定位与聚焦：增强语音语言模型中的术语翻译

Suhang Wu[1,4*] **Jialong Tang**[2] **Chengyi Yang**[1] **Pei Zhang**[2]
Baosong Yang[2] **Junhui Li**[3] **Junfeng Yao**[1,4†] **Min Zhang**[3] **Jinsong Su**[1,4†]
[1]Department of Digital Media Technology, Xiamen University [2]Tongyi Lab [3]Soochow University
[4]Key Laboratory of Digital Protection and Intelligent Processing of Intangible
Cultural Heritage of Fujian and Taiwan (Xiamen University), Ministry of Culture and Tourism, China
wusuhang@stu.xmu.edu.cn, tangjialong.tjl@alibaba-inc.com
{ yao0010,jssu } @xmu.edu.cn

Abstract

直接语音翻译（ST）如今获得了越来越多的关注，但在语句中准确翻译术语仍然是一个巨大挑战。在这方面，当前的研究主要集中于将各种翻译知识融入到 ST 模型中。然而，这些方法往往在处理来自不相关噪声的干扰时遇到困难，并且不能充分利用翻译知识。为了解决这些问题，本文提出了一种用于术语翻译的新颖的定位与聚焦方法。它首先有效地定位语句中包含术语的语音片段，以构建翻译知识，尽量减少对 ST 模型的无关信息。随后，它将翻译知识与语句和假设在音频和文本两种模式下关联，使 ST 模型在翻译过程中能够更好地聚焦于翻译知识。在各种数据集上的实验结果表明，我们的方法能够有效地定位语句中的术语，并提高术语翻译的成功率，同时保持稳健的通用翻译性能。我们的代码和数据将在 https://github.com/DeepLearnXMU/Locate_and_Focus_ST 上提供。

1 引言

直接语音翻译（ST）旨在将源语言的语音直接转换为目标语言的文本，近期的进展受到了语音大型语言模型（LLMs）(Papi et al., 2023; Gupta et al., 2024; Peng et al., 2024; Hussein et al., 2024; Sethiya and Maurya, 2025) 出现的推动。尽管已取得显著进展，但主流的直接 ST 模型在术语翻译方面仍表现不佳，如人名和药品名等，这对于有效的信息传递和专业交流至关重要 (Ailem et al., 2022; Semenov et al., 2023; Bogoychev and Chen, 2023; Conia et al., 2024; Yin et al., 2024; Liu et al., 2025) 。

为了解决这个问题，研究人员提出了各种结合外部翻译知识的方法。如图 1 所示，这些方法大致可以归类为以下两种模式：1）收集和整合 (Gaido et al., 2023; Chen et al., 2024) 。它收集语料库中的所有文本术语及其翻译作为上下文，以提供给 ST 模型。2）Retrieve-and-Demonstrate (Li et al., 2024a) 。这种范式使用检索器来获取包含与源语句相同术语的语句-翻译对，然后将这些对作为上下文学习的示例提供 (Brown et al., 2020) 。

尽管取得了一些成功，但上述范式仍存在两个缺点。一方面，它们引入了大量的无关信息。具体而言，收集与整合范式将所有语料库术语纳入上下文，通常包括许多不相关的内容，如“语音翻译”和“边缘计算”，如图 1 所示。而检索与演示范式检索的语句-翻译对包含与术语翻译无关的句子部分，例如“在文本分析中起着至关重要的作用”。另一方面，由于模态或说话者的不同，ST 模型难以充分利用翻译知识。注意，收集与整合范式从文本模态引入翻译知识，这与源语句的音频模态有显著差异。此外，对于检索与演示，检索到的语句和源语句通常来自不同的说话者，具有不同的口音和情感。因此，有效地结合外部翻译知识以改善直接 ST 中的术语翻译面临重大挑战。

为了解决这些挑战，我们提出了一种新颖的基于大型语言模型的术语翻译方法，称为“定位与聚焦”方法，该方法包括两个关键步骤。术语片段定位步骤采用基于滑动窗口的检索方法，从翻译知识库中高效识别术语并定位它们在话语中的相应语音片段。这个过程使得语音大型语言模型能专注于包含术语的部分，从而减少来自无关部分的干扰。随后的术语聚焦翻译步骤结合音频和文本形式的内容，将翻译知识与话语及假设关联在一起，促进语音大型语言模型聚焦于翻译知识。具体而言，我们用我们定位的话语中的片段替换从检索到的翻译知识中的语音片段。这个过程确保了话语和翻译知识共享共同的语音片段，从而引导语音大型语言模型聚焦于翻译知识。此外，我们鼓励语音大型语言模型在翻译术语之前预测一个特殊标记，作为专注于翻译知识的自我提醒。

由于缺乏针对语音任务的术语翻译数据集，我们从现有的 ST 数据集 CoVoST2 (Wang et al., 2020) 、MuST-C (Cattoni et al., 2021) 和 MSLT (Federmann and Lewis, 2016, 2017) 中收集了一个定制的数据集。该数据集包含英语到中文和英语到德语的翻译方向。结果表明，我们的方

[*] Work done during internship at Tongyi Lab.
[†] Corresponding authors

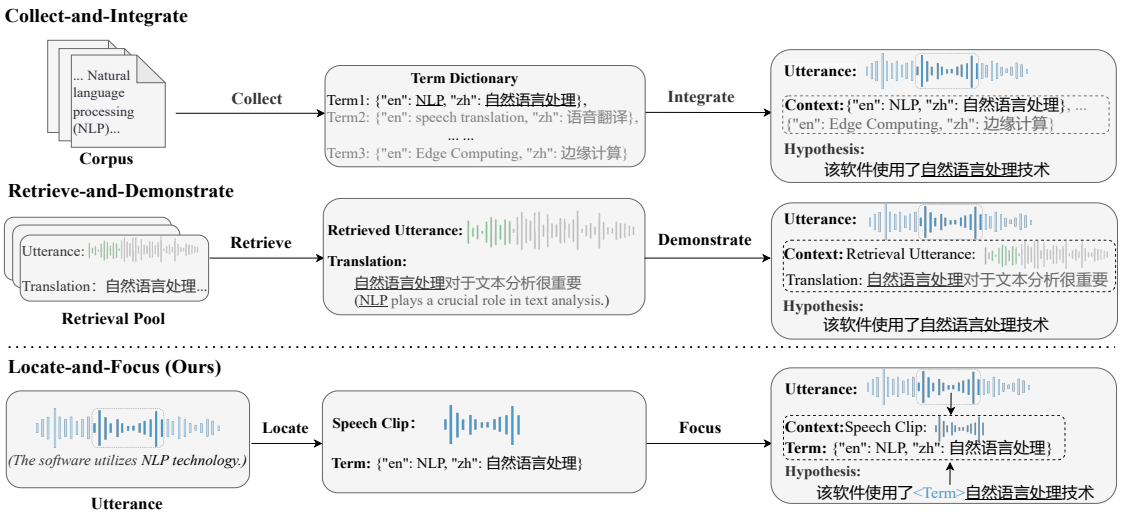

Figure 1: Locate-and-Focus 与现有范式的差异。我们用 灰色 来表示与术语翻译无关的信息。与术语翻译相关的句子和假设中的部分用 蓝色 突出显示。

法不仅能有效定位语句中的术语，还提高了术语翻译的成功率，并保持了稳健的通用翻译性能。

总之，我们在这项工作中的贡献主要有三个方面：

- 我们提出了定位与聚焦的方法用于术语翻译，该方法不仅通过精确定位包含术语的语音片段来减少无关信息的引入，还能有效地引导语音大语言模型利用翻译知识。
- 我们构建了一个高质量的术语翻译数据集，以评估英译中文和英译德文方向的术语翻译性能。
- 实验结果表明，我们的方法能够准确定位话语中的术语，从而显著提高术语翻译的质量，同时保持一般翻译的质量。

2 相关工作

与我们的研究相关的工作涵盖以下两个方面：

基于文本的术语翻译。在这种情况下，主要方法可以大致分为三种类型。第一类侧重于优化解码过程 (Hokamp and Liu, 2017; Post and Vilar, 2018; Hasler et al., 2018)，通过扩展搜索空间或有限状态自动机来提高一致性，尽管这通常会导致翻译质量欠佳。第二种方法涉及对网络结构的修改 (Chen et al., 2021; Wang et al., 2022)，但网络结构的重大变化可能限制其可用性。最后，数据增强方法包括占位符法和混码法。占位符方法在翻译前后，通过替换源文本和目标文本中的术语为有序标签来进行翻译。混码方法在输入模型前，直接用翻译替换源文本中的术语 (Dinu et al., 2019; Bergmanis and Pinnis, 2021)。此外，Zhang et al. (2023) 结合占位符和混码法以获得更好的结果。

请注意，Placeholder 和 Code-switch 不能直接应用于直接 ST，因为将话语的部分替换为文本标签或翻译可能导致跨模态不一致。此外，与这些在源文本中用标签或翻译替换术语的方法不同，我们在模型的假设中加入了特殊标记来改进术语翻译。

语音任务中的术语翻译。与基于文本的术语翻译相比，由于整合了更多的模态，语音任务中的术语处理更加复杂。在端到端自动语音识别（ASR）中，Li et al. (2024b) 引入了 CB-Whisper，通过开放词汇关键字识别来识别术语。Hu et al. (2024) 提出了 VHASR，这是一个多模态语音识别系统。在语音翻译中，主流的方法可以大致分为两种范式：收集与整合 (Gaido et al., 2023; Chen et al., 2024) 和检索与展示 (Li et al., 2024a)。作为前者的代表，Gaido et al. (2023) 提出了一种检测器，用于识别文本术语是否出现在语音中。类似地，Chen et al. (2024) 将高频术语的文本翻译以细粒度的水平整合到提示中，以帮助模型进行术语翻译。然而，这些方法并未引入多模态的翻译知识。作为后者范式的代表，Li et al. (2024a) 检索语音-翻译对，并通过上下文学习增强术语翻译能力。

与上述研究形成对比，我们的工作有两个关键优势。首先，它能够有效地识别包含术语的语句中的语音片段，从而减少噪声干扰。其次，我们的方法鼓励模型关注来自双模态的翻译知识。据我们所知，我们是第一个端到端的术语翻译方法，可以检索并充分利用多模态以及细粒度的多模态知识用于语音大模型。

3 方法

在本节中，我们详细描述了我们提出的方法。如图 2 所示，我们的方法主要包括两个步骤：

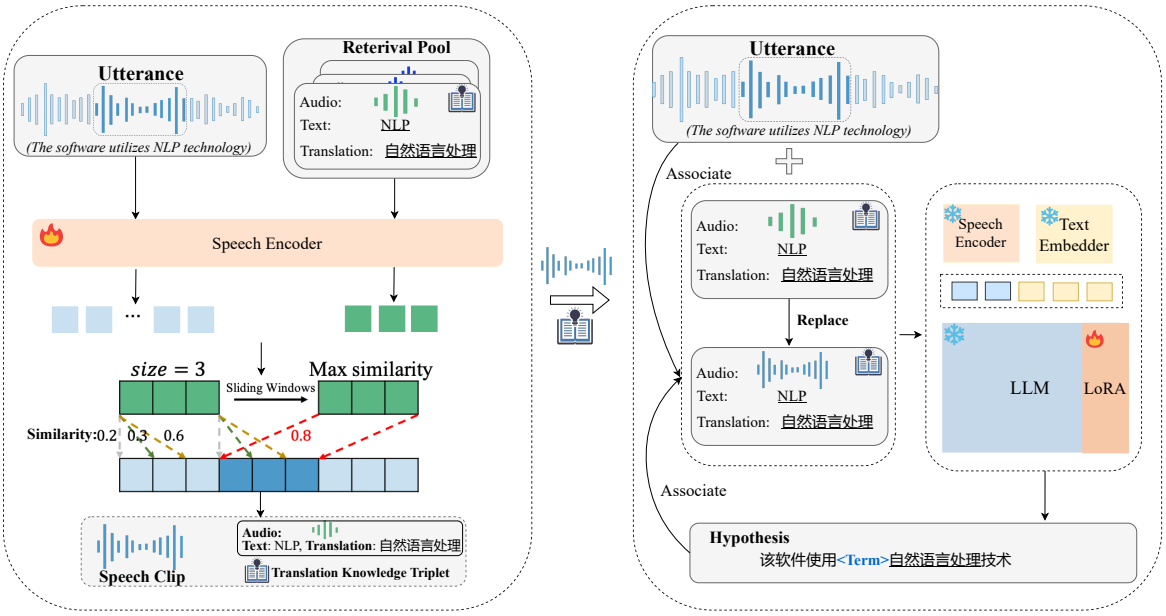

Figure 2: 说明“定位与聚焦”方法，包含语音术语剪辑定位和术语聚焦翻译步骤。对于给定的语句，第一步有效识别并定位包含该术语的语句内的语音剪辑。随后，第二步通过音频替换，利用共同的语音剪辑将语句与翻译知识关联起来。这也鼓励模型在翻译术语之前预测 <Term> 标签，这有助于其专注于翻译知识。

术语片段定位和术语聚焦翻译。我们将在节 3.1 和 3.2 中详细说明这些步骤，接着在节 ?? 中讨论训练过程。

3.1 术语片段定位

在此步骤中，我们旨在准确检索话语中的术语，并定位其在话语中对应的语音片段。通过在话语中定位这些术语相关的片段，语音语言模型可以更容易地聚焦于这些关键部分，从而有效地最小化无关信息。

设 $\mathcal{P}$ 为外部翻译知识库，作为检索池，其中每个元素是一个术语翻译知识三元组 $K = (x, c, y)$，其中 x 表示术语的文字记录，c 表示其对应的语音片段，y 代表其翻译。鉴于在同一音频模式下的检索往往优于跨模态检索 (Li et al., 2024a)，我们使用 c 计算与需要翻译的测试集中的语句 u 的相似性。

滑动检索。由于只有发音的某些部分包含术语，因此直接计算 c 和 u 之间的相似性以检索术语是具有挑战性的。为了解决这个问题，我们提出了一种基于滑动窗口的相似性匹配方法，称为滑动检索。该方法不仅能够更好地计算相似性，还能定位源发音中最可能出现术语的语音片段。

具体而言，我们使用语音编码器 SE 对 c 和 u 进行编码：$\boldsymbol{z}^{\mathrm{c}} = \mathrm{SE}(c), \boldsymbol{z}^{\mathrm{u}} = \mathrm{SE}(u)$，其中 $\boldsymbol{z}^{\mathrm{c}} \in \mathbb{R}^{|c| \times d}$ 和 $\boldsymbol{z}^{\mathrm{u}} \in \mathbb{R}^{|u| \times d}$ 表示 d 维嵌入，长度分别为 $|c|$ 和 $|u|$。随后，我们使用一个大小为 $|c|$、步长为 1 的滑动窗口将 u 分割为语音子序列 $[\boldsymbol{z}^u_1, \ldots, \boldsymbol{z}^u_{|u|-|c|+1}]$ [1]。对于每个子序列，我们在 $\boldsymbol{z}^u_i$ 和 $\boldsymbol{z}^{\mathrm{c}}$ 上执行最大池化，然后计算它们的余弦相似度。获得的最大相似度将表示 u 和 c 之间的相似性，从而指示术语 c 出现在 u 中的可能性。此过程形式上定义为：

$$\mathrm{sim}(u, c) = \max_i \{\mathrm{Cosine}(\mathrm{MaxPool}(\boldsymbol{z}^c), \mathrm{MaxPool}(\boldsymbol{z}^u_i))\} \quad (1)$$

请注意，我们针对知识库中的所有翻译知识三元组计算相似度得分，然后选择得分最高的前 k 个三元组，认为这些三元组的术语最有可能出现在话语中。同时，我们识别出展现最大相似度的语音子序列，并将其对应的语音片段记为 s，这可能包含该术语。

3.2 术语聚焦翻译

在此步骤中，我们制定了两种策略，将翻译知识与语音和文本模态中的话语和假设关联起来，使语音大语言模型更好地关注翻译知识。

如图 2 所示，我们首先将检索到的翻译知识三元组 $K = (x, c, y)$ 中的语音剪辑 c 替换为定位的语音剪辑 s，从而形成新的翻译知识三元组 $K' = (x, s, y)$。这种替换创建了一个锚点，

[1] 请注意，不同子序列的相似性计算可以并行化，因此只会导致延迟的轻微增加。有关更多详细信息，请参阅第 ?? 节。

使话语和翻译知识能够共享相同的声学特征。当语音 LLM 在处理话语时遇到这个锚点时，它可以更有效地关注相关的翻译知识。然后，我们提供这个新的三元组作为附加上下文，与话语 u 一起构建输入到语音 LLM 的指令。

为了进一步增强术语翻译，我们引入了特殊标记作为提示，建立模型假设与翻译知识之间的连接。具体来说，我们通过在每个术语的翻译前添加一个特殊标记 `<Term>` 来修改训练数据的参考。如图 2 所示，由于 “NLP ” 是一个术语，参考 “软件利用 NLP 技术” 将被修改为 “软件集成 `<Term>` NLP 技术”。随后，我们使用这些修改后的训练数据以自回归方式训练语音大语言模型。通过这种方式，当语音大语言模型在推理过程中预测 `<Term>` 时，它提示语音大语言模型关注外部翻译知识三元组 K'，以准确翻译术语。

请注意，如果没有事先训练，我们的术语片段定位步骤可能会产生不令人满意的语音片段，进而削弱以术语为中心的翻译步骤的效果。因此，我们依次训练这两个步骤。

术语剪辑定位步骤训练的目标是确保 SE 与我们的滑动检索方法对齐。为了实现这一点，我们使用对比学习进行 SE 训练。形式上，我们的训练目标 $\mathcal{L}_{\text{SE}}$ 是在最大化与正例相似度的同时最小化与负例相似度：

$$\mathcal{L}_{\text{SE}} = -\log \frac{e^{\text{sim}(u,c^+)}}{e^{\text{sim}(u,c^+)} + \sum_{i=1}^{n} e^{\text{sim}(u,c_i^-)}}, \quad (2)$$

其中 c^+ 表示术语在 u 中出现的语音剪辑，视为正例，而 c_i^- 表示第 i 个随机采样的术语语音剪辑，被视为负例。

随后，我们训练模型以专注于术语的翻译，确保在翻译过程中能够有效利用提供的翻译知识。根据之前的研究 (Rajaa and Tushar, 2024; Chen et al., 2024)，我们应用 LoRA (Hu et al., 2022) 进行微调。正式而言，我们使用标准的下一个令牌预测损失来训练语音 LLM，如下所示：

$$\mathcal{L}_{\text{LLM}} = -\frac{1}{N}\sum_{i=1}^{N} \log P(w_i|K', u, w_{<i}), \quad (3)$$

其中 N 是翻译中的令牌总数，w_i 是目标令牌，$P(w_i|K', u, w_{<i})$ 是 w_i 的预测概率。

4 数据收集

鉴于当前的语音翻译数据集通常缺乏标注的术语翻译知识，我们创建了一个专门用于术语翻译的数据集。具体来说，我们从现有的 ST 数据集中收集数据，包括 CoVoST2 (Wang et al., 2020)、MuST-C (Cattoni et al., 2021) 和 MSLT (Federmann and Lewis, 2016, 2017)。生成的数据显示了英语到中文和英语到德语翻译方向的标注术语翻译。

	EN → ZH		EN → DE	
Split	# utt.	# term.	# utt.	# term.
CoVoST2-train	10000	14191	10000	14664
CoVoST2-test	671	1227	656	1270
MuST-C-test	220	335	220	355
MSLT-test	213	294	164	280

Table 1: 我们收集的数据集的统计信息。# utt. 表示话语的数量，# term. 表示术语的数量。

为实现这一目标，我们利用 Qwen2.5-72B-Instruct (Yang et al., 2024) 从现有 ST 数据集的转录和翻译中提取平行术语对，然后人工检查提取的术语对以确保质量。为了更好地支持 ST，我们使用文本到语音（TTS）模型 CosyVoice2 (Du et al., 2024) 生成术语的对应语音片段。为了保证生成语音的质量，我们采用 ASR 模型 SenseVoice (An et al., 2024) 转录合成的语音片段，并将这些转录与源术语进行比较。需要注意的是，我们只保留在修改距离为 3 或更小的术语的语音片段。在此初步过滤后，我们还进行人工审核以进一步确保片段的质量。关于我们的收集过程的更多细节请参见附录 C。

我们收集的数据详情如表格 1 所示。对于 CoVoST2 (Wang et al., 2020)，我们收集了训练集和测试集的数据，而对于 MuST-C (Cattoni et al., 2021) 和 MSLT (Federmann and Lewis, 2016, 2017)，我们只收集了测试集的数据。注意，我们仅保留包含术语的翻译样本。在后续过程中，我们仅使用 CoVoST2 的训练集进行模型训练，而 MuST-C 和 MSLT 用作领域外的测试集。

5 实验

基础模型 在我们的实验中，我们使用 Whisper-medium (Radford et al., 2023) 作为语音编码器，Qwen2-Audio-Instruct (Chu et al., 2024) 作为语音 LLM。在训练语音编码器时，我们每个例子使用 4 个负样本，并进行 3 个周期的训练。为了确保语音 LLM 的翻译质量，我们将原始的 CoVoST2 训练集与术语翻译数据结合进行训练。对于需要外部翻译知识的方法，我们使用在第 4 节中构建的翻译知识库。有关进一步的实现细节，请参阅附录 A。

我们使用代表性的方法作为我们的基线。

在我们的实验中，我们使用两种不同的设置来为语音 LLM 提供翻译知识：为了研究不同因素对我们方法的影响，我们考虑了以下变体

	EN → ZH						EN → DE					
	CoVoST2		MuST-C		MSLT		CoVoST2		MuST-C		MSLT	
	TSR	BLEU	TSR	BLEU	TSR	BLEU	TSR	BLEU	TSR	BLEU	TSR	BLEU
Base Model	24.12	35.82	27.61	25.73	39.80	31.30	40.38	26.35	53.24	14.33	49.72	18.10
Translation Training	27.30	40.66	32.68	27.02	45.24	31.48	45.52	29.36	48.31	20.45	60.79	19.11
Oracle Knowledge Setting												
SALM	76.53	55.97	69.01	32.10	68.03	31.81	85.91	43.64	76.56	21.15	72.30	16.16
Retrieval-and-Demonstration	60.88	50.22	58.87	30.18	70.06	31.34	57.95	36.09	57.06	19.46	53.95	15.18
Locate-and-Focus	90.13	58.49	94.09	34.52	91.84	33.76	96.35	45.60	87.85	22.06	86.33	17.30
w/o Audio Replacement	89.67	58.37	90.07	33.43	91.50	33.25	93.83	45.20	87.00	21.20	85.37	17.07
w/o Tag Cue	89.00	58.25	88.17	31.09	90.14	32.05	90.74	44.97	85.94	21.36	83.74	17.24
w/o Replacement and Cue	88.59	58.32	86.14	31.44	89.14	30.05	91.00	43.29	81.92	21.88	76.61	16.67
End-to-End Setting												
SALM	28.20	39.82	37.18	27.16	46.40	30.27	41.17	31.16	48.31	15.02	34.17	8.35
Retrieval-and-Demonstration	32.93	41.02	38.31	26.87	56.80	30.54	45.37	32.40	51.97	16.05	52.88	15.48
Locate-and-Focus	65.53	49.30	75.78	31.35	75.51	30.58	77.12	39.66	77.40	21.05	72.66	16.98
w/o Sliding Retrieval	58.02	44.82	72.91	30.72	72.39	28.10	71.49	38.98	75.14	20.92	70.02	16.35
w/o Audio Replacement	63.49	49.52	74.62	31.12	73.91	32.24	75.25	39.21	77.11	20.60	71.94	17.05
w/o Tag Cue	63.73	48.78	72.91	30.74	72.95	30.08	73.28	39.36	74.62	20.83	69.98	16.36
w/o Replacement and Cue	62.95	48.73	71.26	30.79	71.76	30.42	70.54	37.93	72.98	20.29	69.86	16.37

Table 2: 不同方法在语音术语翻译中的性能比较，包括我们方法的变体。我们使用粗体文本来表示每个指标的最佳性能。

进行消融研究。

- 没有滑动检索。在这个变体中，检索器使用最大池化来计算话语与语音片段之间的相似性，而不是采用我们提出的滑动检索方法。
- 无音频替换。此变体将检索到的知识三元组直接提供给语音 LLM，而不使用从话语中找到的片段替换 TTS 生成的音频。
- 无标签提示。在此变体中，我们在训练过程中排除了使用特殊标签，这意味着模型不能使用特殊标签作为提示来预测何时输出术语翻译。
- 无替换和提示。在训练和推理过程中，此变体省略了音频替换和标签提示。

为了评估检索性能，我们使用 Hits@N 来评估正确的项目是否包含在前 n 个检索结果中，其中 n 被设置为 1、5 或 10。为了评估术语翻译的质量，遵循以前的研究，我们采用 BLEU 和术语成功率（TSR）。术语成功率量化了在一次发言中准确翻译的术语比例。

如表 2 所示，我们报告了不同方法及其变体的性能，从中我们可以得出以下结论：

首先，提供外部翻译知识可以显著提高术语翻译的成功率。在 Oracle 知识设置中，所有结合外部知识的方法都优于基础模型和通过翻译训练增强的模型。这表明仅仅增强翻译能力对于有效的术语翻译而言是不够理想的。我们将这归因于术语的长尾分布，使其稀疏且在训练中难以获取。因此，整合外部知识成为一种有效的方法。

其次，外部知识的质量对于准确的术语翻译至关重要。在端到端设置中，例如，由于高频术语往往无法与当前话语中的术语对齐，SALM 的性能下降。基于检索的方法也面临类似问题。由于检索器性能的不完善，“检索与示范”方法的性能也有所下降，在 CoVoST2 英语到中文数据集中的得分从 60.88 降至 32.93。因此，我们认为进一步提高检索性能对于有效的术语翻译是必不可少的。

第三，Locate-and-Focus 超越了现有的方法。在 CoVoST2 英译中数据集的端到端设置中，它实现了 65.53 的 TSR，显著优于 SALM 的 28.20 和 Retrieve-and-Demonstrate 的 32.93。此外，与通过翻译训练增强的模型相比，它通常取得更高的 BLEU 分数。这种优势归因于关键术语的准确翻译，这对于整体翻译质量至关重要。

最后，我们的消融研究强调了我们方法中每个组件的重要性。我们发现，无论是移除音频替换还是标签提示，都会导致性能显著下降。例如，在 MuST-C 英译中数据集的 Oracle 知识设置中，移除音频替换会使 TSR 从 94.09 下降到 90.07，而移除标签提示则会降至 88.17，同时移除两者则进一步降到 86.14。同样，移除滑动检索也会导致性能下降，我们将证明这与检索性能有关。

5.1 检索性能

鉴于缺乏有效的跨粒度语音检索方法，我们将滑动检索方法与基本池化方法进行比较，如表

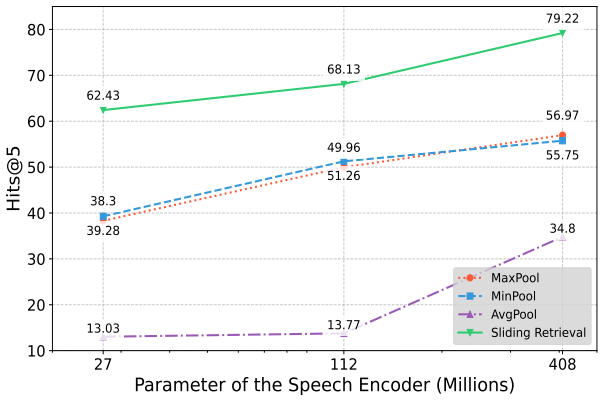

Figure 3: 使用不同大小的语音编码器的方法的 Hits@5 分数比较。

	CoVoST2		MuST-C	
	TSR	BELU	TSR	BELU
EN → ZH				
Top-1	51.67	47.63	58.02	29.90
Top-5	65.53	49.30	75.78	31.35
Top-10	55.01	47.57	60.56	29.45
EN → DE				
Top-1	63.89	37.62	69.49	20.90
Top-5	77.12	39.66	77.40	21.05
Top-10	69.52	38.22	69.77	19.74

Table 3: 我们方法在不同检索设置下的表现，其中 Top- N 表示包含得分最高的 N 个翻译知识三元组。

格 **??** 所示，所有方法都使用相同的数据集来训练语音编码器。实验结果表明，我们的方法在 Hits@1 上实现了大约 60 % 的准确率，在 Hits@10 上约为 85 %。与基于池化的方法相比，滑动检索在所有检索指标上均表现出显著的改进。

为了进一步验证我们方法在不同模型规模上的有效性，我们在 CoVoST2 的英译中子集上进行实验，使用 Whisper-base（约 27M 参数）、Whisper-small（112M 参数）、Whisper-medium（408M 参数）作为语音编码器 [2] 。如图 3 所示，我们的方法在所有模型规模上都持续取得显著更好的性能。

定位剪辑的质量 请注意，滑动检索不仅提高了检索性能，还有效地定位了相应的语音片段。为了验证其有效性，我们对定位的语音片段进行了全面评估。

我们使用英译中数据集，采用 Whisper-medium 定位包含真实术语的语音片段。接着请人类注释员验证这些片段是否准确捕捉目标术语，从而使我们能够计算成功率。结果显示，该方法表现出强大的性能，在 CoVoST-2、MSLT 和 MuST-C 数据集上的术语识别成功率

[2]我们只使用 Whisper 的编码器，并报告编码器的参数数量。

	EN → ZH	EN → DE
Base Model	38.22	23.61
Translation Training	43.64	29.79
SALM	43.39	29.47
Retrieval-and-Demonstration	43.08	29.43
Locate-and-Focus	43.48	29.62

Table 4: 在标准 CoVoST2 测试集上的方法性能。

分别为 88.10%，92.56%，和 93.98%，证实了该方法在精确术语定位上的有效性。

如表 3 所示，我们研究了为语音大型语言模型提供不同数量检索到的译文知识对术语翻译性能的影响。结果表明，使用 top-1 检索通常表现最差，而 top-10 也不如 top-5 有效。这是因为 top-1 检索的准确性较差，Hits@1 在英译中文 CoVoST2 数据集上仅达到 61.04，明显低于 76.22 的 Hits@5 和 85.00 的 Hits@10，如表 **??** 所示。尽管 top-10 检索达到最高的召回率，它也引入了更多无关的译文知识噪声。相反，top-5 检索通过提供噪声最少的译文知识找到了一个平衡，从而实现了更好的性能。

5.2 一般翻译表现

在本节中，我们研究了增强术语翻译能力对普通语音翻译性能的潜在影响。我们在标准 CoVoST2 测试集上进行了全面评估，BLEU 分数如表 4 所示。实验结果表明，我们的方法不仅在特定术语翻译任务中表现优异，同时也保持了强健的普通语音翻译性能。例如，我们的 Locate-and-Focus 方法在英译汉测试集上取得了 43.48 的 BLEU 分数，接近翻译训练方法 (43.64) 的表现，同时超过了其他基于检索的方法，如 SALM (43.39) 和检索与示范方法 (43.08)。

考虑到语音翻译系统的关键实时约束，我们对我们的滑动检索方法的计算效率进行了全面评估。使用单个 NVIDIA A100 80GB GPU，我们预先计算和存储由语音编码器生成的语音表示，然后系统地测量使用单个话语从检索池中检索结果所需的时间。我们的分析包含 5000 个样本，平均处理时间在表 **??** 中报告。

结果表明，与 MaxPool 基准相比，我们的滑动检索方法仅引入了可忽略的计算开销。具体来说，在使用 Whisper-medium 时，MaxPool 方法每个查询平均为 0.152 毫秒，而我们的滑动检索方法仅需 0.217 毫秒，显示出微小的差异。请注意，相对于 Qwen2-Audio-Instruct 在翻译过程中所需的 621.951 毫秒，检索延迟实际上是微不足道的。此外，我们的分析显示，扩展语音编码器参数对系统延迟影响最小，其中滑动检索在 Whisper-base 和 Whisper-medium 中平均分别为 0.195 毫秒和 0.217 毫秒。

在本文中，我们探讨了在语音翻译中准确翻译术语的关键挑战。我们提出了定位聚焦法，该方法有效地减少噪声并充分利用翻译知识。该方法包括两个核心步骤：术语片段定位和术语聚焦翻译。在第一步中，我们识别并定位包含术语的语音片段。随后，在术语聚焦翻译步骤中，我们将翻译知识与来自音频和文本模式的语句和假设相结合，引导模型关注翻译知识。实验结果表明，我们的方法显著提高了各个数据集上的术语翻译成功率，并保持了强大的通用翻译性能。在未来的工作中，我们将拓展术语在其他语音任务中的应用，并研究在传统神经机器翻译研究中广泛研究的强健机器翻译。

6

局限性 在本节中，我们讨论了我们工作的主要局限性以及未来研究如何能够解决这些问题。

依赖预定义术语 我们的方法依赖于预定义的术语集合，这可能最初不包括所有潜在术语。这一限制在某种程度上限制了方法的灵活性。未来，探索自动构建全面且高质量的术语知识库的方法将是必不可少的。

语言覆盖率 我们的方法仅在英语到中文及英语到德语的翻译中进行了测试。未来，我们计划在更多语言中进行实验，以进一步展示方法的有效性。

在其他语音任务中的探索 我们的方法目前专注于翻译任务，但在未来，它可以应用于其他语音任务，如自动语音识别。

7

致谢 本项目得到了中国国家自然科学基金（编号 62036004，编号 62276219），中国福建省自然科学基金（编号 2024J011001），厦门市公共技术服务平台项目（编号 3502Z20231043）以及阿里巴巴研究实习生计划的支持。我们也诚挚感谢审稿人所提出的深思熟虑且富有洞察力的意见。

References

Melissa Ailem, Jingshu Liu, and Raheel Qader. 2022. Encouraging neural machine translation to satisfy terminology constraints. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale, TALN-RECITAL 2022, Avignon, France, June 27 - July 1, 2022*, page 446. ATALA.

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Haoneng Luo, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, Changhe Song, Jiaqi Shi, Xian Shi, Hao Wang, Wen Wang, Yuxuan Wang, Zhangyu Xiao, Zhijie Yan, Yexin Yang, Bin Zhang, Qinglin Zhang, Shiliang Zhang, Nan Zhao, and Siqi Zheng. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *CoRR*, abs/2407.04051.

Toms Bergmanis and Marcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3105–3111. Association for Computational Linguistics.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 890–896. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Comput. Speech Lang.*, 66:101155.

Guanhua Chen, Yun Chen, and Victor O. K. Li. 2021. Lexically constrained neural machine translation with explicit alignment guidance. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12630–12638. AAAI Press.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C. Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg.

2024. SALM: speech-augmented language model with in-context learning for speech recognition and translation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 13521–13525. IEEE.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *CoRR*, abs/2407.10759.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 16343–16360. Association for Computational Linguistics.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3063–3068. Association for Computational Linguistics.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *CoRR*, abs/2412.10117.

Christian Federmann and William D. Lewis. 2016. Microsoft speech language translation (MSLT) corpus: The IWSLT 2016 release for english, french and german. In *Proceedings of the 13th International Conference on Spoken Language Translation, IWSLT 2016, Seattle, WA, USA, December 8-9, 2016*. International Workshop on Spoken Language Translation.

Christian Federmann and William D. Lewis. 2017. The microsoft speech language translation (MSLT) corpus for chinese and japanese: Conversational test data for machine translation and speech recognition. In *Proceedings of Machine Translation Summit XVI, Volume 1: Research Track, MTSummit 2017, September 18-22, 2017, Nagoya, Aichi, Japan*, pages 72–85.

Marco Gaido, Yun Tang, Ilia Kulikov, Rongqing Huang, Hongyu Gong, and Hirofumi Inaguma. 2023. Named entity detection and injection for direct speech translation. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Mahendra Gupta, Maitreyee Dutta, and Chandresh Kumar Maurya. 2024. Direct speech-to-speech neural machine translation: A survey. *CoRR*, abs/2411.14453.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 506–512. Association for Computational Linguistics.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1535–1546. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jiliang Hu, Zuchao Li, Ping Wang, Haojun Ai, Lefei Zhang, and Hai Zhao. 2024. VHASR: A multimodal speech recognition system with vision hotwords. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 14791–14804. Association for Computational Linguistics.

Amir Hussein, Brian Yan, Antonios Anastasopoulos, Shinji Watanabe, and Sanjeev Khudanpur. 2024. Enhancing end-to-end conversational speech translation through target language context utilization. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 11971–11975. IEEE.

Siqi Li, Danni Liu, and Jan Niehues. 2024a. Optimizing rare word accuracy in direct speech translation with a retrieval-and-demonstration approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 12703–12719. Association for Computational Linguistics.

Yuang Li, Yinglu Li, Min Zhang, Chang Su, Jiawei Yu, Mengyao Piao, Xiaosong Qiao, Miaomiao Ma, Yanqing Zhao, and Hao Yang. 2024b. Cb-whisper: Contextual biasing whisper using open-vocabulary keyword-spotting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation,*

LREC/COLING 2024, 20-25 May, 2024, Torino, Italy, pages 2941–2946. ELRA and ICCL.

Jiarui Liu, Iman Ouzzani, Wenkai Li, Lechen Zhang, Tianyue Ou, Houda Bouamor, Zhijing Jin, and Mona Diab. 2025. Towards global ai inclusivity: A large-scale multilingual terminology dataset (gist).

Sara Papi, Marco Turchi, and Matteo Negri. 2023. Alignatt: Using attention-based audio-translation alignments as a guide for simultaneous speech translation. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 3974–3978. ISCA.

Jing Peng, Yucheng Wang, Yu Xi, Xu Li, Xizhuo Zhang, and Kai Yu. 2024. A survey on speech large language models.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1314–1324. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Shangeth Rajaa and Abhinav Tushar. 2024. Speech-LLM: Multi-Modal LLM for Speech Understanding.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the wmt 2023 shared task on machine translation with terminologies. In *Proceedings of the Eight Conference on Machine Translation (WMT)*. Association for Computational Linguistics.

Nivedita Sethiya and Chandresh Kumar Maurya. 2025. End-to-end speech-to-text translation: A survey. *Comput. Speech Lang.*, 90:101751.

Changhan Wang, Anne Wu, and Juan Miguel Pino. 2020. Covost 2: A massively multilingual speech-to-text translation corpus. *CoRR*, abs/2007.10310.

Shuo Wang, Zhixing Tan, and Yang Liu. 2022. Integrating vectorized lexical constraints for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7063–7073. Association for Computational Linguistics.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *CoRR*, abs/2409.12122.

Yongjing Yin, Jiali Zeng, Yafu Li, Fandong Meng, and Yue Zhang. 2024. Lexmatcher: Dictionary-centric data curation for llm-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 14767–14779. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 5364–5376. Association for Computational Linguistics.

Huaao Zhang, Qiang Wang, Bo Qin, Zelin Shi, Haibo Wang, and Ming Chen. 2023. Understanding and improving the robustness of terminology constraints in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6029–6042. Association for Computational Linguistics.

Model	Method	CoVoST2			MuST-C			MLST		
		Hits@1	Hits@5	Hits@10	Hits@1	Hits@5	Hits@10	Hits@1	Hits@5	Hits@10
Whisper-base	MaxPool	23.80	38.30	44.25	29.30	47.89	54.65	36.73	56.46	65.99
	MinPool	24.69	39.28	46.54	32.39	52.39	57.75	39.8	58.16	67.01
	AvgPool	5.94	13.03	18.01	9.86	20.28	25.63	13.61	24.83	30.61
	Sliding Retrieval	40.59	62.43	72.13	46.76	69.86	76.33	47.96	75.17	84.36
Whisper-small	MaxPool	36.59	49.96	56.56	49.01	60.28	68.45	53.4	73.13	81.29
	MinPool	37.73	51.26	58.03	48.45	62.53	69.85	56.12	73.81	81.63
	AvgPool	6.85	13.77	17.03	16.62	27.89	34.65	14.29	30.27	37.76
	Sliding Retrieval	45.31	68.13	78.24	52.96	76.9	85.92	55.68	91.15	94.90
Whisper-medium	MaxPool	45.07	56.97	62.18	55.12	68.17	74.37	69.05	83.33	87.76
	MinPool	45.80	55.75	61.53	53.80	63.66	70.70	61.56	83.00	87.41
	AvgPool	22.66	34.80	40.91	38.87	54.37	58.87	46.93	63.27	70.75
	Sliding Retrieval	61.04	79.22	85.00	64.23	82.54	89.58	71.09	92.86	97.62

Table 5: 检索器在英译中数据集上的性能。

Model	Method	CoVoST2			MuST-C			MLST		
		Hits@1	Hits@5	Hits@10	Hits@1	Hits@5	Hits@10	Hits@1	Hits@5	Hits@10
Whisper-base	MaxPool	22.97	35.47	42.52	25.42	43.79	52.26	28.78	45.69	55.03
	MinPool	23.52	39.19	46.08	32.20	48.31	56.78	28.78	45.32	52.88
	AvgPool	5.62	11.96	14.81	7.34	18.36	24.29	8.99	16.91	23.02
	Sliding Retrieval	41.4	61.28	71.18	48.02	64.69	76.55	49.64	73.74	84.17
Whisper-small	MaxPool	35.78	48.14	53.99	44.35	57.62	63.27	48.56	61.15	66.18
	MinPool	37.45	50.12	55.67	47.17	59.60	66.10	48.20	62.23	69.06
	AvgPool	5.14	11.95	14.80	10.45	18.64	23.44	11.51	19.78	26.25
	Sliding Retrieval	44.34	63.34	74.82	52.14	79.94	88.42	53.67	87.05	92.08
Whisper-medium	MaxPool	46.08	56.85	62.00	56.21	68.64	72.31	57.55	72.30	76.62
	MinPool	44.41	55.03	60.81	54.52	66.67	71.75	53.96	67.99	74.46
	AvgPool	20.66	34.92	39.67	35.31	49.15	55.08	37.77	51.08	58.63
	Sliding Retrieval	58.19	76.32	84.40	67.51	87.57	93.79	72.66	89.21	93.88

Table 6: 检索器在英语到德语数据集上的表现。

A 实现细节

在我们的实验中，我们采用 Whisper-medium 作为主要的检索器。我们为每个例子加入 4 个负样本，并进行 3 轮的训练，学习率设为 1×10^{-5}，批量大小为 16。该过程可以在单个 NVIDIA A100 80G GPU 上执行，大约需要 6 小时完成。

在提取语音片段时，我们关注的是具有最高相似性的隐藏状态。在 Whisper 中，每个隐藏状态大约代表 0.02 秒，这使我们能够精确地分割语音的相关部分。

为了微调语音 LLM，我们采用 SWIFT 框架[3]，使用 LoRA，等级为 16，alpha 为 32，丢弃概率为 0.05。批量大小设置为 96，学习率配置为 $1e-4$。我们针对 `q_proj`、`k_proj` 和 `v_proj` 模块。此训练过程在八个 NVIDIA A100 80G GPU 上执行，大约需要 16 小时完成。

B 补充实验结果

在表格 5 和 6 中，我们详细展示了在采用各种检索方法时，Whisper-base、Whisper-small 和 Whisper-medium 模型的性能。从结果中可以明显看出，滑动检索方法在所有模型和数据集上始终表现出色。例如，在 CoVoST2 数据集中，应用于 Whisper-base 模型的滑动检索方法达到了 40.59 的 Hits@1 值，超过了 MaxPool 方法的 23.80。这一显著提升强调了滑动检索方法的优越性。在 MuST-C 和 MLST 数据集中也观察到了类似的趋势。这些发现说明，滑动检索不仅能巧妙地适应不同规模的模型，还能在多个领域的数据集中保持稳健的优化。

B.1 提供的翻译知识数量

表 8 展示了我们的方法在不同翻译知识检索配置下的性能。结果表明选择前 5 条翻译知识条目通常能够获得最佳性能。这突出了在检索准确性与减少无关信息之间平衡的重要性。

例如，在 CoVoST2 数据集上的英译中任务中，提供前 5 个知识条目使得 TSR 达到 65.53，BELU 达到 49.30，优于前 1 和前 10 的设置。这表明包含更多高度相关的翻译选项可以显著提高准确性和流畅性。然而，尽管前 10 的设置似乎可以提供更多的多样性，但它往往引入不必要或分散注意力的信息，从而导致性能下降。这在 MSLT 数据集的英译中任务中特别明

[3] https://github.com/modelscope/ms-swift

	EN → ZH						EN → DE					
	CoVoST2		MuST-C		MSLT		CoVoST2		MuST-C		MSLT	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Base Model	35.82	82.46	25.73	78.90	31.30	78.02	26.25	80.61	14.33	64.97	18.10	72.93
Translation Training	40.66	83.23	27.02	79.26	31.48	77.99	29.36	82.26	20.45	72.42	19.11	72.52
Oracle Knowledge Setting												
SALM	55.97	88.01	32.10	78.83	31.81	75.75	43.64	86.47	21.15	72.21	16.16	65.16
Retrieval-and-Demonstration	50.22	86.57	30.18	76.65	31.34	74.93	36.09	84.56	19.46	71.42	15.18	63.93
Locate-and-Focus	58.49	88.78	34.52	79.71	33.76	76.62	45.60	86.93	22.06	72.57	17.30	66.70
End-to-End Setting												
SALM	39.82	76.93	27.16	74.08	30.27	72.94	31.16	81.66	15.02	67.17	8.35	5.05
Retrieval-and-Demonstration	41.02	83.57	26.87	74.00	30.54	74.82	32.40	83.05	16.05	70.10	15.48	64.33
Locate-and-Focus	49.30	84.51	31.35	77.29	30.58	75.90	39.66	84.07	21.05	73.76	16.98	65.32

Table 7: 语音术语翻译中不同方法的性能比较，包括我们方法的变体。我们使用粗体文本来表示每个指标的最佳性能。

	CoVoST2		MuST-C		MSLT	
	TSR	BELU	TSR	BELU	TSR	BELU
EN → ZH						
Top-1	51.67	47.63	58.02	29.90	68.03	32.29
Top-5	65.53	49.30	75.78	31.35	75.51	30.58
Top-10	55.01	47.57	60.56	29.45	59.18	24.24
EN → DE						
Top-1	63.89	37.62	69.49	20.90	68.70	16.61
Top-5	77.12	39.66	77.40	21.05	72.66	16.98
Top-10	69.52	38.22	69.77	19.74	64.38	15.31

Table 8: 我们的方法在不同检索设置下的表现。Top-N 代表在我们的检索设置中提供得分最高的前 N 个翻译知识三元组。

显，其中 TSR 和 BELU 分别下降到 59.18 和 24.24。

考虑到 BLEU 指标在与人工判断的相关性上存在差距，我们在评估中补充了 COMET 翻译指标[4]，其结果显示在表 7 中。实验结果表明，我们的方法在该指标上仍表现良好，且趋势与使用 BLEU 观察到的一致。

C 数据收集的详细信息

C.1 手动标注的细节

我们聘请了三位精通英语和中文的专家，以及三位精通英语和德语的专家来帮助标注测试数据。他们的工作主要涉及三个任务。首先，他们核实由 LLM 提取的术语是否合理，确保它们是有意义的实体名称并且翻译正确。每位专家独立审核术语以防止偏见。其次，他们检查文本到语音生成的音频是否包含这些术语，确保发音准确且自然，并剔除低质量音频。最后，他们核实我们的方法定位的音频是否包含真实术语，剔除任何不完全符合标准的音频。每个样本都需要三位专家的一致同意才能保留，以确保高质量和可靠性。

[4]我们使用 wmt22-comet-da（https://huggingface.co/Unbabel/wmt22-comet-da/）。

C.2 数据样本

Instruction for Locate-and-Focus : I've provided a selection of words along with their audio from a dictionary. You can utilize these words for the upcoming speech translations. But please note that some of them may include information unrelated to the utterance. Bilingual words: Word: ..., Audio: <audio>...</audio>, Translation: ..., ..., Word: ..., Audio: <audio>...</audio>, Translation: ... . Translate from English to Chinese: <audio>common-voice-en.mp3</audio>

Instruction for SALM : I've provided a selection of words from a dictionary. You can utilize these words for the upcoming speech translations. But please note that some of them may include information unrelated to the utterance. Bilingual words: Word: ..., Audio: <audio>...</audio>, Translation: .... Translate from English to Chinese: <audio>common-voice-en.mp3</audio>

Instruction for Retrieve-and-Demostration : I have provided a pair of sentences that include important entities. You can use these entities for the upcoming speech translations. But please note that some of them may include information unrelated to the utterance. Audio: <audio>...</audio>, Translation: ... . Translate from English to Chinese: <audio>common-voice-en.mp3</audio>

Instruction for Terminology Extraction Please meticulously extract uncommon person and entity name pairs from the provided source sentences and their corresponding translations, organizing them into a list where each pair is formatted as [term - translated term] per line. Ensure the output contains no additional text or explanations. This task requires keen attention to accurately representing terms, including names, locations, and specific domain vocabulary, to ensure that each extracted pair reflects the correct relationship between the original text and its translation.
During this process, strictly follow the output format requirements, maintaining a "A - B" structure without any extra content, to ensure clarity and precision. For clarity, consider this example: when given specific source sentences and their translations, your task is to extract and list these uncommon name pairs accurately as "Term1 - Translation1" followed by "Term2 - Translation2," and so on.
If your analysis does not uncover any name pairs that are sufficiently distinctive or significant, return "None" to indicate this outcome.

C.3 收集的术语类型

为了更好地分析我们收集的术语翻译数据集，我们使用表现良好的 NER 模型 GliNER-large-v2.1 (Zaratiana et al., 2024) [5] 来检验数据中出现的术语类型。结果展示在表格 9 和表格 10 中。通过比较这两个表中的数据分布，我们发现了一些趋势。在英译中文和英译德文数据集中，“人物”和“地名”类别的术语显著多于其他类别。这表明这些类别中的术语具有重要性，并且在语音翻译任务中经常出现。此外，与其他类别相比，“食物”、“公司”和“文化”相关的术语在两个数据集中都不太常见，这可能是因为这些术语在典型的口语对话中不太常见。

[5] https://huggingface.co/urchade/gliner_large-v2.1

Category	CoVoST2	Must-C	MSLT
Person	313	191	129
Location	297	41	53
Food	12	2	3
Company	16	10	7
Biology	2	1	0
Organization	27	11	2
Health	3	1	0
Culture	22	1	2
Transport	13	4	0
Religion	62	7	0
Fashion	5	0	5
Science	2	3	2
Geography	9	0	2
Language	26	2	2
History	18	3	2
Politics	5	0	1
Architecture	5	2	0
Military	17	4	7
Environment	1	0	1
Education	29	4	3
Sport	2	0	5
Book	4	1	0
Physics	0	1	0
Game	0	0	1
Literature	1	0	0
Art	2	2	0
Music	2	0	1
Entertainment	4	0	2
Award	5	3	1

Table 9: 术语在英中数据的各种类别中的分布。

Category	CoVoST2	MUST-C	MSLT
Person	613	205	128
Location	237	32	42
Food	10	1	8
Company	13	9	10
Biology	1	1	1
Organization	6	11	2
Health	2	2	1
Culture	12	2	2
Transport	5	4	1
Religion	51	5	4
Fashion	5	0	8
Medicine	0	2	0
Science	1	1	1
Geography	0	0	1
Language	14	2	4
History	11	3	1
Architecture	1	4	0
Military	11	1	4
Environment	0	0	1
Education	14	6	3
Sport	1	0	1
Law	0	1	0
Book	1	1	0
Game	1	0	0
Literature	1	0	0
Art	1	1	0
Music	1	0	2
Entertainment	3	0	0
Award	0	3	0

Table 10: 英语到德语测试数据中不同类别的术语分布。