arXiv:2301.05494v1 [cs.CL] 13 Jan 2023

使用世界语言和适配器融合进行多语言值得核查声明检测*

Ipek Baris Schlicht[1,2], Lucie Flek[3], 和 Paolo Rosso[1]

[1] PRHLT研究中心，瓦伦西亚理工大学，西班牙[2] DW Innovation，德国[3] CAISA实验室，马尔堡大学，德国{ibarsch@doctor,prosso@dsic}.upv.es，lucie.flek@uni-marburg.de

摘要。 值得核查性检测旨在识别值得事实核查人员调查的声明。非世界语言资源匮乏以及模型学习成本，仍是创建支持多语言值得核查性检测的模型所面临的主要挑战。本文提出在一组世界语言上进行跨语言训练适配器，并通过适配器融合将其结合起来，以检测在全球多种语言中出现的声明。（1）世界语言拥有大量可用标注者，且适配器模型存储效率高，因此这种方法更具成本效益。模型可以更频繁地更新，从而保持最新状态。（2）适配器融合能够深入揭示各适配器模型对特定语言的影响，并支持对其进行解读。在我们的基准任务中，所提出的解决方案表现通常优于顶尖的多语言方法。

关键词： 事实核查 · 可核查性检测 · 多语言 · 适配器。

1 引言

人们对支持事实核查人员和调查记者的自动化工具的需求日益增长，尤其是在突发新闻或争议性新闻出现时 [11,19]。识别并优先筛选适合进行事实核查的声明，即所谓的可核查声明（CW 声明）检测，是此类自动化系统的首要任务 [9]。这项任务有助于引导事实核查人员关注可能造成危害的声明，以便进一步调查。如表 1 所示，CW 声明是可验证的，关乎公众利益，且可能引发情绪反应 [20]。该领域的大多数研究都聚焦于单语方法，主要使用英语数据集来训练模型。对于调查非英语资源的事实核查人员而言，多语言能力已成为一项必不可少的特性 [8]。

尽管 Transformer 在多项多语言应用中取得了有竞争力的成果 [4,3]，包括 CW 检测 [21,18], ，但 CW 任务仍面临两项主要挑战：（1）由于任务复杂，目前公开的.....

* 本文已被 ECIR 2023 接收。

表 1：CT21 中一个值得核查的声明示例。

那些谎称疫苗 100% 安全、嘲笑反疫苗者的人，将不得不依赖那些拒绝宣扬官方谎言的人，才能承认大规模接种疫苗*确实*至关重要。挪威有 23 人在接种新冠疫苗后死亡： [URL] 来源： [MENTION]

现有的多语言数据集。要更新多语言模型，使其能够检测近期出现的声明，需要进行数据标注并及时重新训练。寻找事实核查专家来标注样本并非易事，在低资源语言中尤其如此。（2）为每种受支持的语言分别存储独立的微调模型副本，需要海量存储空间。由于非营利组织和媒体机构的预算有限，频繁更新模型并不可行。

世界语言（WL）是指在多个国家使用的语言，而非母语者仍可将其作为外语彼此交流。英语、阿拉伯语和西班牙语都是世界语言（WL）的例子 [4]。显然，与低资源语言相比，为世界语言收集样本时更容易找到专家标注人员。

作为一种资源效率更高的替代方案，适配器 [14,31] 近期被提出，可用于替代对 Transformer 模型进行完全微调。适配器是轻量、模块化的神经网络，能够以更少的参数学习任务，并在不同任务 [14,31,22] 和语言 [24] 之间迁移知识。与完全微调的预训练模型相比，微调后的适配器占用的存储空间更少。

本文提出使用适配器对 WL 数据集进行跨语言训练，以缓解资源匮乏问题并提供经济高效的解决方案。我们首先为每种 WL 训练任务适配器（TA） [23] ，并引入可解释的适配器融合（AF） [22] ，将各 WL 的 TA 结合起来，从异质来源中有效迁移知识。

本文的贡献总结如下 [5]:

– 我们对多语言 CW 声明检测任务中的 WL AF模型进行了广泛分析，并通过零样本学习（即训练期间未见过目标语言）评估这些模型。结果表明，这些模型的表现优于单语任务适配器和完全微调的模型。在某些语言中，它们的表现也超过了相关的最佳方法。此外，WL AF模型能够实现零样本学习。– 我们构建了一套评估方法，用于量化模型在不同语言中针对全球／本地主题声明的表现。我们用于整理评估集的方法也可用于评估其他多语言社交任务。– 我们还进行了详细的消融研究，以了解 AF模型的局限性及其在 CW 任务中的行为。

[4] https://bit.ly/3eMIZ9q [5] 我们在 https://bit.ly/3rH6yXu 分享了源代码。

2 相关工作

2.1 识别可核查性声明

早期研究采用特征工程，并结合机器学习模型和神经网络模型，识别政治辩论和演讲中的 CW 声明。Claim-Buster [13,12] 是最早采用支持向量机（SVM）分类器的方法之一，该分类器使用词汇和情感特征集进行训练。Gencheva 等人 [7]结合政治辩论中的句子级信息和上下文信息，训练神经网络架构。Lespagnol 等人 [17] 采用信息营养标签 [6] 和词嵌入作为特征。Vasileva 等人 [35] 利用来自不同事实核查组织的多任务学习，判断一项陈述是否属于 CW 声明。Jaradat 等人 [15] 使用 MUSE 词嵌入，支持英语和阿拉伯语的 CW 检测任务。

CheckThat!(CT) 自 2020 [2,21,18] 起组织多语言 CW 检测任务。他们每年支持更多语言，提交的多语言系统数量也不断增加。Schlicht 等人 [28]提出了一种支持数据集中所有语言的模型。他们使用多语言句子Transformer [26]，随后在语言识别任务上对模型进行联合微调。类似地，Uyangodage 等人 [34] 在一个包含各语言均衡样本的数据集上对 mBERT 进行了微调。最近，Du 等人 [5] 通过插入提示，在 CW 检测任务上对 mT5 [37] 进行了微调，并将其与多项相关任务和语言联合训练。所有上述方法都仅限于其训练所涵盖的语言。

Kartal 和 Kutlu [16] 评估了 mBERT 在英语、阿拉伯语和土耳其语上的零样本学习能力，并观察到跨语言训练中的性能下降。跨语言训练需要更有效的方法。此外，这些方法均未解决资源效率问题。本文中，我们使用适配器，并仅在 WL 上训练，以提高资源效率；同时评估未见过的语言，并利用适配器融合了解数据集中哪些目标语言能够从 WL 迁移的知识中获益。

2.2 适配器

适配器已成功应用于预训练模型，以便高效微调并适用于各种应用。早期研究 [14,22] 将适配器用于英语任务适配。Pfeiffer 等人 [22] 提出了一个适配器融合模块，用于学习如何组合任务适配器，作为多任务学习的替代方案。一些研究人员 [33,24,25]利用语言专用适配器进行跨语言迁移。本文以 [22,25]的研究为基础。我们利用跨语言训练集学习任务适配器，然后使用适配器融合有效地组合这些适配器，并说明哪些源语言任务适配器对目标语言有效。

表2：CT21 和 CT22 的统计数据。

		CT21		CT22	
	拆分	总数	%经核查的说法	总数	%经核查的说法
ar	训练集	3444	22.15	2513	38.28
	Dev	661	40.09	235	42.55
	测试集	600	40.33	682	35.28
es	训练集	2495	8.02	4990	38.14
	Dev	1247	8.74	2500	12.20
	测试集	1248	9.62	4998	14.09
en	训练集	822	35.28	2122	21.07
	Dev	140	42.86	195	22.56
	测试集	350	5.43	149	26.17
tr	Test	1013	18.07	303	4.62
保加利亚语	测试集	357	21.29	130	43.85
荷兰语	测试集	-	-	666	47.45

表3：全球和本地主题统计

		CT21		CT22	
	拆分	总数	%值得核查	总数	%值得核查
ar	全球	269	43.12	116	46.55
	本地	40	42.5	-	-
es	全球	1208	9.93	148	22.97
	本地	-	-	917	12.43
荷兰语	全球	-	-	103	56.31
	本地	-	-	15	46.67
英语	全球	349	5.44	14	28.57
tr	全球	887	8.91	-	-
保加利亚语	全球	356	21.35	25	32

3 个数据集

3.1 任务数据集

我们为实验寻找了使用 WL 和其他语言的多语言数据集。CT21 [29] 和 CT 22 [18] 是唯一符合这一要求的公开数据集。CT21 包含英语、阿拉伯语、土耳其语、西班牙语和保加利亚语样本。该数据集主要涉及新冠肺炎事件，但西班牙语样本除外，它们只关注政治。与 CT21 相比，CT22 还包含荷兰语样本。该数据集中的英语、阿拉伯语、保加利亚语和荷兰语样本基于有关新冠肺炎19 [1] 的语料库。研究人员收集了新的土耳其语和西班牙语样本。西班牙语样本通过 CT21 进行了扩充。

数据集的统计信息见表2。数据集包含英语、西班牙语和阿拉伯语这几种 WL。两个数据集都存在类别不平衡问题，即各语言中的 CW样本占比较低。尽管英语被视为一种高资源语言，但英语样本数量明显少于其他语言的样本数量。由于数据集中的部分样本可能重叠，我们按数据集分别开展了研究实验。

3.2 主题评估数据集

有些国家在文化上存在差异，或可能有不同的政治议程，因此各国的 CW 主题可能有所不同。例如，在 COVID-19 大流行期间，疫苗接种是全球范围内与 CW 相关的主题，但某些 COVID-19 谣言只有少数群体相信 [30]。理想的多语言系统应当既能处理全球主题，也能处理局部主题。

全球主题是指存在于所有数据集的所有语言中的主题，而本地主题仅存在于一种语言中。我们创建了专用数据集，用于评估 WL 模型在全球和

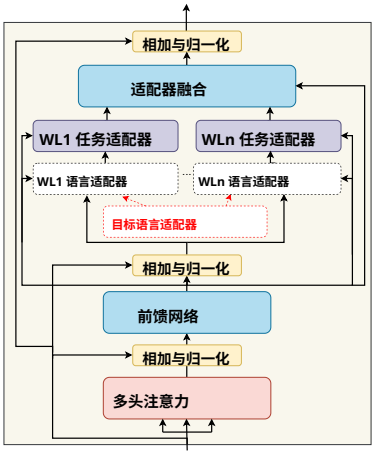


图1：变换器中 WL+AF 模型的架构。WL+AF+LA 是将语言适配器（虚线框所示）堆叠到任务适配器中的设置。

局部主题中识别 CW 声明的性能。该评估数据集仅包含全球和局部主题，创建过程如下。为了学习跨数据集的主题，我们首先将数据集翻译成英语，并使用句子Transformer [26]对其进行编码。其次，我们使用 BERTopic [10] 在 WL 训练数据集上进行主题建模，为所有语言的测试样本分配主题。将主题存在于所有 WL 语言中的测试样本添加到评估集中，作为可能的全球主题。BERTopic 将与已学习主题无关的主题标记为 -1。为了筛选具有局部主题的样本，我们从评估集中选出标记为 -1 的样本，并对其重新进行主题建模。将主题未出现在任何其他语言的测试数据集中的测试样本（即局部主题）添加到局部主题评估集中。局部和全球评估集的统计数据见表3。

4 方法

4.1 世界语言适配器融合

本节介绍 WL+适配器融合模型。WL+适配器融合模型是包含名为适配器的附加模块的 Transformer。在训练期间，仅训练适配器，并冻结预训练模型的权重。因此，这是一种参数高效方法。我们对两种类型的 WL+适配器融合模型进行了实验。WL+适配器融合是组合 WL 任务适配器的标准设置。WL+适配器融合+语言适配器通过堆叠任务适配器，加入了额外的语言适配器（LAs） [24]。我们在图 1 中展示了这些模型的架构。该架构以文本为输入，输出一个用于描述输入内容核查价值的概率分数。

Transformer 编码器。为了提供跨语言表征，我们使用多语言 BERT（mBERT）[4] 和 XLM-Roberta（XLM-R）[3][6]。这是因为相关研究此前使用过这些 Transformer [34,1]。此外，已有公开可用的适用于这些 Transformer 的语言适配器。在使用 Transformer 对文本进行编码之前，我们首先清除其中的网址、提及和转推，然后将文本截断或填充至 128 个词元。

任务适配器。标注可能会受到各国不同文化背景和事件的影响，例如 CW 声明。此外，有些标注可能由遵循人工事实核查流程的记者创建，而另一些则源自众包 [21,18] 基于这些原因，我们将每个 WL 数据集视为一个独立任务。然后，我们通过在相应的 WL 数据集上进行优化来获得任务适配器。任务适配器由一个下投影层、一个 ReLU 激活和一个上投影层组成 [23]。每个 WL 的任务适配器都在相应的数据集上进行微调。

适配器融合。为了利用不同任务适配器中的知识，对未见过的语言或新主题中的样本进行预测，我们需要有效地将它们结合起来。适配器融合是一种组合多个任务适配器的方法，并能缓解多任务学习中的常见问题，例如灾难性遗忘，以及为支持新任务而对模型进行完整的重新训练 [22]。适配器融合包含一个注意力模块，用于学习如何动态组合任务适配器中的知识。通过分析适配器融合的注意力权重，我们可以了解哪些源任务适配器对测试预测的贡献更大。我们将基于 WL 训练的任务适配器与适配器融合相结合。随后，在 WL 混合数据集上对适配器融合进行微调，以便用于目标语言。

语言适配器。为了从各种跨语言来源学习 CW 声明检测，我们应确保模型学到的不是语言分类任务，而是如何识别 CW 声明。因此，我们需要将语言特定知识与任务特定知识分离。近期一项研究 [24]表明，将在一种源语言中执行的任务所涉及的知识迁移到另一种语言时，语言适配器取得了更好的效果。为了将语言特定信息编码到变换器模型中，并考察语言适配器对 AF模型性能的影响，我们使用了 Adapter Hub [23]中的语言适配器。这些语言适配器的架构与任务适配器类似。不过，这些语言适配器是通过掩码语言模型目标，在未标注的维基百科 [24] 数据上预训练的。在使用任务适配器学习任务特定变换的同时，每个世界语言适配器都堆叠在相应的任务适配器上。语言适配器的权重保持冻结，因此不会再次进行训练。在目标任务的推理过程中，源语言适配器会被目标语言适配器替换。

4.2 实现细节

我们从 Huggingface [36] 下载预训练的 Transformer。在英语、阿拉伯语和西班牙语训练样本（WL）上微调 TA，并通过在其他

[6] 我们使用模型的基础版本，该版本由 12 层组成。

语言上进行测试，评估模型的零样本学习能力。对于数据集 [24], 中阿拉伯语、英语、西班牙语和土耳其语样本所用的 LA，我们从 Adapter Hub [23] 下载了两种 Transformer 的预训练 LA。但是，Adapter Hub 中没有保加利亚语和荷兰语的 LA，因此我们使用英语 LA。

我们使用 AdapterHub [23] 的训练器进行完整微调和适配器调优。我们将训练轮次设为 10，针对 CT21 使用 1e-4 的学习率，针对 CT22[7], 使用 2e-5 的学习率，批量大小设为 16。我们根据数据集指标选用开发集上表现最好的模型。我们使用不同的随机种子重复实验十次，使用的显卡为 NVIDIA GeForce RTX 2080。

5 基线

我们将 WL AF模型的性能与多种基线进行比较：

– **表现最佳的系统：**我们选择了 CT21 和 CT22 上表现最佳的系统，这些系统使用单一模型进行多语言CW检测，而非包含针对特定语言的模型。Schlicht 等人 [28] 是 CT21 上的亚军 [8]系统。其原始实现使用句子Transformer，并在所有语言的数据集上对模型进行微调。为确保比较公平，我们将语言识别任务设置为 WL，并用 mBERT 和 XLM-R 替换句子 Transformer。CT22 上的最先进模型基于 mT5-xlarge [5], 这一多任务文本到文本Transformer。与 Schlicht 等人的模型一样，该模型通过多任务学习在所有语料库上进行微调。由于计算资源有限，我们无法仅使用 WL 对该模型进行微调。我们报告了来自 [5]的结果。

– **完全微调（FFT）变换器（单语言）：**我们在 WL 的单一语言数据集上进行微调，以评估跨语言学习的效率：AR+FFT、EN+FFT 和 ES+FFT。我们还加入 BG+FFT、ES+FFT 和 NL+FFT，作为零样本学习的基线。

– **任务适配器（单语言）：**这些基线包含一个任务适配器，后接一个 LA，这是使用适配器进行跨语言迁移学习的常用设置 [24]：AR+TA+LA、EN+TA+LA、ES+TA+LA。将我们的模型与这些基线进行比较，有助于了解使用 WL 进行跨语言训练是否有效。

– **其他 WL 模型：**我们将 AF 模型与一种包含在 WL 上训练的任务适配器的模型（WL+TA）进行比较。除了这一基线模型外，为了确定是否需要复杂的融合方法，我们还使用 WL+TA+LA+Mean，该方法对 AR+TA、EN+TA 和 ES+TA 的预测结果取平均值。最后，我们分析了在多个 WL 上进行的适配器调优，并将其与 mBERT 的完全微调进行比较：WL+FFT。

[7] 2e-5 在 CT22[8] 的开发集上取得了更好的结果。BigIR 是当前最先进的方法，但没有相关论文或代码对该系统进行描述。

表 4：WL 中 CW 检测在 CT21 上的 MAP 分数和在 CT22 上的 F1分数。**粗体** 表示最佳分数，下划线表示第二佳分数。总体而言，AF模型在多种语言上表现良好，而其他模型的性能则容易受到训练集特征的影响。

		CT21			CT22			
		ar	es	en	ar	es	en	avg
	Du 等人[5]	-	-	-	**62.8** 57.1		**51.9**	-
mBERT	阿拉伯语+快速傅里叶变换	50.17	15.78	6.03	55.52	17.85	42.92	31.38
	西班牙语+快速傅里叶变换	41.22	20.30	6.80	37.30	54.05	45.28	34.16
	英语+快速傅里叶变换	51.63	15.90	10.80	40.97	22.49	44.29	31.01
	AR+文本增强+拉丁语	58.20	19.97	8.62	18.13	17.61	45.73	28.04
	ES+文本增强+拉丁语	48.07	18.93	11.06	37.30	54.05	45.73	35.86
	英语+文本增强+拉丁语	50.81	40.81	**21.21**	56.39	24.63	12.93	34.46
	弱标签+快速傅里叶变换	47.93	51.50	13.85	51.50	63.20	39.94	44.65
	Schlicht 等人[28]	51.51	31.04	7.87	45.93	66.48	34.18	39.50
	弱标签+文本增强	53.77	46.58	14.44	39.54	62.69	37.19	42.37
	弱标签+文本增强+拉丁语+平均值	54.89	35.72	12.96	0.00	33.21	51.03	31.30
	弱标签+适配器融合	55.13	46.29	16.05	36.45	64.32	39.73	42.96
	弱标签+适配器融合+拉丁语	55.32	46.58	15.66	39.87	64.64	37.27	43.22
XLM-R	阿拉伯语+快速傅里叶变换	43.55	12.80	5.88	38.72	6.83	41.29	24.85
	西班牙语+快速傅里叶变换	41.22	15.90	6.80	40.25	21.69	43.51	28.23
	英语+快速傅里叶变换	43.64	10.49	5.64	31.46	64.66	45.69	33.60
	AR+文本增强+拉丁语	58.16	25.87	7.64	6.93	0.06	1.40	16.68
	西班牙语+文本增强+拉丁语	50.27	**52.51**	11.53	41.24	65.45	38.79	43.30
	英语+文本增强+拉丁语	56.39	24.63	12.93	12.82	0.06	28.74	22.60
	弱标签+快速傅里叶变换	47.93	13.85	6.37	44.53	64.75	50.93	38.06
	Schlicht 等人[28]	51.56	21.61	7.32	42.59	**67.13**	36.40	37.77
	弱标签+文本增强	58.02	50.76	11.53	42.91	63.36	31.69	43.05
	弱标签+文本增强+拉丁语+平均值	**59.32**	46.11	10.49	25.32	35.55	35.28	35.35
	弱标签+适配器融合	58.39	49.42	13.29	39.84	65.66	46.96	**45.59**
	弱标签+适配器融合+拉丁语	58.83	47.26	16.06	35.17	65.80	43.46	44.43

6 结果与讨论

本节将展示并分析 WL AF（+LA）模型的结果。我们比较这些模型在以下任务中的 CW 检测表现：（1）WL；（2）零样本语言；（3）局部和全局主题。最后，我们比较 WL+AF 和 WL+AF+LA 的表现，以探究 LA 是否有助于提升模型性能。

如表 4 所示，如果训练集规模足够大，或者训练集与测试集的语言相同，那么仅使用单一语言训练的模型也能够在其他 WL 上取得良好表现。此外，Schlicht 等人[28]和 WL+快速傅里叶变换仅在 CT22 上表现良好；总体而言，AF模型在多种语言上表现出色。Du 等人 [5] 在 CT22 的阿拉伯语和英语样本上优于 AF模型，但在西班牙语样本上的表现较差。

如表 5 所示，AF模型在零样本语言的目标集上取得了良好的结果。这表明，利用适配器融合多个来源

表5：零样本语言中 CW 检测的**CT21** MAP和**CT22** F1值。**粗体** 表示最高分，下划线表示次高分。AF模型表现出色，平均得分甚至超过了 WL+快速傅里叶变换和 Schlicht 等人 [28] 的方法，以及部分单语方法。

		CT21		CT22			
		tr	bg	tr	bg	nl	avg
	Du 等人[5]	-	-	17.3	61.7	**64.2**	-
	保加利亚语+快速傅里叶变换	-	35.64	-	57.16	-	-
	土耳其语+快速傅里叶变换	28.47	-	14.66	-	-	-
	荷兰语+快速傅里叶变换	-	-	-	-	49.80	-
mBERT	阿拉伯语+快速傅里叶变换	30.14	22.12	10.71	54.84	45.55	32.67
	西班牙语+快速傅里叶变换	23.17	24.15	8.19	47.67	33.82	27.4
	英语+快速傅里叶变换	42.80	33.20	13.57	54.58	54.32	39.69
	阿拉伯语+文本增强+拉丁语	58.08	26.76	8.04	57.43	58.99	41.86
	西班牙语+文本增强+拉丁语	49.09	28.09	8.58	47.67	42.16	35.12
	英语+文本增强+拉丁语	54.16	44.98	8.19	33.92	33.82	35.01
	弱标签+快速傅里叶变换	27.61	24.29	10.90	58.53	29.03	30.07
	Schlicht 等人[28]	27.81	24.41	7.86	51.94	29.33	28.27
	弱标签+文本增强	54.11	37.04	9.73	48.02	36.95	37.17
	弱标签+文本增强+拉丁语+平均值	62.32	34.52	12.02	47.17	39.91	39.19
	弱标签+适配器融合	50.46	39.80	9.63	53.55	38.76	38.44
	弱标签+适配器融合+拉丁语	50.94	40.27	9.73	52.75	43.07	39.35
	保加利亚语+快速傅里叶变换	-	24.68	-	43.47	-	-
	土耳其语+快速傅里叶变换	24.57	-	**18.90**	-	-	-
	荷兰语+快速傅里叶变换	-	-	-	-	58.61	-
XLM-R	阿拉伯语+快速傅里叶变换	23.40	21.86	11.62	45.10	38.73	28.14
	西班牙语+快速傅里叶变换	23.17	24.15	9.07	62.72	31.24	30.07
	英语+快速傅里叶变换	22.63	21.76	18.43	49.25	45.62	31.54
	阿拉伯语+文本增强+拉丁语	56.19	21.37	0.48	9.75	5.22	18.60
	西班牙语+文本增强+拉丁语	44.98	23.86	8.04	46.72	30.52	30.82
	英语+文本增强+拉丁语	58.38	41.61	15.19	8.42	25.83	29.89
	弱标签+快速傅里叶变换	27.61	24.29	14.02	63.79	36.34	33.21
	Schlicht 等人[28]	25.86	22.19	9.55	53.75	24.44	27.16
	弱标签+文本增强	59.37	39.72	15.60	**66.14**	35.98	43.36
	弱标签+文本增强+拉丁语+平均值	63.65	32.28	9.90	31.27	19.91	31.40
	弱标签+适配器融合	57.46	46.86	12.73	59.12	40.83	43.4
	弱标签+适配器融合+拉丁语	**61.74**	41.78	17.77	63.88	37.59	**44.55**

有助于知识迁移，并且优于其他融合方法 WL+TA+LA+平均值。值得注意的是，Du 等人 [5] 在训练数据集的所有样本上进行了训练，因此不具备零样本学习能力。尽管 Du 等人在荷兰语样本上表现更好，但 AF模型在其他语言上也能取得相近的结果。在资源效率方面，如表 8 所示，与 WL+FFT 和 mT5 相比，AF模型存储新权重所需的空间更少，因此对于预算有限的新闻编辑部而言，AF模型比更新 mT5 更适用。

我们比较了使用多种 WL 训练的模型在识别有关全局或局部主题的 CW 声明方面的性能。我们使用第 3.2 节所述的评估集，以 F1 分数衡量这一实验。我们取这些模型的平均值

表6：各数据集在全球主题上的模型 F1分数。适配器训练比完全微调更有效。尽管 WL+文本增强在某些语言上的表现优于适配器融合模型，但适配器融合模型的平均表现更好。

	CT21				CT22						
	tr	es	ar	en	bg	es	ar	en	bg	nl	avg
WL+快速傅里叶变换	0.00	0.00	2.60	0.59	0.10	77.59	**52.13**	**52.05**	**48.52**	38.53	27.21
Schlicht 等人[28]	18.15	17.92	58.84	7.18		19.75	**81.81**	48.12	28.33	36.55	30.34	34.70
弱标签+文本增强+拉丁语	**52.37**	38.09	61.01	**13.58**	**38.44**	77.28	45.47	22.57	41.31	**45.09**	38.52	
弱标签+适配器融合	45.02	**41.57**	61.51	13.36	36.23	79.14	44.61	50.96	40.58	43.19	**45.62**	
弱标签+适配器融合+拉丁语	48.52	37.44	**61.79**	13.38	28.72	77.95	42.84	44.23	41.09	43.99	44.00	

表7：各数据集在局部主题上的模型 F1分数。AF模型的结果与 WL+文本增强相近，且优于 WL+快速傅里叶变换。

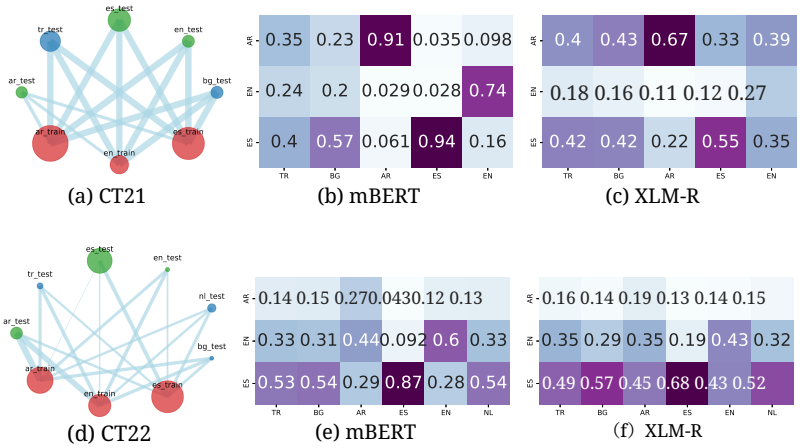
	CT21	CT21		
	ar	es	nl	avg
WL+快速傅里叶变换	0.93	60.64	31.11	30.89
Schlicht 等人[28]	46.80	**64.65**	14.44	41.96
弱标签+文本增强+拉丁语	**50.88**	61.65	38.47	**50.33**
弱标签+适配器融合	48.15	62.01	34.38	48.18
弱标签+适配器融合+拉丁语	46.41	63.07	**41.16**	50.21

表 8：各模型的训练参数数量和文件大小比较。mT5 比 mBERT 和 XLM-R 更大。

模型	基础模型参数	模型大小	
WL+快速傅里叶变换	mBERT	178 M	711.5 MB
	XLM-R	278 M	1.1 GB
Schlicht 等人[28]	mBERT	179 M	716.3 MB
	XLM-R	279 M	1.1 GB
文本增强 &弱标签+文本增强+拉丁语	mBERT	150 万	6 MB
	XLM-R	150 万	6 MB
AF	mBERT	22 M	87.4 MB
	XLM-R	22 M	87.4 MB
LA	mBERT	-	147.78 MB
	XLM-R	-	147.78 MB
	mT5	3.7 B	15 GB

，这些模型分别采用 mBERT 和 XLM-R 编码，并分别在表6和表7中展示其在全局和局部主题上的结果。总体而言，对于这两类主题，AF模型的表现均优于 WL+快速傅里叶变换和 Schlicht 等人[28] 。然而，在 CT21 中预测阿拉伯语样本的局部陈述时，WL+文本增强的表现与 WL+适配器融合+拉丁语相近。

最后，我们比较 WL+AF和WL+AF+LA的性能，以探究 LA 是否能提升模型性能。我们计算了每项实验中各语言的 AF模型的 Fleiss Kappa分数。总体分数为



(a) CT21 (b) mBERT (c) XLM-R

(d) CT22 (e) mBERT (f) XLM-R

图2：左侧图像为 CT21（左上）和 CT22（左下）的主题关系图。在这些图中，节点大小随样本数量变化，边的粗细取决于重叠主题。热力图的 x 轴表示任务适配器，y 轴表示不同语言的测试样本。（b）、（c）为 mBERT 的注意力热力图，（e）、（f）为 XLM-R 的注意力热力图。主题分布和训练数据集的样本量会影响任务适配器的激活。尤其是，XLM-R 的文本增强比 mBERT 更敏感。

0.63，属于中等一致性。我们进一步研究了卡帕系数低于 0.5 的分歧。这些冲突主要出现在零样本语言和英语中，这些语言在两个数据集上的 CW样本数均最少。由于有时 WL+AF+LA 的表现优于 WL+AF，反之亦然，因此我们得出结论：LA 在我们的实验中并未发挥作用。预训练LA是在维基百科文本 [24] 上训练的。因此，它们可能无法捕捉社交文本的特性，而社交文本大多包含噪声。

7 进一步分析

本节将进一步分析适配器融合模型。我们将考察 AF注意力，并对模型的预测结果进行错误分析。

融合注意力的解读。AF模型可以帮助判断在将知识迁移到目标数据集时，哪些源语言任务适配器可能有用。此类分析有助于新闻编辑部的数据科学家决定应收集哪些弱标签，以更新模型并管理新的资源。为检查 WL+TA+LA+AF 的 AF 行为，我们对融合层中每个任务适配器所在层的 softmax概率取了平均值。概率越高，说明该任务对确定标签 [22] 越有用。此外，为了将注意力与源数据集关联起来，我们创建了一张图，展示源数据集之间的主题关系

表9：由弱标签文本增强适配器融合正确预测的 CW 声明+文本增强+适配器融合，示例为西班牙语和保加利亚语。每个示例中文本的顺序为：1）mBERT 上的可视化，2）XLM-R 上的可视化，3）红色文本为译文。模型对地缘政治实体（例如国家名称）的关注度高于其他实体类型。我们根据声明的积分梯度对其进行着色 [32]。

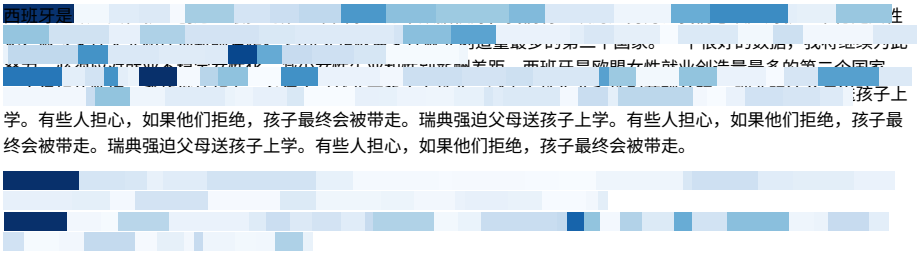


表10：AF模型在实体类型预测方面的表现，以平均 F1值和标准差表示。模型对地缘政治实体的预测比对其他实体类型更准确。

	地缘政治实体	组织	数字	人物
F1值	46.83 ± 17.37	40.44 ± 18.62	40.99 ± 22.52	37.88 ± 18.83

以及目标集。在图中，节点代表单语数据集；边代表源数据集与目标数据集之间的重叠主题，其权重由涉及重叠主题的样本百分比决定。节点大小根据样本量缩放。图2展示了两个数据集的图，以及 mBERT 和 XLM-R 任务适配器的注意力权重。主题分布和源数据集的规模会影响激活哪个任务适配器。XLM-R 的任务适配器对源数据规模和主题关系更加敏感。例如，CT22 中的西班牙语测试集与阿拉伯语和英语源数据集的联系较弱，XLM-R 的西班牙语任务适配器激活程度低于 mBERT 的任务适配器。

错误分析。最后，我们分析了两个适配器融合模型分类错误和分类正确的样本。如表 9 所示，我们注意到 AF模型关注地缘政治实体（GPE）。如表10所示，与包含其他类型实体的声明相比，模型更擅长对包含地缘政治实体（GPE）的声明进行分类。模型可以从 WL语料库中学习地缘政治实体（GPE）的重要性，因为其中的 CW样本包含了不可忽略数量的此类实体（例如，CT21 和 CT22 阿拉伯语源数据集中的地缘政治实体（GPE）分别占 76% 和 77%）。然而，模型无法预测需要本地语境的声明，尤其是在零样本语言中。此外，模型无法识别那些真实性会随时间变化、进而不再属于 CW 的声明。表11列出了一些示例。

训练效率。 我们测量了模型在数据集上完成一个训练轮次所需的训练时间。TA训练平均比完全微调少4分钟

表11：被适配器融合模型错误分类的 CW 声明。黑色文字为土耳其语声明。红色文字为这些声明的译文，蓝色文字则是对这些声明的解释。

我以前一直在写这件事，现在还会再写一遍。部长们的家属、他们的父母所需的全套医疗治疗（包括整形美容）都由国家预算承担，但为什么不为SMA患儿的治疗设立预算¨呢？¨UNK¨UNK.# 让国家拯救SMA I've always written this and I will write it again. 部长们的家属及其父母的全套医疗治疗（包括整形美容）由国家预算承担，但为什么不为SMA患儿的治疗设立预算.UNK.UNK¨# 让国家拯救 示例：一个需要本地语境的CW声明。SMA是一种影响儿童的疾病，SMA的治疗在土耳其是一个颇具争议的问题。新冠肺炎疫情期间，病例数达到3万.UNK.# Corona # 新冠肺炎 # coronavirus 新冠肺炎疫情期间的病例数达到3万[# Corona # 新冠肺炎 # coronavirus 一个CW声明示例，其真实性可能会随时间而改变。

调优。不过，不使用语言适配器的适配器融合训练耗时多 3 分钟，使用语言适配器的训练耗时则比 WL+快速傅里叶变换的训练时间多 9 分钟；后者约为 22 分钟。AdapterDrop [27] 等方法可以加快适配器融合训练。

8 结论与未来工作

本文研究了在世界语言上以高效方式跨语言训练适配器融合模型，以检测多种语言声明的核查价值。所提出的方案在多种语言上表现良好，即使在零样本学习场景下也是如此。得益于适配器融合，我们得以发挥适配器在特定语言上的效用。

一些任务适配器的注意力似乎取决于源数据集中的主题和样本分布。在世界语言之间确保主题平衡似乎十分重要。我们的错误分析结果表明，检测本地声明需要本地语境。我们建议通过注入背景知识来检测本地声明。

未来，我们希望研究如何将背景知识注入适配器，并在更多领域（例如战争领域）验证我们的结果，同时采用更多语言（如德语），并专注于零样本学习。

致谢 我们谨向匿名审稿人、Joan Plepi、Flora Sakketou、AkbarKarimi 和 Nico Para 致谢，感谢他们提出的建设性意见。Ipek Schlicht 的工作是 KID2项目（由 DW Innovation 牵头并由 BKM 共同资助）的一部分。Lucie Flek 的工作是 BMBF 项目 DeFaktS 和 DynSoDA 的一部分。Paolo Rosso 的工作是在 IBERIFIER（ INEA/CEF/ICT/A202072381931 n.2020-EU-IA-0252）、XAI Disinfodemics（ PLEC2021-007681）和 MARTINI（PCI2022-134990-2）的框架下开展的。

参考文献

1. Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Da San Martino, G., Abdelali, A., Durrani, N., Darwish, K., 等：应对新冠肺炎信息疫情：模拟记者、事实核查人员、社交媒体平台、政策制定者及社会的视角。载于：计算语言学协会论文集：EMNLP 2021。第611–649页（2021） 2. Barr´on-Cede˜no, A., Elsayed, T., Nakov, P., Martino, G.D.S., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A., Shaar, S., Ali, Z.S.：CheckThat! 2020概述：社交媒体声明的自动识别与核验。载于：CLEF。计算机科学讲义，第12260卷，第215–236页。施普林格（2020） 3. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzm´an, F., Grave, ´E., Ott, M., Zettlemoyer, L., Stoyanov, V.：大规模无监督跨语言表征学习。载于：计算语言学协会第58届年会论文集。第8440–8451页（2020） 4. Devlin, J., Chang, M., Lee, K., Toutanova, K.：BERT：用于语言理解的深度双向Transformer预训练。载于：Burstein, J., Doran, C., Solorio, T.（编）计算语言学协会北美分会2019年会议论文集：人类语言技术，NAACL-HLT 2019，美国明尼阿波利斯，2019年6月2日至7日，第1卷（长篇与短篇论文）。第4171–4186页。计算语言学协会（2019）。https://doi.org/10 1423 5 18653 1 19 1423, 10 18653 1 19 2022 . /v /n - https://doi.org/ . /v /n - . Du, S.M., Gollapalli, S.D., Ng, S.K.：NUS-IDS在CheckThat!中的工作：使用CheckThat5识别推文是否值得核查。CLEF工作论文集（2022） 6. Fuhr, N., Giachanou, A., Grefenstette, G., Gurevych, I., Hanselowski, A., J¨arvelin, K., Jones, R., Liu, Y., Mothe, J., Nejdl, W., Peters, I., Stein, B.：在线文档的信息营养标签。SIGIR论坛**51**(3), 46–66 (2017). https://doi.org/10.1145/3190580.3190588, https://doi.org/10.1145/3190580.3190588 7. Gencheva, P., Nakov, P., M`arquez, L., Barr´on-Cede˜no, A., Koychev, I.：一种用于检测政治辩论中值得核查主张的上下文感知方法。载于：Mitkov, R., Angelova, G.（编）国际会议论文集《自然语言处理最新进展》，RANLP 2017，8，2017 2 267 276 2017 10 26615 978 954 452 Varna, Bulgaria, September - . pp. – . INCOMA有限公司 ( ). https://doi.org/ . / - - - 049 6 037, 10 26615 978 954 452 049 6_037 - https://doi.org/ . / 8 - - - - . Ginsborg, L., Gori, P.：关于事实核查人员对新冠疫苗及虚假信息看法的调查报告。技术报告，欧洲数字媒体观测站（EDMO）（2021），https://cadmus.eui.eu/handle/1814 70917 9 / . Graves, L.：理解自动化事实核查的前景与局限（2018） 10. Grootendorst, M.：BERTopic：采用基于类别的TF-IDF流程进行神经主题建模。arXiv预印本 arXiv:2203.05794（2022） 11. Hassan, N., Adair, B., Hamilton, J.T., Li, C., Tremayne, M., Yang, J., Yu, C.：自动化事实核查的探索。载于：2015年计算新闻学研讨会论文集（2015） 12. Hassan, N., Arslan, F., Li, C., Tremayne, M.：迈向自动化事实核查：通过ClaimBuster检测值得核查的事实主张。载于：第23届ACM SIGKDD知识发现与数据挖掘国际会议论文集，加拿大哈利法克斯，2017年8月13日至17日。第1803–1812页。ACM（2017）。https://doi.org/10.1145/3097983.3098131, https://doi.org/10. 1145/3097983.3098131

13. Hassan, N., Li, C., Tremayne, M.：检测总统辩论中值得核查的事实主张。载于：Bailey, J., Moffat, A., Aggarwal, C.C., de Rijke, M., Ku- mar, R., Murdock, V., Sellis, T.K., Yu, J.X.（编）《第24届ACM信息与知识管理国际会议论文集》，CIKM 2015，澳大利亚维多利亚州墨尔本，2015年10月19日至23日。第1835–1838页。ACM（2015）。https://doi.org/10.1145/2806416.2806652，https://doi.org/10.1145/2806416.2806652 14. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.：面向自然语言处理的参数高效迁移学习。载于：ICML。《机器学习研究论文集》，第97卷，第2790–2799页。PMLR（2019） 15. Jaradat, I., Gencheva, P., Barr´on-Cede˜no, A., M`arquez, L., Nakov, P.：ClaimRank：检测阿拉伯语和英语中的可核查性声明。载于：《2018年计算语言学协会北美分会会议论文集：演示》。第26–30页。计算语言学协会，新奥尔良，路易斯安那州（2018年6月）。https://doi.org/10.18653/v1/N18-5006，https: // aclanthology.org/N18 5006 16 - . Kartal, Y.S., Kutlu, M.：分享前先三思：关于可核查性声明优先级排序的综合研究。IEEE计算社会系统汇刊（2022） 17. Lespagnol, C., Mothe, J., Ullah, M.Z.：利用信息营养标签和词嵌入估计信息的核查价值。载于：《第42届ACM SIGIR信息检索研究与发展国际会议论文集》。第941–944页。SIGIR'19，计算机协会，纽约，美国（2019）。https://doi.org/10.1145/3331184.3331298，https://doi.org/10. 1145/3331184.3331298 18. Nakov, P., Barr´on-Cede˜no, A., Martino, G.D.S., Alam, F., Struß, J.M., Mandl, T., M´ıguez, R., Caselli, T., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., Shahi, G.K., Mubarak, H., Nikolov, A., Babulkov, N., Kartal, Y.S., Beltr´an, J.：《CLEF-2022 CheckThat!实验室：应对新冠肺炎信息疫情和检测假新闻》。载于：欧洲信息检索会议（2022） 19. Nakov, P., Corney, D.P.A., Hasanain, M., Alam, F., Elsayed, T., Barr´on-Cede˜no, A., Papotti, P., Shaar, S., Martino, G.D.S.：辅助人工事实核查员的自动化事实核查。载于：IJCAI。第4551–4558页。ijcai.org（2021） 20. Nakov, P., Da San Martino, G., Elsayed, T., Barr´ on-Cedeno, A., M´ıguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Babulkov, N., 等：CLEF-2021 CheckThat!实验室：检测可核查性声明、已核查声明和假新闻。载于：欧洲信息检索会议。第639–649页。施普林格（2021） 21. Nakov, P., Da San Martino, G., Elsayed, T., Barr´on-Cede˜no, A., M´ıguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Mansour, W., 等：CLEF-2021 CheckThat!实验室概述：检测可核查性声明、已核查声明和假新闻。载于：欧洲语言跨语言评测论坛国际会议。第264–291页。施普林格（2021） 22. Pfeiffer, J., Kamath, A., R¨uckl´e, A., Cho, K., Gurevych, I.：AdapterFusion：用于迁移学习的非破坏性任务组合。载于：Merlo, P., Tiedemann, J., Tsarfaty, R.（编）《计算语言学协会欧洲分会第16届会议论文集：主会场》，EACL 2021，在线，2021年4月19日至23日。第487–503页。计算语言学协会（2021）。https://doi.org/10.18653/v1/2021.eacl-main.39，https://doi.org/10.18653/v1/ 2021.eacl-main.39

23. Pfeiffer, J., R¨uckl´e, A., Poth, C., Kamath, A., Vuli´c, I., Ruder, S., Cho, K., Gurevych, I.：AdapterHub：用于适配Transformer的框架。载于：《2020年自然语言处理实证方法会议：系统演示》。第46–54页。计算语言学协会，在线（2020年10月）。https://doi.org/10.18653/v1/2020.emnlp-demos.7，https://aclanthology.org/2020.emnlp-demos.7 24. Pfeiffer, J., Vuli´c, I., Gurevych, I., Ruder, S.：MAD-X：用于多任务跨语言迁移的基于适配器的框架。载于：《2020年自然语言处理实证方法会议（EMNLP）论文集》。第7654–7673页。计算语言学协会，在线（2020年11月）。https://doi.org/10.18653/v1/2020.emnlp-main.617，https://aclanthology.org/2020.emnlp-main.617 25. Pfeiffer, J., Vuli´c, I., Gurevych, I., Ruder, S.：无处不在的未知词：将多语言语言模型适配到新文字系统。载于：《2021年自然语言处理实证方法会议论文集》。第10186–10203页。计算语言学协会，在线及多米尼加共和国蓬塔卡纳（2021年11月）。https://doi.org/10.18653/v1/2021.emnlp-main.800，https://aclanthology.org/2021.emnlp-main.800 26. Reimers, N., Gurevych, I.：Sentence-BERT：使用孪生BERT网络生成句子嵌入。载于：《2019年自然语言处理实证方法会议暨第九届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集》。第3982–3992页（2019） 27. R¨uckl´e, A., Geigle, G., Glockner, M., Beck, T., Pfeiffer, J., Reimers, N., Gurevych, I.：AdapterDrop：Transformer中适配器的效率研究。载于：《2021年自然语言处理实证方法会议论文集》。第7930–7946页。计算语言学协会，在线及多米尼加共和国蓬塔卡纳（2021年11月）。https://doi.org/10.18653/v1/2021.emnlp-main.626，https://aclanthology.org/2021.emnlp-main.626 28. Schlicht, I.B., de Paula, A.F.M., Rosso, P.：UPV参加CheckThat! 2021：缓解文化差异，以识别多语言值得核查的主张。载于：CLEF（工作笔记）。CEUR研讨会论文集，第2936卷，第465–475页。CEUR-WS.org（2021） 29. Shaar, S., Hasanain, M., Hamdan, B., Ali, Z.S., Haouari, F., Nikolov, A., Kutlu, M., Kartal, Y.S., Alam, F., Martino, G.D.S., Barr´on-Cede˜no, A., M´ıguez, R., Beltr´an, J., Elsayed, T., Nakov, P.：CLEF-2021 CheckThat!实验室任务1概述：推文和政治辩论中的核查价值估计。载于：CLEF（工作笔记）。CEUR研讨会论文集，第2936卷，第369–392页。CEUR-WS.org（2021） 30. Singh, K., Lima, G., Cha, M., Cha, C., Kulshrestha, J., Ahn, Y.Y., Varol, O.：40个国家的新冠肺炎错误信息、可信度与疫苗接受度：新冠肺炎信息疫情初期阶段的启示。PLOS ONE **17**(2), e0263381（2022） 31. Stickland, A.C., Murray, I.：BERT及其伙伴：用于多任务学习高效适配的投影注意力层。载于：ICML。《机器学习研究论文集》，第97卷，第5986–5995页。PMLR（2019） 32. Sundararajan, M., Taly, A., Yan, Q.：深度网络的公理化归因。载于：ICML。《机器学习研究论文集》，第70卷，第3319–3328页。PMLR（2017） 33. ¨Ust¨un, A., Bisazza, A., Bouma, G., van Noord：UDapter：面向真正通用依存句法分析的语言适配。载于：《2020年自然语言处理实证方法会议（实证方法

语言处理）。第 2302–2315 页。计算语言学协会，在线（2020年11月 180 34 2020），https://www.aclweb.org/anthology/ .emnlp-main.。Uyangodage, L., Ranasinghe, T., Hettiarachchi, H.：多语言变换器能否对抗新冠肺炎信息疫情？载于：自然语言处理最新进展国际会议论文集（RANLP 2021）。第 1432–1437 页。INCOMA有限公司，在线举行（2021年9月），https://aclanthology.org/2021. ranlp-1 160 35。。Vasileva, S., Atanasova, P., M`arquez, L., Barr´on-Cede˜no, A., Nakov, P.：九个人才能闻出一只老鼠：用于值得核查性预测的神经多任务学习。载于：Mitkov, R., Angelova, G.（编）《自然语言处理最新进展国际会议论文集》，RANLP 2019，保加利亚瓦尔纳，2019年9月2日至4日。第 1229–1239 页。INCOMA有限公司（2019）。https://doi.org/10.26615/978-954-452 056 4-141, 10 26615 978 954 452 056 4_141 -- 36 https://doi.org/ . / - - - -。Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.：Transformers：最先进的自然语言处理技术。载于：2020年自然语言处理实证方法会议：系统演示论文集。第 38–45 页。计算语言学协会，在线（2020年10月）。https://doi.org/10.18653/v1/2020.emnlp-demos.6, https://aclanthology. org/2020 6 37 .emnlp-demos.。Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.：mT5：一种大规模多语言预训练文本到文本变换器。载于：2021年计算语言学协会北美分会人类语言技术会议论文集。第 483–498 页（2021）