Mobvoi NLU for SMP2019-ECDT-Task-1

作者：张贺, 祝文博, 孟振南, 齐乔松, 赵广璞, 林士翔

出门问问信息科技有限公司

摘要

自然语言理解模块是人机对话系统中必不可少的部分，在SMP2019-ECDT任务1中，本文在BERT预训练模型的基础上，结合出门问问知识图谱、数据增强以及语义框架检测，搭建了一套端到端的自然语言理解系统，该系统在SMP2019-ECDT任务1上句准率为72.23%，取得了任务1第一名的成绩。

关键词：**中文人机对话, 自然语言理解, BERT, 出门问问, 知识图谱, 数据增强**

1.引言

随着人工智能领域的崛起，人机对话系统在生活中的应用越来越广泛，自然语言理解模块作为人机对话系统的核心模块之一备受工业界和学术界关注。SMP2019-ECDT任务1是专门为自然语言理解模块设计的评测任务，在该任务中，凭借出门问问NLU团队自主研发的QAP平台(Query Analysis Platform)快速迭代，我们提交了50个评测版本，尝试了多种模型和方法，例如BERT预训练模型、出门问问知识图谱、数据增强、语义框架检测、传统机器学习模型等，最终的提交系统是在BERT预训练模型的基础上，结合出门问问知识图谱、数据增强以及语义框架检测。在本文中，除了会分享最终版本的技术方案，也会分享比赛过程中其他尝试的经验。

2.模型及方法介绍

在SMP2019-ECDT任务1中，我们尝试了多种模型和方法，例如BERT预训

练模型、出门问问知识图谱、数据增强、语义框架检测、传统机器学习模型和语义规则等，通过将不同模型和方法进行组合，搭建了多套自然语言理解系统，下面会从中选择部分具有代表性的系统进行详细介绍。

2.1 基于传统机器学习模型的自然语言理解系统

基于传统机器学习模型的自然语言理解系统，主要通过传统机器学习模型进行NLU处理，分类任务使用最大熵模型，序列标注任务使用CRF模型。系统框架如图1所示。

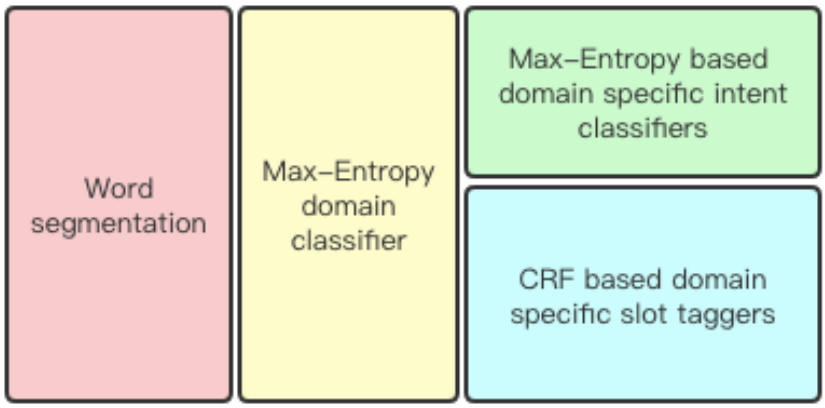

图1. 基于传统机器学习模型的自然语言理解系统

在该系统中，我们首先会对query进行分词预处理，然后经过基于最大熵的Domain分类模型，得到Domain类别之后，再分别进行Intent分类和Slot标注。其中Intent分类模块包含多个基于最大熵的Intent分类模型，每个Domain都有一个对应的Intent分类模型；Slot标注模块包含多个基于CRF的Slot标注模型，每个Domain都有一个对应的Slot标注模型。以SMP2019-ECDT任务1为例，该系统共包含1个29 Domain分类模型，29个Intent分类模型，29个Slot标注模型。

为方便描述，下文以ML-NLU表示基于传统机器学习模型的自然语言理解系统。

2.2 基于传统机器学习模型+知识图谱的自然语言理解系统

基于传统机器学习模型+知识图谱的自然语言理解系统，在ML-NLU的基础

上引入了出门问问知识图谱。出门问问知识图谱是针对人机对话领域而专门设计的知识图谱，包含实体图谱和概念图谱两部分，经过近7年的沉淀，实体图谱积累了1000万实体和2000万关系，概念图谱积累了5000万实体、50万概念和2.5亿关系。知识图谱在NLU模块中有多种用途，其中两个用途就是用于文本分类和序列标注，用于文本分类的知识图谱是从一个通用的小规模知识图谱，这个通用的小规模知识图谱包含人机对话系统支持的各个垂直领域，可以根据人机对话系统支持垂直领域的不同进行定制；用于序列标注的知识图谱是各领域定制的小规模知识图谱。引入知识图谱后的ML-NLU系统框架如图2所示。

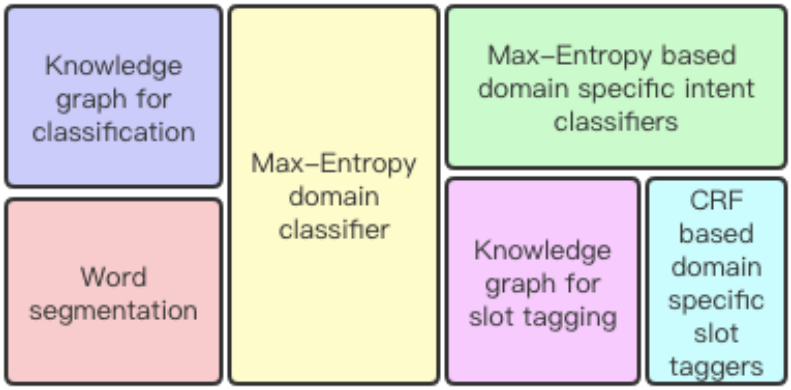

图2. 基于传统机器学习模型+知识图谱的自然语言理解系统

在该系统中，预处理阶段除了进行分词处理，还要根据知识图谱检测query中出现了哪些实体及其概念，实体信息会作为重新调整分词的依据，概念信息会作为文本分类的特征用于Domain分类和Intent分类。在Slot标注时，还会将领域定制知识图谱提供的概念信息作为CRF特征。

为方便描述，下文以ML-KG-NLU表示基于传统机器学习模型+知识图谱的自然语言理解系统。

2.3 基于BERT的自然语言理解系统-baseline

基于BERT的自然语言理解系统-baseline，是在BERT预训练模型的基础上，通过fine-tuning来训练Domain分类、Intent分类和Slot标注任务。该系统中，fine-tuning之后的BERT预训练模型是被Domain分类、Intent分类和Slot标注三个任务

共同使用的，而不是为Domain分类、Intent分类和Slot标注三个任务分别fine-tuning一个BERT模型。系统框架如图3所示。

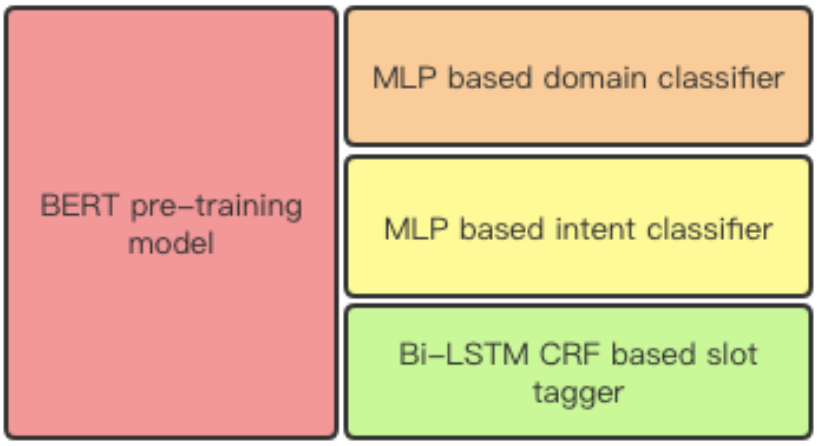

图3. 基于BERT的自然语言理解系统-baseline

该系统的架构比较简单，在BERT预训练模型之后，拼接一个全连接层用于Domain分类，拼接一个全连接层用于Intent分类，拼接一个Bi-LSTM-CRF用于Slot标注。如果以训练数据集中的语义框架为标准，那么该系统有一个明显缺陷，由于所有Domain的Intent分类和Slot标注分别共用一个模型，所以会出现不符合语义框架的情况出现，例如Domain A中只有Slot_1和Slot_2，但是系统可能会标注出Slot_3，针对这类问题，我们从模型和语义框架的角度提出了解决方案，具体方案会在后边介绍。

为方便描述，下文以BERT-baseline-NLU表示基于BERT的自然语言理解系统-baseline。

2.4 基于BERT的自然语言理解系统-BIE-restriction

基于BERT的自然语言理解系统-BIE-restriction，是在BERT-baseline-NLU的基础上，针对Slot标注进行优化。系统框架和BERT-baseline-NLU基本一致，不同的是，在Slot标注部分采用BIE形式的标记，后续只选取满足BIE限制的标记。这种方法虽然会导致Slot标记数量翻倍，但是可以明显降低Slot模型的误召回。系统框架如图4所示。

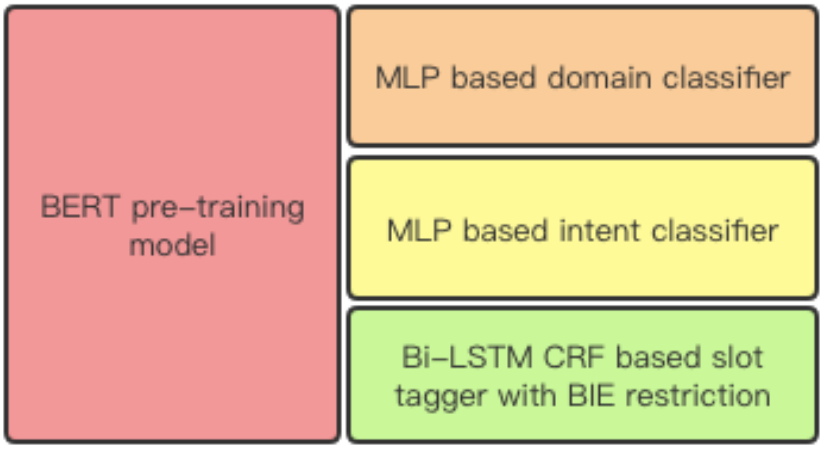

图4. 基于BERT的自然语言理解系统-BIE-restriction

为方便描述，下文以BERT-BIE-NLU表示基于BERT的自然语言理解系统-BIE-restriction。

2.5 基于BERT的自然语言理解系统-domain-intent-joint

基于BERT的自然语言理解系统-domain-intent-joint是在BERT-BIE-NLU的基础上进行改进，主要是为了解决2.3节提到的Intent分类可能不符合语义框架的情况。具体的方法就是联合Domain分类任务和Intent分类任务，通过在BERT预训练模型后拼接一个全连接层用于Domain-Intent分类，这样模型就不会预测出不属于Domain的Intent。系统框架如图5所示。

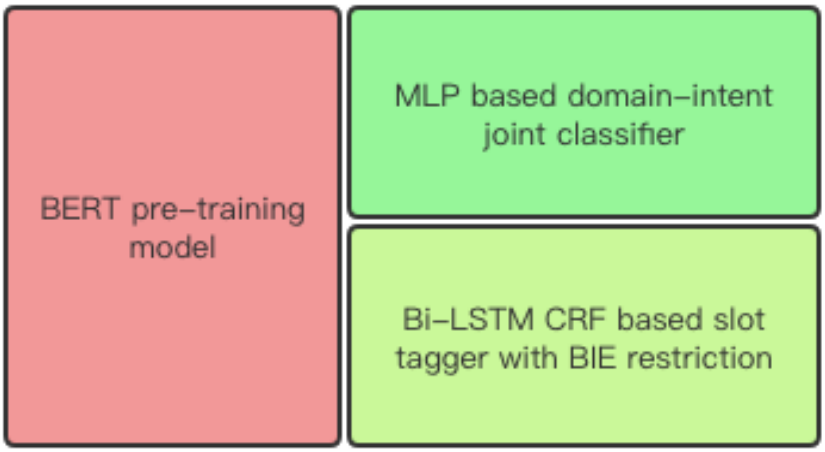

图5. 基于BERT的自然语言理解系统-domain-intent-joint

为方便描述，下文以BERT-JOINT-NLU表示基于BERT的自然语言理解系统-

domain-intent-joint。

2.6 基于BERT的自然语言理解系统-NLU-frame-filter

基于BERT的自然语言理解系统-NLU-frame-filter是在BERT-JOINT-NLU的基础上进行改进，主要是为了解决2.3节提到的Slot标注可能不符合语义框架的情况。具体的做法就是根据训练数据集归纳得出语义框架，再用语义框架过滤模型预测的Slot标注结果，过滤掉不属于Domain的Slot标记，如果有足够的数据或者明确的语义框架定义，不仅可以做Slot标记过滤，还可以做Slot标记转换。系统框架如图6所示。

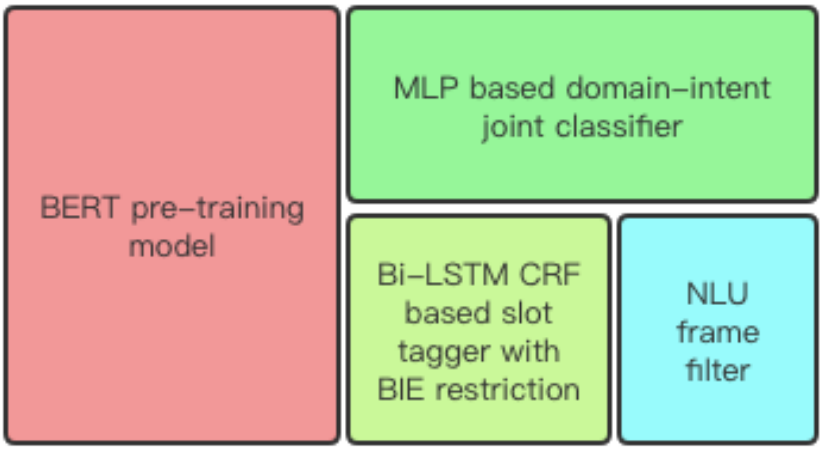

图6. 基于BERT的自然语言理解系统-NLU-frame-filter

为方便描述，下文以BERT-FRAME-NLU表示基于BERT的自然语言理解系统-NLU-frame-filter。

2.7 基于BERT+KG的自然语言理解系统-classification

基于BERT+KG的自然语言理解系统-classification是在BERT-JOINT-NLU的基础上引入出门问问知识图谱，用于提升Domain-Intent分类效果。系统框架如图7所示。

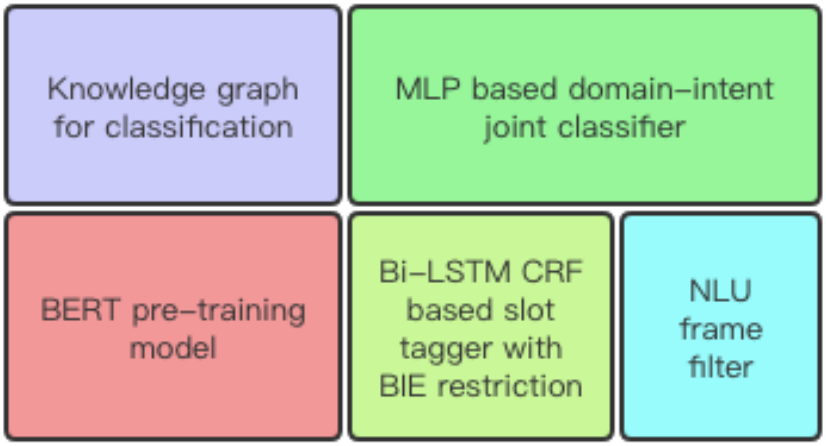

图7. 基于BERT+KG的自然语言理解系统-classification

该系统中，对于Domain-Intent分类器部分，MLP的输入主要包括两部分：BERT输出的句子嵌入和query匹配知识图谱的概念嵌入。分类器结构如8所示。

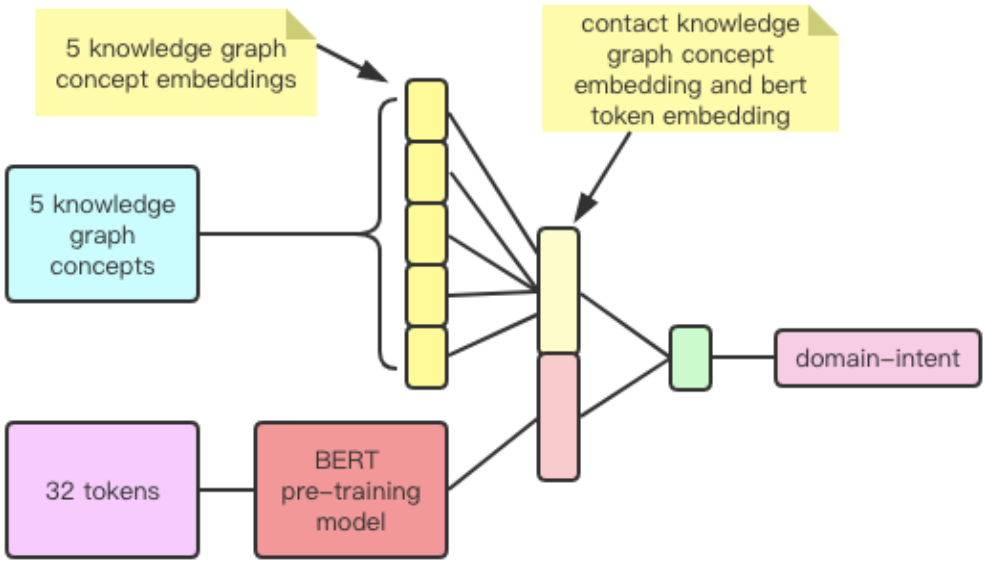

图8. 基于BERT+KG的Domain-Intent分类模型

在Domain-Intent分类模型中，我们限制一个query至多匹配5个知识图谱概念，并将每个知识图谱概念表示成300维分布式向量，然后将5个知识图谱概念嵌入拼接在一起，再连接一个全连接层，转换为768维向量，再和BERT预训练模型输出的768维向量拼接在一起，再连接一个全连接层，进行Domain-Intent分类。

为方便描述，下文以BERT-CLS-KG-NLU表示基于BERT+KG的自然语言理解系统-classification。

2.8 基于BERT+KG的自然语言理解系统-end-2-end

基于BERT+KG的自然语言理解系统-end-2-end是在BERT-CLS-KG-NLU的基础上进一步引入出门问问知识图谱，用于提升Slot标注的效果。系统框架如图9所示。

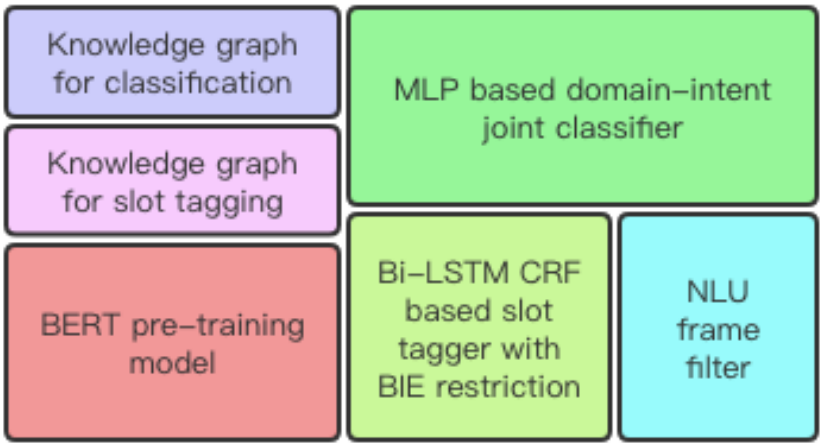

图9. 基于BERT+KG的自然语言理解系统-end-2-end

该系统中，对于Slot标注部分，Bi-LSTM CRF的输入主要包括两部分：BERT输出的token嵌入和每个token匹配的知识图谱概念嵌入。关于知识图谱概念嵌入的使用方式和Domain-Intent分类模型类似，我们限制每个token至多匹配5个知识图谱概念，并将每个知识图谱概念表示成300维分布式向量，然后将5个知识图谱概念嵌入拼接在一起，再连接一个全连接层，转换为768维向量，再和BERT预训练输出的token对应的768维向量拼接在一起，作为Bi-LSTM CRF的输入，进行Slot标注。

为方便描述，下文以BERT-E2E-KG-NLU表示基于BERT+KG的自然语言理解系统-end-2-end。

2.9 其他技巧

SMP2019-ECDT任务1包含29个Domain，主办方共提供了2579条训练数据，训练数据分布如图10所示。

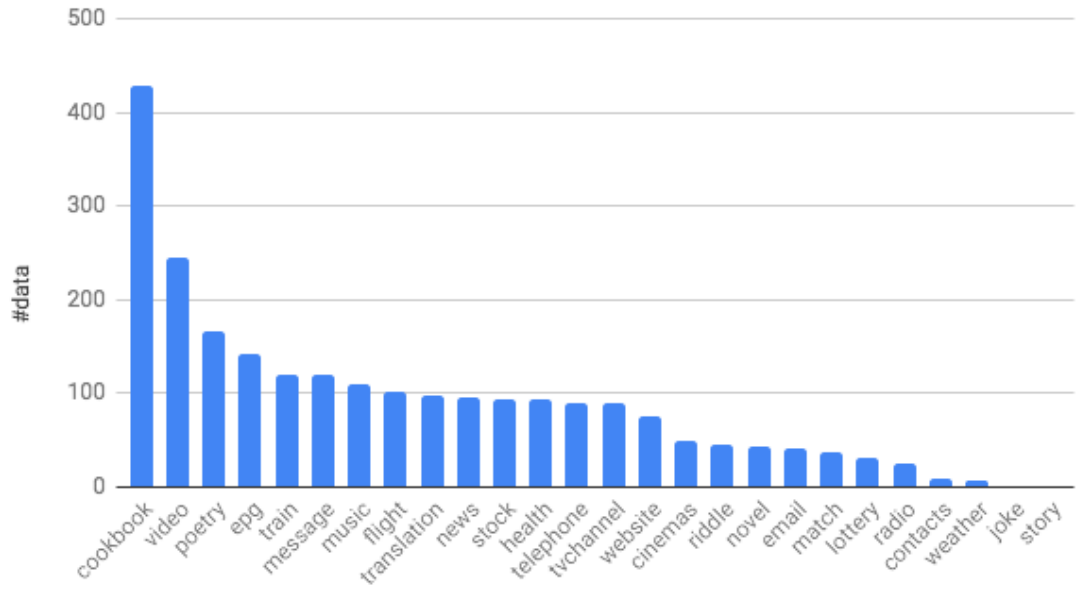

图10. 主办方训练数据集各Domain数据分布

从图10可以看出，相对于SMP2019-ECDT任务1的复杂程度而言，主办方提供的训练数据规模较小并且分布不均衡，例如story垂直领域，仅包含1条训练数据。此外，部分垂直领域的界限模糊，训练数据中存在较严重的噪声，例如contacts垂直领域和telephone垂直领域都包含类似“查下xx的号码”的训练数据。为此，结合我们多年的NLU经验，探索了一些方法用于缓解该问题。

2.9.1 数据增强

为了弥补数据的不足，我们进行了数据增强，数据增强主要包括以下三方面：

1. SMP2019-ECDT任务1的部分语义框架和讯飞AI平台上的某些技能定义类似，所以我们参考了部分讯飞AI平台上的语义框架定义，例如weather、cookbook、message等十余个垂直领域。通过分析讯飞AI平台上的技能，不仅能够清晰的理解该垂直领域的功能，还能从中推理出一些主办方训练集没有覆盖的情况，特别是一些Slot的可能取值范围。

2. 出门问问深耕人机对话领域多年，也积累了相当丰富的垂直领域，综合分析主办方训练集中的语义框架、出门问问语义框架和讯飞AI平台语义框架，我们将出门问问语义框架中的部分数据应用到SMP2019-ECDT任务1中，例如故事类型，新闻类型，笑话类型，电影名称，音乐名称等。

3. 根据前两方面作为指导方向，多名经验丰富的NLU工程师在2579条训

练数据的基础上进行了适当的扩展，人工生成了约5000条高质量训练数据，然后以这批高质量训练数据作为依据，通过语义相似度计算，从出门问问海量用户日志中筛选出50万数据作为训练数据的补充。

2.9.2 交叉验证

数据集不足的情况下，交叉验证是一个很有效的发现模型问题的策略，交叉验证主要包括以下两方面：

1. 在训练模型阶段进行交叉验证，并保留错误实例，通过分析错误实例，可以发现模型存在的多种问题，例如模型预测不符合语义框架定义、垂直领域之间定义模糊等。

2. 在模型提交阶段，我们也借鉴交叉验证的思想，用于消解歧义较大的训练数据，例如将“查xx的号码”这种问法都分到contacts或者telephone准确率怎么样，为此我们提交了多个版本，通过官方评测结果辅助明确相似Domain之间的差异。

2.10 QAP平台

QAP平台(Query Analysis Platform)是出门问问NLU团队针对人机对话系统自主设计研发语义理解平台，该平台具有灵活复用、方便扩展、跨平台、跨场景等特点，通过简单配置即可将出门问问知识图谱应用到自然语言理解任务中，该平台还集成了最大熵、CRF、SVM等传统机器学习模型及CNN、Bi-LSTM、BERT、BERT wwm等深度学习模型，同时能够方便的进行模型交叉验证以及实验分析。在此次SMP2019-ECDT任务1中，该平台大大提高了出门问问NLU参赛队伍的模型迭代效率。

3.实验结果及分析

在SMP2019-ECDT任务1中，我们共提交了50次评测，下面会根据前文提到

的模型从中选择部分有代表性结果进行分析。各版本NLU系统在官方评测集上的表现如表1所示。

表1. 各版本NLU系统在官方评测集上的表现

系统名称	Domain Acc	Intent Acc	Slot F-score	Sentence Acc
ML-NLU	82.73%	75.12%	49.25%	34.71%
ML-KG-NLU	93.06%	85.73%	71.14%	60.75%
BERT-baseline-NLU	90.36%	81.87%	55.61%	43.78%
BERT-BIE-NLU	90.45%	84.09%	69.47%	58.24%
BERT-JOINT-NLU	92.19%	86.60%	70.57%	58.05%
BERT-FRAME-NLU	93.64%	86.79%	74.82%	63.36%
BERT-CLS-KG-NLU	94.12%	87.27%	77.45%	67.31%
BERT-E2E-KG-NLU	97.30%	91.71%	81.28%	72.23%

从表1中ML-NLU和ML-KG-NLU两个系统对比、BERT-FRAME-NLU和BERT-E2E-KG-NLU两个系统对比均可以看出，无论是传统机器学习模型，还是深度学习模型，如果能引入高质量的外界知识，对于系统本身的提升都是非常明显的。从ML-NLU的评测结果可以看出，传统机器学习模型在少量训练数据上的表现欠佳，句子准确率仅有34.71%，在引入出门问问知识图谱之后，ML-KG-NLU系统的准确率得到了显著的提升，因为在训练数据量较少的情况下，传统机器学习模型难以学到足够的特征并且容易产生过拟合的现象，在训练数据上表现良好，但是在开发集以及测试集上表现不足，引入知识图谱之后，知识图谱会为模型提供query中的实体信息以及概念信息，这些信息能够大大降低模型对数据规模的依赖。从BERT-FRAME-NLU和BERT-E2E-KG-NLU两个系统对比也可以看出，在深度学习模型中，如果能引入外界知识，也能显著提升系统效果，主要原因在于BERT预训练模型是一个通用的预训练模型，而每个人机对话系统的NLU都是一个高度定制化的任务，fine-tuning是一种针对定制化任务优化的策略，但是如果能引入定制化任务的外界知识，系统能够提升更多。从BERT-CLS-KG—NLU和BERT-E2E-KG-NLU的对比可以看出，引入用于Domain-Intent分类的知识图谱，不仅Domain-Intent有所提升，Slot F-Score也有提升，一方面原因是此次评测的Slot F Score依赖于Domain-Intent准确率，Domain-Intent准确率提升之后，Slot F-Score也必然会有提升，另一方面原因，

Domain-Intent分类和Slot标注同时fine-tuning BERT预训练模型，引入Domain-Intent分类知识图谱后，对BERT fine-tuning也有所影响，进而间接提升了Slot标注效果。

从表1中ML-NLU和BERT-baseline-NLU两个系统对比可以看出，在同样少量数据的情况下，BERT预训练模型的效果要明显优于传统机器学习模型，因为BERT预训练模型是在大量语料的基础上训练得出的，并且BERT模型架构本身也是预训练模型中比较合理的架构，所以BERT预训练模型中已经隐式的包含了一部分世界知识。

从表1中BERT-baseline-NLU、BERT-BIE-NLU、BERT-JOINT-NLU和BERT-FRAME-NLU四组系统对比可以看出，以下几项改进都是有效的：

- 在Slot标注模型中添加BIE标记限制；
- 联合预测Domain-Intent标记，减少不符语义框架的Intent；
- 通过语义框架后处理，减少不符语义框架的Slot；

此外，数据增强和交叉验证是在系统迭代的过程中逐步进行的，后期系统的提升，除了模型和方法，数据增强和交叉验证也起到了关键作用。

最终版本BERT-E2E-KG-NLU在官方测试集上表现为Domain Acc=97.30%，Intent Acc=91.71%，Slot F-Score=81.28%，Sentence Acc=72.23%，取得了第一名的成绩，SMP2019-ECDT任务1官方评测结果前15名如表2所示。

表2. SMP2019-ECDT任务1 官方评测结果前15名

系统名称	机构名称	Domain Acc	Intent Acc	Slot F-score	Sent. Acc
MobvoiNLU	出门问问信息科技有限公司	97.30%	91.71%	81.28%	72.23%
decodeNL	北京沃丰时代数据科技有限公司	94.50%	87.17%	75.97%	65.19%
coffeeNLU	南京大学Websoft实验室	92.77%	86.60%	74.41%	62.58%
FutureMiracle	深思考人工智能机器人科技（北京）有限公司	92.86%	85.05%	71.17%	59.88%
nlu_sys	cvte中央研究院	91.42%	83.80%	73.57%	59.40%
baseline	华南理工大学-CIKE实验室	89.20%	82.64%	68.40%	58.05%
mi&tNLU	哈尔滨工业大学机器智能与翻译	92.29%	81.97%	69.35%	57.76%
dev65	南京大学	90.74%	82.26%	70.90%	57.18%
scauNLU	华南农业大学口语对话系统研究室	96.62%	88.52%	70.25%	57.18%
baseline1	欧文科技有限公司	92.67%	81.77%	69.09%	54.96%
fuxiNLU	网易伏羲实验室	89.48%	80.04%	65.57%	53.04%
baseline_1	哈尔滨工业大学网络智能研究室	90.26%	80.52%	65.60%	52.16%
V3.2_eh	北京来也网络科技有限公司	90.84%	83.12%	62.92%	51.01%
tra_baseline	杭州华卓科技	89.01%	79.17%	60.84%	49.18%
NLU_MODEL	台達電子知識管理部門	87.85%	79.46%	62.09%	47.93%

4.总结

在SMP2019-ECDT任务1中，凭借出门问问NLU团队自主研发的QAP平台快速迭代，我们尝试了传多种模型和方法，包括BERT预训练模型、出门问问知识图谱、数据增强、语义框架检测和传统机器学习模型等，重点对比了基于传统机器学习模型的自然语言理解系统和基于BERT预训练模型的自然语言理解系统，引入知识图谱前后的基于传统机器学习模型的自然语言理解系统和基于BERT预训练模型的自然语言理解系统，并验证了在基于BERT的端到端自然语言理解系统中，添加BIE限制标记、联合预测Domain-Intent分类以及语义框架后处理的有效性。