arXiv:2305.01620v2 [cs.CL] 6 May 2023

面向 STOP 质量挑战的流水线与端到端口语理解系统集成研究：用于口语语义解析

Siddhant Arora[1], Hayato Futami[2], Shih-Lun Wu[1],Jessica Huynh[1],Yifan Peng[1],Yosuke Kashiwagi[2], Emiru Tsunoo[2], BrianYan[1], Shinji Watanabe[1]

[1]卡内基梅隆大学， [2]索尼集团株式会社，日本

摘要

近年来，研究人员一直在尝试为口语语言理解（SLU）引入新的基准任务，例如语义解析。本文介绍了我们为口语语言理解大挑战赛质量赛道（赛道 1）提出的口语语义解析系统。该挑战赛是 ICASSP 2023 信号处理大挑战赛的一部分。针对这一任务，我们分别对端到端系统和流水线系统进行了实验。在我们的口语理解框架中，采用 Whisper 等性能强大的自动语音识别（ASR）模型，以及 BART 等预训练语言模型（LM），以提升系统性能。我们还研究了在输出层面组合多种模型的方法，最终获得了 80.8 的精确匹配准确率，并在该挑战赛中取得了第一名。

索引词——STOP 挑战、口语理解、端到端系统

1. 引言

口语理解大挑战，即口语任务导向解析（STOP）挑战，是 ICASSP 2023 信号处理大挑战的一部分，旨在构建能够将口语话语转换为语义解析序列的系统，以便语音助手执行任务。本文介绍了我们 PittOsaki 团队在赛道 1 中采用的方法，旨在利用开源模型和自动语音识别数据集提升生成的语义解析质量。

在这项工作中，我们尝试了多种流水线和端到端（E2E）口语理解方法。我们的口语理解框架采用了 WavLM 和 Hubert 等预训练自监督语音（SSL）表征。我们还纳入了 BART large 等预训练语言模型。最后，多种模型的系统组合相比基线系统显著提升了性能。

2. 方法

我们将语义解析这一口语理解任务表述为一个统一的序列到序列问题。输入是从原始音频中提取的语音特征序列，输出则是以线性化树结构表示的语义解析结果。我们的端到端（E2E）口语理解方法采用基于注意力机制的编码器—解码器架构。我们的自动语音识别模型同样基于编码器—解码器架构。在自动语音识别和口语理解的训练中，我们使用 WavLM 等自监督学习表征作为前端。我们对多个隐藏状态进行加权求和，并在训练期间冻结参数。我们还尝试在口语理解框架中使用近期发布的 Whisper 模型。由于初步实验中 Whisper 表现更优，我们对整个 Whisper 模型进行微调，而非将其用作前端。我们的自然语言理解模型通过微调预训练语言模型构建。与先前的工作类似 [1], ，我们采用两遍口语理解方法 [2], , 提升端到端口语理解模型的语义建模能力；在第二遍中，结合声学信息与由自动语音识别假设生成的预训练语言模型语义信息。受任务组合性原则启发，我们还训练了组合式端到端口语理解模型 [3] ，先将口语话语转换为一系列词元表征，再将其用于传统的自然语言理解框架。为合并多个模型的假设，我们直接采用识别器输出投票错误率降低（ROVER）方法，并根据合并后的语义解析序列计算完全匹配（EM）准确率。

3. 实验设置

我们采用 STOP 基准测试中的评估指标，即精确匹配准确率。我们的端到端口语语言理解模型的编码器为 12 层 Conformer，解码器为 6 层 Transformer。自注意力层的头数和维度分别设为 8 和 512。编码器和解码器的线性单元数均为 2048。我们使用速度扰动和 SpecAugment 进行数据增强。我们的自动语音识别模型与端到端口语语言理解模型采用相同的架构。我们研究使用 LibriSpeech 和 Common Voice 等外部自动语音识别数据集进行预训练。对于两遍式口语语言理解模型，我们的审议编码器由 4 层 Conformer 组成，第二遍解码器与自动语音识别解码器采用相同的架构。我们的组合模型在自动语音识别组件中采用与自动语音识别模型相同的架构，并使用 6 层 Transformer

	预训练模型	测试 EM (↑)
STOP 基准测试 [5]		
E2E	Wav2Vec2	68.7
E2E	Hubert	69.23
流水线	Wav2Vec2+BART-L	72.36
我们的端到端方法		
采用自监督学习	Hubert	71.6
	WavLM	73.3
采用两遍口语理解	WavLM+BART-L	75.8
采用组合式端到端口语理解	WavLM	75.9
我们的流水线方法	WavLM+BART-L	78.2

表1：STOP数据集上语义解析的完全匹配（EM）准确率。

	预训练模型	测试 EM（↑）
我们端到端的方法		
搭配 SSL	Whisper	78.8
	Whisper large	79.1
采用两遍式口语理解	采用 Whisper 转写文本	77.0
采用组合式端到端口语理解	采用 Whisper 转写文本	77.4
我们的流水线方法	最佳自动语音识别+BART-L	80.5
	最佳自动语音识别+T5	80.1
我们的系统组合	4 个最佳模型	80.8

表2：使用 Whisper 在 STOP数据集上进行语义解析的完全匹配（EM）准确率。

编码器，并在自然语言理解组件中使用 6 层 Transformer 解码器。我们采用 Dropout 和标签平滑。对于预训练语言模型，我们使用 BART large 和 T5 large，并通过 HuggingFace Seq2Seq Trainer 对其进行训练。我们在词汇表中添加了用于槽位和意图标签的特殊标记。

我们还使用 Whisper medium 进行自动语音识别，并使用 Whisper medium 和 large 模型进行口语理解。对于自动语音识别，我们在 3 种设置下进行训练：(a) 使用小写转录文本；(b) 使用原始大小写的转录文本（如“话语”字段所示），称为带原始大小写的Whisper；(c) 使用原始大小写的转录文本，并更频繁地保存检查点（即每 1k 次迭代保存一次，而非每个训练轮次保存一次），称为使用高频检查点的Whisper。与预训练语言模型类似，在训练端到端口语理解模型时，我们会在词汇表中添加用于槽位和意图标签的特殊标记。有关模型和配置文件的更多详情将作为 ESPnet-SLU [4] 工具包的一部分公开发布。

4. 结果

不使用 Whisper 模型时，我们在语义解析任务上的结果如表1所示。两遍式和组合式端到端口语理解模型的表现优于传统端到端模型。我们进一步观察到，流水线模型显著优于所有端到端口语理解模型。

预训练模型	预训练数据集	测试词错误率（↓）
STOP 基准测试 [5]		
Wav2Vec2	STOP	4.45
Hubert	STOP	4.26
我们的自动语音识别模型		
Hubert	STOP	3.8
WavLM	STOP	3.3
带语言模型的 WavLM	STOP	3.1
带语言模型的 WavLM	LibriSpeech+Common Voice+STOP	2.7
Whisper	STOP	2.4
带原始大小写的Whisper	STOP	2.3
Whisperw/ 频率检查点	STOP	2.3
我们的系统组合	4个最佳模型	2.2

表3：STOP数据集上的词错误率（WER）。

表2显示，对 Whisper 进行微调有助于提升端到端口语理解的性能，因为 Whisper 使用大量带标签的数据进行过训练。表3同样显示，Whisper 在自动语音识别任务上取得了显著提升。我们还尝试在推理过程中直接将 Whisper 转录文本用于两遍式和组合式端到端口语理解模型，并观察到性能显著提升。由于时间有限，我们未能研究如何将 Whisper 模型整合到组合式和两遍式端到端口语理解模型中，但将在未来的工作中进行研究。

最后，我们使用 ROVER 对由表现最好的 4 个自动语音识别模型生成的假设结果（表 3）进行融合，将词错误率（WER）降至 2.2。利用这一自动语音识别转录文本，我们能够显著提升流水线系统的性能。此外，我们进一步使用 ROVER 融合表现最好的 4 个口语理解模型（表 2），使完全匹配率（EM）达到 80.8。

5. 致谢

本研究通过先进网络基础设施协调生态系统：服务与支持（ACCESS）计划的项目编号 CIS210014，使用了 PSC 的 Bridges2 系统和 NCSA 的 Delta 系统。该计划由美国国家科学基金会资助，资助编号为 #2138259、#2138286、#2138307、#2137603 和 #2138296。Jessica Huynh 获得了美国国家科学基金会研究生研究奖学金资助，资助编号为 DGE1745016 和 DGE2140739。本文所表达的观点不一定代表该资助机构的观点。

6. 参考文献

[1] D. Le、A. Shrivastava、P. D. Tomasello 等人，《用于设备端口语理解的审议模型》，载于 Interspeech 会议论文集，2022 年，第 3468–3472 页。[2] S. Arora、S. Dalmia、X. Chang 等人，《两遍低延迟端到端口语理解》，载于 arXiv 预印本 arXiv:2207 06670，2022 年。[3] S. Arora、S. Dalmia、B. Yan 等人，《使用组合式端到端模型进行口语理解的词元级序列标注》，载于 EMNLP 会议论文集，2022 年。

[4] S. Arora、S. Dalmia、P. Denisov 等人，《ESPnet-SLU：通过 ESPnet 推进口语理解》，载于 ICASSP 会议论文集，2022 年，第 7167–7171 页。[5] P. Tomasello、A. Shrivastava、D. Lazar 等人，《STOP：一个面向口语任务的语义解析数据集》，载于 CoRR。