所得的精确率、召回率和 F1分数。请注意，“使用真实尾实体的分类”使用真实标注中的尾实体作为关系预测模型的输入，因此不代表完整的流水线

。预测针对对齐的尾实体进行，并据此计算评估指标。

4.2 端到端生成式抽取

我们还针对事实抽取问题提出了一种端到端方法（图3），该方法能够联合抽取尾实体及其与头实体之间的关系。该领域以往针对单语言事实抽取的研究表明，与相互独立的方法相比，同时执行尾实体抽取和关系抽取的模型更有可能取得更好的效果（Li 等， 2021年）。与上述流水线方法相比，该方法的优势在于尾实体抽取和关系预测之间存在双向交互，从而提升了两项任务的性能，因为二者并非彼此独立。

我们将这个问题设定为文本到文本任务，并使用 mT5（Xue 等人，2020）自回归序列到序列模型，在给定头实体和输入文本作为输入时生成关系和尾实体。我们使用交叉熵损失训练该模型。生成式模型具有更强的泛化能力，并支持更开放的信息抽取，即关系和尾实体的集合不受限制。

我们对该流程的 3 种变体进行了实验。在所有变体中，事实都被线性化为目标文本：将头实体和尾实体用特殊标记连接起来。因此，对于给定句子 S，如果对应的事实是 i $[h, r_1, t_1]$, $[h, r_2, t_2]$....$[h, r_i, t_i]$，那么目标文本将是 $< R > r_1 < T > t_1 < R > r_2 < T > t_2 .... < R > r_i < T > t_i$。

第一种方案是对预训练的 mT5 模型进行微调，使其在所有语言上执行事实抽取任务。对于第二个实验，我们采用文字系统统一方法，将除英语以外所有语言的输入文本音译为天城文。这样做的想法是，统一的文字系统输入有助于模型训练，因为多种印度语言的词汇存在高度重叠。在第三种方案中，我们训练多个双语事实抽取模型，每种语言各训练一个。关于这些模型和 TERC(4.1) 的实现细节见附录 A.3。

5 结果与讨论

表 1总结了第 4 节中提到的多种事实抽取方法的结果。

可以看出，开放式方法的 F1 分数表现最佳，同时也能为可能涉及的实体和关系提供完全的灵活性。另一个发现是，对于英语和孟加拉语这两种语言，分别训练双语模型的策略优于组合模型。这是因为这两种语言在我们的数据集中出现频率最高，合计占训练数据的 54.44%。因此，由于印度语言之间可以共享学习成果，多语言训练总体上证明是有益的。我们还发现，文字统一（将输入文字转写为天城文）尤其有利于数据集中的所有达罗毗荼语系语言（te、ta、kn）。

需要指出的是，模型的实际表现可能优于数字所显示的结果。原因在于，我们目前采用严格的评估方案，只有当预测尾实体与实际尾实体之间存在词匹配时，才会将预测判定为正确。然而，这种评估方式未能涵盖预测尾实体与真实尾实体完全同义的情况，例

如，模型将职业预测为“作家”，而真实标签却是“作者”。

6 结论与未来工作

在这项工作中，我们提出了针对英语和另外七种印度低资源语言的多语言及跨语言事实抽取任务。我们得出结论：尽管文字统一有助于某些语言，但单一的多语言端到端生成流程表现更好，整体 F1 分数达到 77.46。这项工作为未来研究如何从低资源印度语言文本中提取知识铺平了道路。未来，我们计划探索能够专门应对数据集部分对齐特性的方案，以取得进一步改进。

参考文献

Tushar Abhishek、Shivprasad Sagare、Bhavyajeet Singh、Anubhav Sharma、Manish Gupta 和 Va- sudeva Varma。2022 年。Xalign：面向低资源语言的跨语言事实到文本对齐与生成。载于 *2022* 年万维网大会配套论文集，WWW ’22，第 171–175 页，纽约，美国纽约州。计算机协会。

Gabor Angeli、Melvin Jose Johnson Premkumar 和 Christopher D. Manning。2015 年。利用语言结构进行开放域信息抽取。载于第 *53* 届计算语言学协会年会暨第 *7* 届自然语言处理国际联合会议论文集（第 *1* 卷：长篇论文），第 344–354 页，中国北京。计算语言学协会。

ChengXiang Zhai、Charu C. Aggarwal。2012 年。文本数据挖掘。

Hady Elsahar、Pavlos Vougiouklis、Arslen Remaci、Christophe Gravier、Jonathon Hare、Frederique Laforest 和 Elena Simperl。2018 年。T-REx：自然语言与知识库三元组的大规模对齐。载于第十一届语言资源与评测国际会议论文集（*LREC 2018*），日本宫崎。欧洲语言资源协会（ELRA）。

Claire Gardent、Anastasia Shimorina、Shashi Narayan 和 Laura Perez-Beltrachini。2017年。为自然语言生成微规划器创建训练语料。载于计算语言学协会第*55*届年会论文集（第*1*卷：长篇论文），第179–188页，加拿大温哥华。计算语言学协会。

Athang Gupte、Saumitra Sapre 和 SheetalSonawane。2021年。使用文本生成知识图谱神经机器翻译技术。载于 *2021*年通信、信息与计算技术国际会议（*ICCICT*），第 1–8 页。Matthew Honnibal 和 Ines Montani。2018 年。spaCy 2：借助 Bloom 嵌入、卷积神经网络和增量解析实现自然语言理解。 即将发表。Simran Khanuja、Diksha Bansal、Sarvesh Mehtani、Savya Khosla、Atreyee Dey、Balaji Gopalan、Dilip Kumar Margam、Pooja Aggarwal、Rajiv Teja Nagipogu、Shachi Dave、Shruti Gupta、Subhash Chandra Bose Gali、Vish Subramanian 和 Partha Talukdar。2021年。MURIL：印度语言的多语言表示。Keshav Kolluru、Martin Rezk、Pat Verga、William W. Cohen 和 Partha Talukdar。2021年。 多语言事实链接。Rik Koncel-Kedziorski、Dhanush Bekal、Yi Luan、Mirella Lapata 和 Hannaneh Hajishirzi。2019 年。利用图变换器从知识图谱生成文本。Jens Lehmann、Robert Isele、Max Jakob、Anja Jentzsch、Dimitris Kontokostas、Pablo N. Mendes、Sebastian Hellmann、Mohamed Morsey、Patrick van Kleef、Sören Auer 和 Christian Bizer。2015 年。 DBpedia——从维基百科中抽取的大规模多语言知识库。 语义网，6(2):167–195。Ren Li、Dong Li、Jianxi Yang、Fangyue Xiang、Hao Ren、Shixin Jiang 和 Luyi Zhang。2021年。 通过实体相关注意力神经模型联合抽取实体和关系。 信息科学，581:179–193。Sirui Li、Kok Kai Wong、Dengya Zhu 和 Chun Che Fung。2022 年。 利用图摘要改进知识图谱问答。Mausam、Michael Schmitz、Stephen Soderland、Robert Bart 和 Oren Etzioni。2012 年。 用于信息抽取的开放式语言学习。载于《*2012* 年自然语言处理与计算自然语言学习经验方法联合会议论文集》，第 523–534 页，韩国济州岛。计算语言学协会。Gowtham Ramesh、Sumanth Doddapaneni、Aravinth Bheemaraj、Mayank Jobanputra、Raghavan AK、Ajitesh Sharma、Sujit Sahoo、Harshita Diddee、Mahalakshmi J、Divyanshu Kakwani、Navneet Kumar、Aswin Pradeep、Srihari Nagaraj、Kumar Deepak、Vivek Raghavan、Anoop Kunchukuttan、Pratyush Kumar 和 Mitesh Shantadevi Khapra。2021年。Samanantar：规模最大的公开可用平行语料库集合，涵盖 11 种印度语言。Saurabh Srivastava、Mayur Patidar、Sudip Cho wdhury、Puneet Agarwal、Indrajit Bhattacharya 和 Gautam Shroff。2021年。 利用机器翻译和

多任务学习。载于《欧洲计算语言学协会第*16*届会议论文集：主卷》，第 3428–3439 页，线上举行。计算语言学协会。

Dianbo Sui、Yubo Chen、Kang Liu、Jun Zhao、Xian- grong Zeng 和 Shengping Liu。2020a年。使用集合预测网络联合进行实体与关系抽取。

Dianbo Sui、Yubo Chen、Kang Liu、Jun Zhao、Xian- grong Zeng 和 Shengping Liu。2020b年。使用集合预测网络联合进行实体与关系抽取。

Denny Vrandeˇci´c 和 Markus Krötzsch。2014年。维基数据：一个免费的协作式知识库。《*ACM* 通讯》，57(10)：78–85。

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。

Zhiheng Yan、Chong Zhang、Jinlan Fu、Qi Zhang 和 Zhongyu Wei。2021年。用于联合实体与关系抽取的分区过滤网络。载于*2021*年自然语言处理实证方法会议论文集，第185–197页，线上举行及多米尼加共和国蓬塔卡纳。计算语言学协会。

Daojian Zeng、Ranran Haoran Zhang 和 Qianying Liu。2019年。Copymtl：通过多任务学习联合抽取实体和关系的复制机制。

Sheng Zhang、Kevin Duh 和 Benjamin Van Durme。2017a。MT/IE：使用神经序列到序列模型进行跨语言开放信息抽取。载于第*15*届欧洲计算语言学协会分会会议论文集：第*2*卷，短篇论文，第64–70页，西班牙瓦伦西亚。计算语言学协会。

Sheng Zhang、Kevin Duh 和 Benjamin Van Durme。2017b。MT/IE：使用神经序列到序列模型进行跨语言开放信息抽取。载于第*15*届欧洲计算语言学协会分会会议论文集：第*2*卷，短篇论文，第64–70页，西班牙瓦伦西亚。计算语言学协会。

Zexuan Zhong 和 Danqi Chen。2020 年。一种令人沮丧地简单的实体和关系抽取方法。

XiaohanZou.2020. 一项关于知识图谱应用的调查. 物理学杂志：会议系列,1487(1):012016.

A 示例附录

A.1 尾实体对齐

从译文中提取的实体会与真实尾实体对齐，以衡量模型在测试集上的表现。所谓对齐，是指为每个提取出的实体分配一个真实标注尾实体，且不重复分配。与真实标注没有任何重叠的部分提取实体会被忽略。部分真实标注实体可能不会分配到任何提取实体，从而导致召回率降低。分配依据相似度分数和阈值进行。两个实体之间的相似度分数由 GloVe 向量的余弦相似度之和与词项的交并比计算得出。当阈值为 0.7 时，我们的精确率为 0.54，召回率为 0.77。

A.2 附加数据集统计

图 4展示了数据集中出现频率最高的前 30 种关系的分布。图 5 展示了数据集中各语言所占的比例。可以看出，该数据集在关系和语言方面都存在严重的类别不平衡。

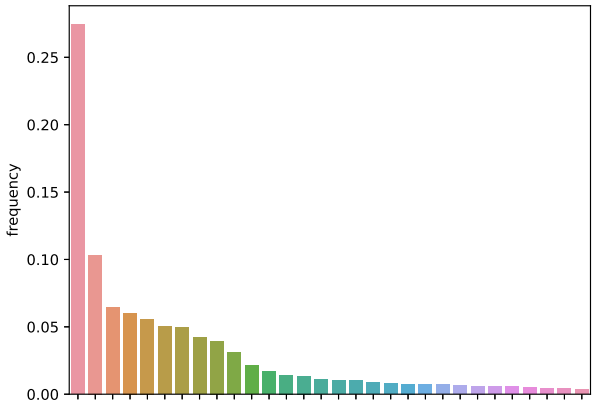


图4：数据集中出现频率最高的前30种关系的分布

A.3 实现细节

两阶段和端到端生成式架构均在 NVIDIA GeForce RTX 2080 Ti 显卡上训练。对于两阶段方法，

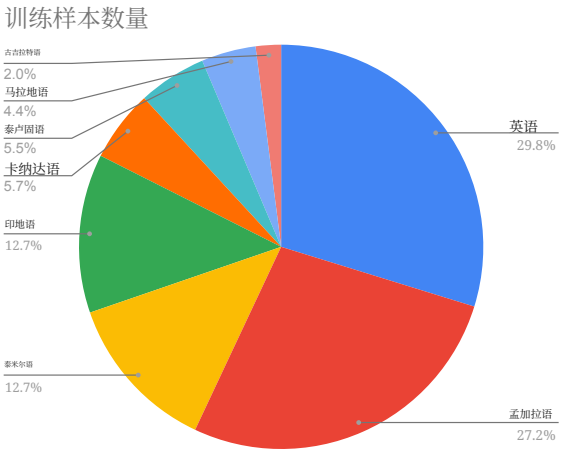


图5：训练集中8种语言的分布

唯一需要训练的模块是关系预测模块。分类器以谷歌的 MURIL 编码器模型为基础，该模型包含 12 个编码层，输出维度为 768。训练期间，MURIL 的第 12 层以及前馈网络中的各层均不被冻结。使用 Adam 优化器，初始学习率为 1e-4，并采用步长为 2、伽马值为 0.3 的步长调度。每批次使用 16 条事实进行训练，以优化交叉熵损失。使用类别的逆对数频率作为交叉熵损失的权重，以抵消数据集中的类别不平衡。在 1 张 GPU 卡上训练关系预测模块耗时 5 小时。

对于生成式方法，我们使用预训练的 mT5 模型，并在所有实验中对其进行 5 个训练轮次的微调。学习率为 0.001，权重衰减为 0.01。为防止训练数据过拟合，丢弃率设为 0.1。在生成过程中，我们使用 Adafactor 优化器来优化交叉熵损失。

A.4 开放信息抽取分析

我们尝试使用斯坦福的开放信息抽取器，从其他语言文本的英语译文中进行抽取。即使考虑到翻译损失，从开放信息抽取中提取的事实仍然没有用，因为关系和实体过于具体。图6是源句“Sindhu is the second Indian after Saina Nehwal to win in badminton after 2012”的一个示例。

```
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'after Saina Nehwal to win after 2012'}
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'after Saina Nehwal to win in badminton'}
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'Saina Nehwal win after 2012'}
|- {'subject': 'Saina Nehwal', 'relation': 'win in', 'object': 'badminton'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'Saina Nehwal to win'}
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'Saina Nehwal to win in badminton'}
|- {'subject': 'Sindhu', 'relation': 'is', 'object': 'Indian'}
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'after Saina Nehwal to win in badminton after 2
012'}
|- {'subject': 'Saina Nehwal', 'relation': 'win medal after', 'object': '2012'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'Saina Nehwal win'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'after Saina Nehwal win after 2012'}
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'after Saina Nehwal win in badminton'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'Saina Nehwal win in badminton'}
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'Saina Nehwal win'}
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'Saina Nehwal win in badminton after 2012'}
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'Saina Nehwal to win'}
|- {'subject': 'Sindhu', 'relation': 'is', 'object': 'second Indian'}
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'after Saina Nehwal win in badminton after 201
2'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'Saina Nehwal win after 2012'}
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'Saina Nehwal to win after 2012'}
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'after Saina Nehwal win'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'Saina Nehwal to win in badminton'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'after Saina Nehwal to win in badminton'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'after Saina Nehwal to win after 2012'}
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'after Saina Nehwal win after 2012'}
|- {'subject': 'Saina Nehwal', 'relation': 'win', 'object': 'Olympic medal'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'after Saina Nehwal to win in badminton after 2012'}
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'Saina Nehwal win in badminton'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'after Saina Nehwal win in badminton after 2012'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'after Saina Nehwal win in badminton'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'Saina Nehwal win in badminton after 2012'}
|- {'subject': 'Saina Nehwal', 'relation': 'win medal in', 'object': 'badminton'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'after Saina Nehwal to win'}
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'Saina Nehwal to win in badminton after 2012'}
|- {'subject': 'Saina Nehwal', 'relation': 'win after', 'object': '2012'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'Saina Nehwal to win after 2012'}
|- {'subject': 'Saina Nehwal', 'relation': 'win', 'object': 'medal'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'after Saina Nehwal win'}
|- {'subject': 'Sindhu', 'relation': 'is second Indian', 'object': 'after Saina Nehwal to win'}
|- {'subject': 'Sindhu', 'relation': 'is Indian', 'object': 'Saina Nehwal to win in badminton after 2012'}
```

图6：开放信息抽取的输出示例