中文口语语义理解大作业报告

常烁晨 521021910369
杜思霖 521021910470
徐阳 521021910363

2024 年 3 月 18 日

目录

1 简介 Introduction **2**
1.1 任务概述 . . . 2
1.2 实现内容 . . . 2
1.2.1 数据增强 . . . 2
1.2.2 模型方法 . . . 2
1.2.3 训练策略 . . . 2

2 数据增强与分析 Data Augmentation and Analysis **3**
2.1 替换关键词 . . . 3
2.2 出错数据增强 . . . 3
2.3 对话历史的理解 . . . 3

3 模型/方法 Methodology **4**
3.1 baseline . . . 4
3.2 BERT . . . 5
3.2.1 BERT-WWM . . . 5
3.2.2 RoBERTa . . . 5
3.2.3 MacBERT . . . 5
3.3 CRF . . . 6

4 训练策略 Training Strategy **7**
4.1 学习率设置 . . . 7
4.2 课程学习 . . . 7

5 分析 Analysis **8**
5.1 数据增强 . . . 8
5.2 模型与方法对比 . . . 8
5.2.1 模型选择 . . . 9
5.2.2 方法选择 . . . 9
5.3 课程学习 . . . 9

6 总结 Conclusion **9**

A 附录 Appendix **11**

1 简介 Introduction

1.1 任务概述

口语语义理解（Spoken Language Understanding, SLU）是任务型对话系统的核心组成部分，它的目标是从用户的口语中准确地提取用户的意图和信息，为对话管理系统提供决策所需的信息。这不仅对于理解用户的请求至关重要，也是确保系统能够顺利完成用户任务的基础。

SLU 通常涉及将用户的口语转换为结构化的语义单元，以便于机器处理。每个语义单元包含三个关键部分：动作（act），语义槽（slot），和槽值（value）。语义槽定义了可以接受的信息类型，例如日期、地点或数量等，而槽值则是这些语义槽的具体填充内容。对话动作描述了用户的话语背后的意图，如请求、确认或提供信息等。这些动作反映了对话中的行为特征，并结合了轮次信息来推动对话向前发展。SLU 系统采用对话语义动作来表达用户的意图，将行为特征与语义槽值对结合，形成如 act(slot) = value 的简明结构，这不仅反映了对话中的动作，还包括了相关的、具体的语义信息。

1.2 实现内容

在本项目中，我们从数据增强、模型方法、训练策略的角度入手对口语语义理解问题进行全面探究，多方面对原有的模型进行改进。我们具体从以下几个方面对中文口语语义理解问题进行探索和改进。

1.2.1 数据增强

通过替换语义三元组中的 value 槽值，我们生成了大量新数据，大大扩充了数据集，增强了模型面对未知数据时的泛化能力，防止数据集不平衡。此外，通过对训练时预测出错的数据进行专门的数据增强，也提高了模型的表现。

1.2.2 模型方法

在课程项目 baseline 提供的 BiLSTM+ 词性标注方法的基础上，我们尝试了 chinese-bert-wwm-ext、chinese-roberta-wwm-ext、chinese-macbert-base 几种不同的 bert [1] 模型，并将条件随机场（CRF）应用于词性标注中，并比较了不同方法的模型效果。

1.2.3 训练策略

我们通过对比测试对模型的最佳学习率进行了探究，并应用了“课程学习（Curriculum Learning）[2]”的训练策略以加快收敛和提高模型表现。

报告的后续部分将详细介绍每种方法的实现原理、实现过程、面临的问题以及性能表现，旨在通过比较不同方法的优势和局限性，更深入地理解口语语义理解问题，为相关领域的研究提供参考和启示。

2 数据增强与分析 Data Augmentation and Analysis

数据集内容为导航的语音对话数据，每条数据包含了原始的语音识别文字数据"asr_1best"，经过人工校正的语音识别数据 "manual_transcript" 以及语句中的语义信息 "semantic"。我们发现分别使用"asr_1best" 和 "manual_transcript" 当做语句的模型表现有着非常大的差异，其原因在于人工校正的语音文字正确率高，而 ASR 识别的文字可能存在错误，对模型效果的影响比较大。

2.1 替换关键词

在语义三元组 <act>(<slot>=<value>) 中，每一个 slot 槽值都有一个对应的 value 槽值，而 value 槽值的待选项均可以在 data 文件夹下的数据中得到。我们发现对于三元组数据，可以用待选项中的 value 替换当前 value 槽值而不会对句子的合理性产生太大影响。

举例来说，对于 "导航到凯里大十字"，其语义三元组为 "inform(操作 = 导航)" 和 "inform(终点名称 = 凯里大十字)"，那我们就可以生成一条内容为 "导航到北京" 的数据，其语义三元组为 inform(操作 = 导航)" 和 "inform(终点名称 = 北京)"。同理，我们还可以替换这里的导航为 "定位"、"途径" 等等。

在实验中，我们对地名、请求类型、路线偏好和对象这几个 slot 槽值对应的 value 进行了替换，使模型的表现得到了一定的提升。

2.2 出错数据增强

事实上，数据集中不是所有数据都需要扩充。过度的数据增强会显著增加训练时间和成本，同时可能引入一些较为罕见的特征，如地名等，这有可能导致数据的过拟合。

因此，我们选择了一种更为针对性的数据增强策略。对于未经过数据增强的模型，我们首先获取其在训练集上预测错误的数据，如图1。然后仅对这些数据进行增强。这种方法增强了模型处理特定难题的能力。经过测试，我们发现，相比于对全数据集进行增强，这种针对出错数据的增强策略能够取得相似的模型效果，同时还大幅减少了训练时间和内存使用。

2.3 对话历史的理解

在多轮对话情景中，对话动作关注于对话参与者的行为 <act>，在此项目的背景下有"inform" 和"deny" 两种。相比之下，槽值 <slot> 和语义槽 <value> 则更多地涉及特定的信息元素，如起点、终点、操作等。这些元素虽然在对话中很重要，但它们通常是静态的，不像对话行为 <act> 那样能直接反映对话历史的动态变化。因此，<slot> 和 <value 与对话历史之间的直接关联远不如 <act> 那样显著。

而此项目中关于导航口语语义三元组的 <act> 只有两种值，且其中有关"deny" 的信息只有几条，数量极少，难以通过对话历史对模型提供较大的改进。

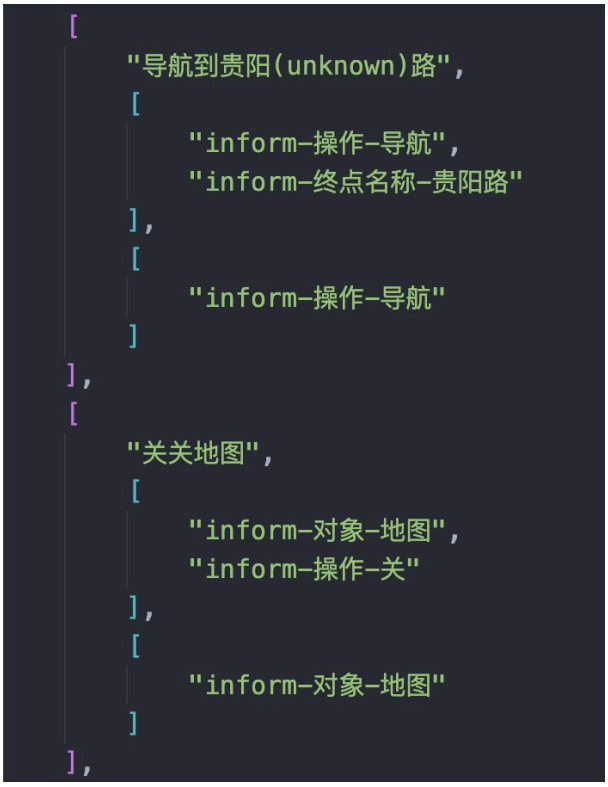

图 1: 出错数据示例

3 模型/方法 Methodology

本部分将依次介绍项目中使用的不同神经网络模型。本项目实现了 baseline 对应的 BiLSTM 方法，以及项目使用的 Transformer 模型，并且针对每一种神经网络结构，项目都实现了条件随机场（Conditional Random Field，简称 CRF）方法。通过对比分析不同模型及方法对语义三元组的识别效果，展示本项目在口语语义理解上做出的改进。

3.1 baseline

双向长短时记忆网络（BiLSTM）是一种序列处理模型，其双向性体现在其结合了两个独立的长短时记忆网络（LSTM）来处理上下文中形成序列的数据。LSTM 是一种特殊的 RNN，旨在解决传统 RNN 在处理长序列时的梯度消失问题。通过引入门控机制（包括输入门、遗忘门和输出门），LSTM 能够更好地捕捉长距离依赖。BiLSTM 包括两个 LSTM 层，一个沿着时间正向运行（处理从头到尾的序列），另一个沿着时间反向运行（处理从尾到头的序列）。这样，每个时间点的输出都包含了过去（正向）和未来（反向）的信息。正向和反向 LSTM 的输出通常在每个时间步被合并（例如，通过拼接）来形成最终的输出，这样每个时间步的输出都包含了整个序列的上下文信息。这种架构适用于需要理解整个输入序列以做出预测的任务，尤其是本项目 SLU 的语义理解任务。

在本项目中，双向长短时记忆网络（BiLSTM）被应用于一种特定的序列标注任务，类似于命名实体识别（NER），采用了广泛使用的 BIO（Begin-Inside-Outside）标注方案。具体来说，项目中的文本首先输入 BiLSTM 网络。该网络的主要任务是对对话中的每个字进行 BIO 标签的预测，从而实现文本中动作（act）和语义槽（slot）的识别，并据此确定相应的值（value）。网络输出的标

签预测结果与实际标签之间的差异通过交叉熵损失函数（Cross-Entropy Loss）进行量化，通过反向传播算法更新 BiLSTM 网络的参数，从而优化模型的性能。

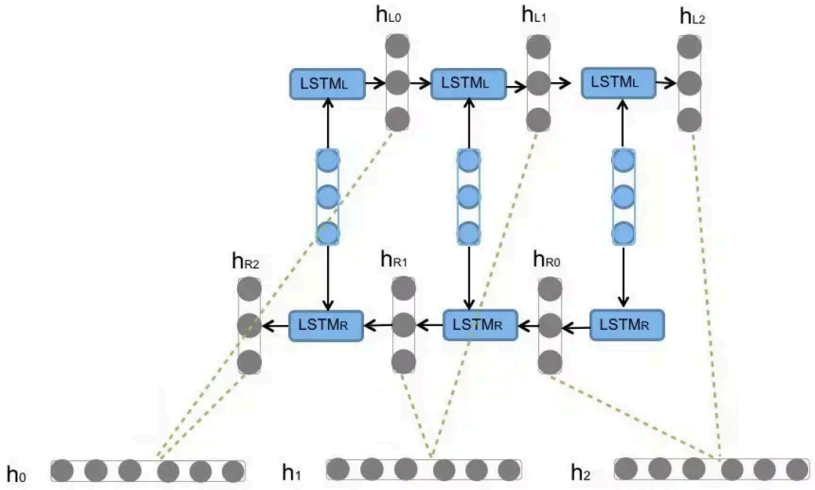

图 2: BiLSTM 模型示意图

3.2 BERT

BERT（Bidirectional Encoder Representations from Transformers）是一个由 google 在 2020 年提出的基于 Transformer 的预训练语言模型。Bert 在预训练过程中应用了“掩码语言模型”（Masked Language Model, MLM）任务，在每次训练过程中，其中一部分输入词被随机掩码，模型需要预测这些掩码词。通过在大规模数据集上的预训练过程以及在少量标注数据上的微调，BERT 能够考虑句子中每个词的左右上下文，从而生成更深层次、更精细的语言表示。

在本项目中，我们尝试使用预训练的 BERT 模型来替换 baseline 中使用的 BiLSTM 模型。BERT 模型有许多变种，针对本项目的中文口语语义理解任务，我们尝试了三种针对中文语言的 BERT 模型。

3.2.1 BERT-WWM

chinese-bert-wwm-ext：BERT-WWM 是一个 BERT 模型的中文版本，其中”WWM” 代表”Whole Word Masking”（全词遮蔽）。在这个版本中，预训练时的遮蔽策略被修改，以确保整个中文词被一起遮蔽，而不是仅仅遮蔽词中的单个字符。

3.2.2 RoBERTa

chinese-roberta-wwm-ext：这是基于 BERT 的 Roberta 模型的中文版本。Roberta（A Robustly Optimized BERT Pretraining Approach）是 BERT 的一个优化版本，它通过更大的批量大小、更长的训练时间以及不使用 Next Sentence Prediction（NSP）任务等调整来优化 BERT 模型。

3.2.3 MacBERT

chinese-macbert-base：MacBERT（Masked and Corrected BERT）是另一种 BERT 的改进版本，在 MacBERT 中，掩码位置不仅仅是简单地被一个随机掩码符号替换，而是用与原始词语语义

相近但不完全相同的词语替换，这样做的目的是为了让模型学习更加深入和准确的上下文信息。

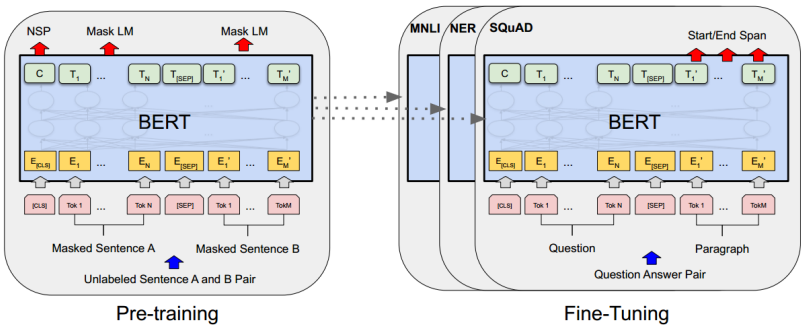

图 3: BERT 模型示意图

3.3 CRF

条件随机场（Conditional Random Field，简称 CRF）是一种用于建模序列数据中元素之间条件关系的统计建模方法。它是一种用于结构化预测的无向图模型，特别适用于处理序列数据，尤其是自然语言处理任务。在本项目中，口语语意理解作为一种经典的 NLP 任务，将 CRF 方法结合到 backbone 之上，是一种可行的对序列信息进行预测优化的方式。

CRF 能够考虑到序列中的上下文信息，使得其在序列标注任务中比独立分类器（如逻辑回归）表现更好。具体而言，CRF 在进行预测时会考虑整个输入序列，通过对上下文信息的综合，可以找到全局最优的标注序列。通过将 CRF 与神经网络架构相结合，模型能够更加有效地捕捉序列中的依赖关系，提高序列标注和其他任务的性能。相应的，CRF 在处理长序列时的训练和推断过程较为计算密集，实际训练时增加了 CRF 处理的模型，在训练时间上有显著的增加。

在 BiLSTM 中，原本的全连接层使用一个简单的前馈神经网络解码器，将 BiLSTM 的输出映射到标签空间，忽略填充标签使用交叉熵损失函数，经过 softmax 后得到预测概率，并通过交叉熵计算损失。引入 CRF 方法后，将原本的前馈神经网络替换成一个 CRF 层，对相应的输入输出向量进行处理。BiLSTM 首先用于处理输入序列，提取特征并捕捉序列中的上下文信息，随后 CRF 层通过对标签之间依赖性的考量，尽可能确保预测的标签序列在整体上是合理的。

在 BERT 中，CRF 与神经网络的结合方式也几乎相同，只不过将模型的 backbone 部分由 BiLSTM 改为 BERT，其余的编程细节非常类似，因此不再赘述。

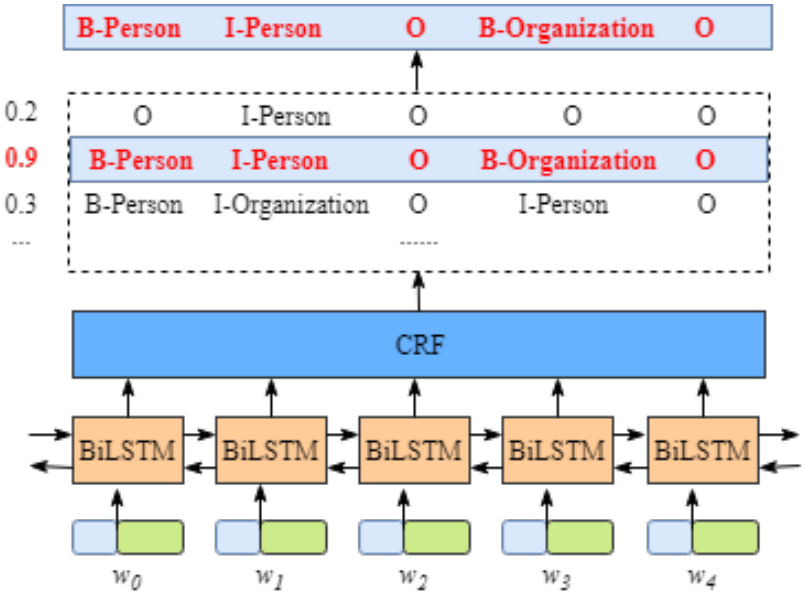

图 4: CRF 应用于 BiLSTM 模型示意图

4 训练策略 Training Strategy

4.1 学习率设置

bert 模型的最佳学习率量级在 1e-5 左右，对 bert 模型 chinese-bert-wwm-ext 设置不同的学习率, 使用"asr_1best" 数据，探究最佳学习率，结果见表1。

lr	Accuracy/%	Precision/%	Recall/%	F1 score/%
5e-6	78.77	84.74	81.65	83.17
1e-5	**79.22**	**84.75**	**82.27**	**83.49**
2e-5	79.22	84.61	81.96	83.26
5e-5	78.21	83.91	81.02	82.44

表 1: 不同学习率下的 bert 性能

由结果可知，同一数量级下学习率对结果影响不大，lr=1e-5 时，模型效果最好。

4.2 课程学习

课程学习（Curriculum Learning）是一种训练策略，它模仿人类课程中有意义的学习顺序，从较容易的数据到较难的数据来训练机器学习模型。作为一个即插即用的模块，CL 策略在计算机视觉、自然语言处理等广泛场景中展示了其在提高各种模型的泛化能力和收敛速度方面的威力。课程学习有很多方法，本实验中，我们选择由易到难训练数据的方法，即将数据按照特定的标准评估学习难度，并按难度排序，初始时由易到难地使用数据训练模型，之后使用打乱的数据继续训练。

对于数据难度的测量，我们采用了多个角度的评估标准：

- 每包含一个语义三元组，难度加一

- 人工校正语音文字“manual_transcript”和 ASR 识别语音文字数据”asr_1best” 不同，难度加一
- “manual_transcript”每包含一个字符，难度加 0.1

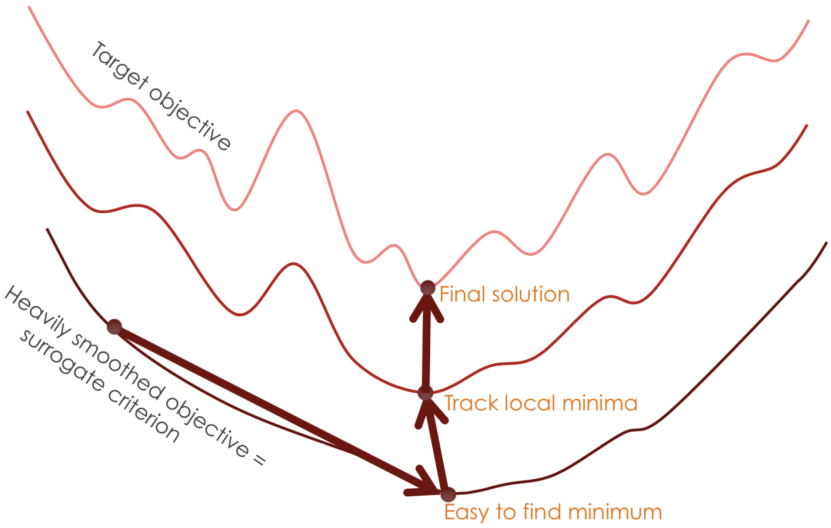

图 5: 课程学习示意图

5 分析 Analysis

5.1 数据增强

在 chinese-bert-wwm-ext 模型中使用了“manual_transcript”数据，在不同的数据增强场景下，模型性能如表2所示，可见数据增强都能提升模型性能，而全数据增强的训练时间约是出错数据增强方法的四倍。相比于对全数据集进行增强，这种针对出错数据的增强策略能够取得相似的模型效果，同时还大幅减少了训练时间和内存使用。

Method	Accuracy/%	Precision/%	Recall/%	F1 score/%
未增强	95.08	95.56	96.66	96.11
全数据增强	95.98	96.19	97.39	96.79
出错数据增强	95.40	95.18	96.87	96.02

表 2: Comparison of Different Methods

5.2 模型与方法对比

本项目实现了四种神经网络架构（BiLSTM、Bert、Roberta、Macbert）的普通版本以及 CRF 版本，下表中展示了不同模型以各自的最优学习率，在“asr——1best”数据集上训练后的测试结果。

接下来将从模型与方法两个角度，对实验结果进行分析和整理。

model(CRF)	Accuracy/%	Precision/%	Recall/%	F1 score/%
BiLSTM	70.95（72.51）	82.24（83.92）	73.41（74.56）	77.58（78.96）
Bert	79.22（79.22）	84.75（83.62）	82.27（81.44）	83.49（82.51）
Roberta	78.10（78.32）	84.02 (82.94)	80.60（81.13）	82.28（82.02）
Macbert	78.10（78.21）	81.80（82.94）	80.60（81.13）	81.20（82.02）

表 3: Comparison of Different Models

5.2.1 模型选择

经过对比可以发现，使用了 Bert 作为预训练语言模型架构的模型，其 Accuracy 整体要比使用 BiLSTM 的模型高出 8% 左右，而 Bert 的三个变体模型，总体而言预测结果差距不大，基础的 chinese-bert-wwm-ext 模型取得了最好的成绩。

使用 BiLSTM 架构的模型，其最优学习率在 1e-3 左右，而 Bert 及其变体的最优学习率都在 1e-5 数量级，这说明 Bert 模型对于数据的变化更加敏感，能更好提取输入文本序列的深层信息。此外，在实际训练过程中发现，即使 baseline 中设置的学习率明显更大，实际收敛速度却不如 Bert。

5.2.2 方法选择

对于上述四个模型，本项目对比了使用普通前馈神经网络作为输出层和使用 CRF 层作为输出层的结果。通过表格可以看到，对于传统的 BiLSTM 结构，引入 CRF 可以得到一定程度上的性能提升，但是对于 Bert 及其变体，引入 CRF 对预测准确率的影响较为微弱。

引起这一结果的原因可能是 Bert 作为强大的预训练掩码语言模型，其本身对于输入文本的上下文有深层次的理解，同时其具有相当的泛化性，直接通过前馈神经网络输出概率就可以提现模型的认知能力，使用 CRF 进行额外信息综合处理，未必能够在性能上有所提升。

但是对于传统的 BiLSTM 而言，CRF 对整个输入序列上下文信息的综合就能够弥补长短时记忆网络对于上下文关系理解的缺憾，因此引入 CRF 方法后获得了较为明显的性能提升。

5.3 课程学习

同样的，我们在 chinese-bert-wwm-ext 模型中使用了”asr_1best” 数据，得到如表4中结果。可以看到在使用课程学习时，模型性能得到了近 1% 的提升。并且当我们使用课程学习时，模型收敛的更快。

Method	Accuracy/%	Precision/%	Recall/%	F1 score/%
Baseline	79.22	84.75	82.27	83.49
Curriculum Learning	80.44	84.78	83.38	84.07

表 4: Comparison of Different Methods

6 总结 Conclusion

在本项目中，我们基于 baseline 中原始的 BiLSTM 方法从数据增强、模型设计以及训练策略三个角度针对中文口语语义理解问题进行了全面深入的分析与改进，并通过测试验证了改进的有效

性。相对于传统模型如 BiLSTM，BERT 及其变体在中文口语语义理解方面表现优异；数据增强和课程学习策略对提高模型性能和加速收敛也有显著作用。

由于尝试时间及内容有限，我们的方法还有很大局限性，当前的研究主要集中于特定的数据集，可能缺乏足够的多样性和复杂性。此外，模型的泛化能力和对不同中文方言的适应性还有待提高。未来改进的方向可以包括：

1. **扩大和多样化数据集：**包含更多样化的口语表达和方言，以提高模型对不同口语风格的适应性和准确性。
2. **深入探索模型结构：**实验不同的模型结构或混合模型，以进一步提升性能。
3. **优化训练策略：**探索更高级的训练技术，如对抗性训练或迁移学习，以增强模型的泛化能力。
4. **实际场景测试：**在更多真实应用场景中测试模型，以评估和优化其实际表现。
5. **性能和资源效率的平衡：**探索减少模型复杂性和计算资源消耗的方法，同时保持高性能。

这些方向有助于在未来的研究中进一步提升中文口语语义理解的准确性和应用范围。

A 附录 Appendix

分工贡献：

- 常烁晨（100%）：模型优化修改、模型测试、报告撰写
- 徐阳（100%）：数据增强、课程学习代码构建与测试、报告撰写
- 杜思霖（100%）：模型框架代码构建与测试、报告撰写

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning, 2021.