基于 BERT 联合训练的自然语言理解评测报告

作者：侯晋峰 李伟 高亨德

北京沃丰时代数据科技有限公司

摘要：人机对话技术可以使人通过自然语言与计算机进行交互，是学术界和产业界需要攻克的难点之一。其中，由于任务型人机对话技术的重要实用价值及应用前景，受到更多的关注。本文针对任务型对话系统中的自然语言理解模块的领域分类、意图识别和语义槽填充三项任务，使用基于 BERT 的联合训练模型，对三个任务做预测，在 SMP2019 的自然语言理解评测中取得第二名的成绩。

关键词：自然语言理解，BERT，联合训练，SMP2019

1.引言

近年来，越来越多的用户通过任务型对话系统来获得便捷高效的服务。任务型对话系统是指以人机对话的形式在特定条件下提供信息或服务的系统。通常情况下是为了满足带有明确目的的用户。例如查天气、点播电影、订餐、订机票火车票等任务型场景。鉴于其广泛的应用前景，在 SMP2019 大会上，哈尔滨工业大学与科大讯飞股份有限公司联合组织并承办了中文人机会话技术评测(ECDT)，为任务型会话提供了一个具有指导意义的评测任务。

自然语言理解任务主要包括下面三个子任务：领域分类、意图识别和语义槽填充。其中，领域分类任务是把语句划分到指定的不同领域标签内[1]。例如：给定一句用户的指示，“我想听周杰伦的菊花台”需要将其划分到领域“music”下，进而根据领域“music”有针对性的对指示给出响应。与领域分类任务类似，意图识别任务是把语句划分到不同的意图标签内。而语义槽填充则是需要识别出用户指令语句中的实体部分，并进行标注。例如，上例中的意图为“PLAY”，语义槽为“artist:周杰伦”，“song:菊花台”。特别的，本次评测任务中，意图识别任务、语义槽填充任务都与领域分类任务有关联。例如：判定领域为“music”的，意图

一定为“PLAY”；在填充语义槽时，同样是始发地与终点，判定领域为“bus”时，其语义槽为“Src”与“Dest”而判定领域为“flight”时，其语义槽为“startLoc_city”，“endLoc_city”。

针对这三个任务，传统模型面临特征选择困难，数据量大时计算效率低等问题， BERT 具有很好的语义表达的能力，可以充分捕捉到句子的上下文信息，考虑到这三个任务之间具有一定的关联性，本文提出一种基于 BERT 的联合训练模型和联合训练的方法，在三个任务中取得了第二名的成绩。

2.模型及方法介绍

本节主要介绍 BERT 基础模型处理相关任务及本文所使用的基于 BERT 的两种联合训练模型。

2.1BERT 基础模型

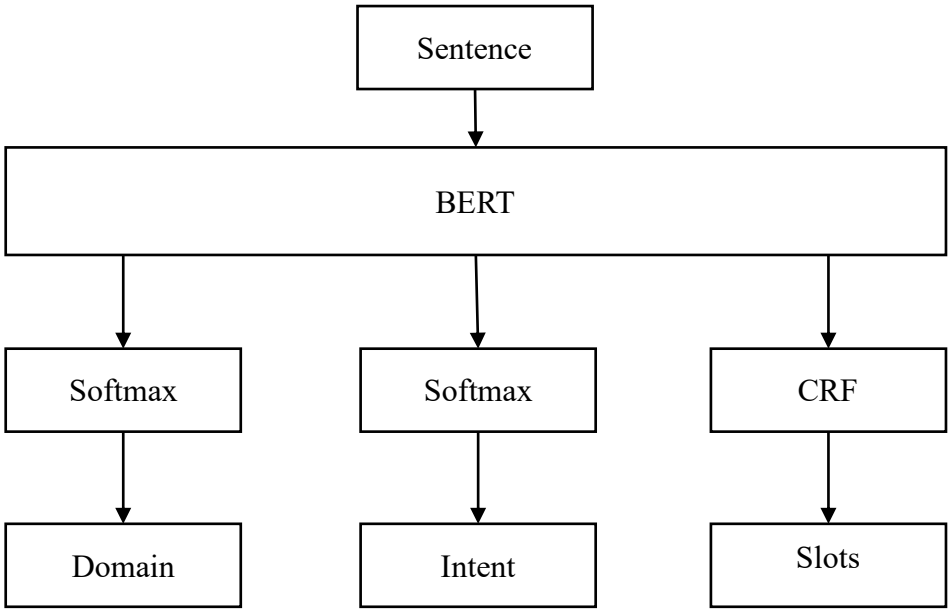
Bidirectional Encoder Representations from Transformers（BERT），是谷歌于 2019 年底提出的一种基于 Transformer 模型的双向编码器[1]。与 Generative Pre-Training(GPT)[2]两阶段训练法类似，BERT 也采用语言模型进行预训练作为第一阶段，第二阶段在下游任务进行微调，其在多种自然语言(NLP) 处理任务上取得了最佳成绩。

BERT 中的特征抽取器完全使用 Transformer 模型[3]，与传统的循环神经网络 (RNN)和卷积神经网络(CNN)作为 encoder-decoder 的其他大多数模型不同，在中文中使用字级别的特征，并在建模时每个字都与句中的其他字建立联系，故而可以结合上下文中较远的关键信息。双向 Transfromer 深度结构的使用极大增强了模型的语言表征能力。

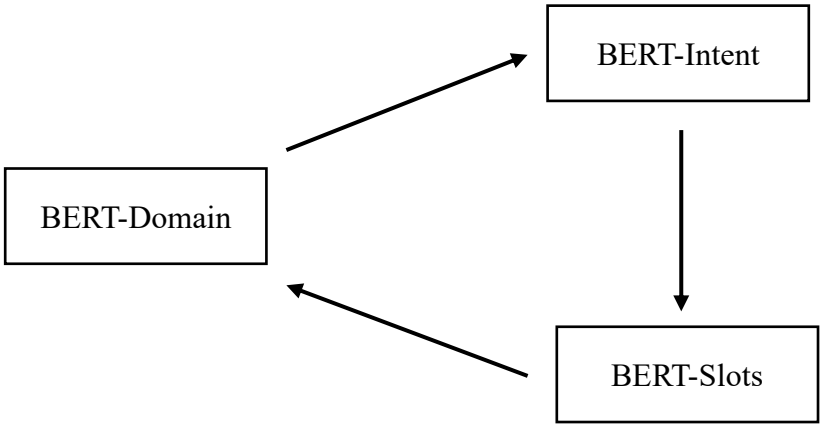
2.2 基于 BERT 联合训练模型 I

为了融合多个任务的信息，我们采用多任务共享模型参数的方式训练多个任务的模型，即多个任务的模型分别训练，但是训练一个模型时以其他模型的训练结果进行模型的初始化，模型结构及训练流程如图 1 所示，三个模型共享 BERT

的参数，采用分别单独循环训练的方式训练模型，即先训练一轮 Domain，再在 Domain 的 BERT 参数基础上训练 Intent，然后在 Intent 的 BERT 参数基础上训练 Slots，一轮之后重新在 Slots 的 BERT 参数基础上训练 Domain；如此往复进行训练，直到各个模型都达到最优。



a) 模型结构



b) 训练流程

图 1 联合训练模型 I

2.3 基于 BERT 联合训练模型 II

模型 II 与模型 I 在代码结构上基本一样，只是训练方式不同，模型 II 采用多个模型共同训练的的方式进行训练，即将多个模型的输出 loss 进行加权求和，作为联合模型的 loss 进行训练，模型的参数同时更新，如图 2 所示。

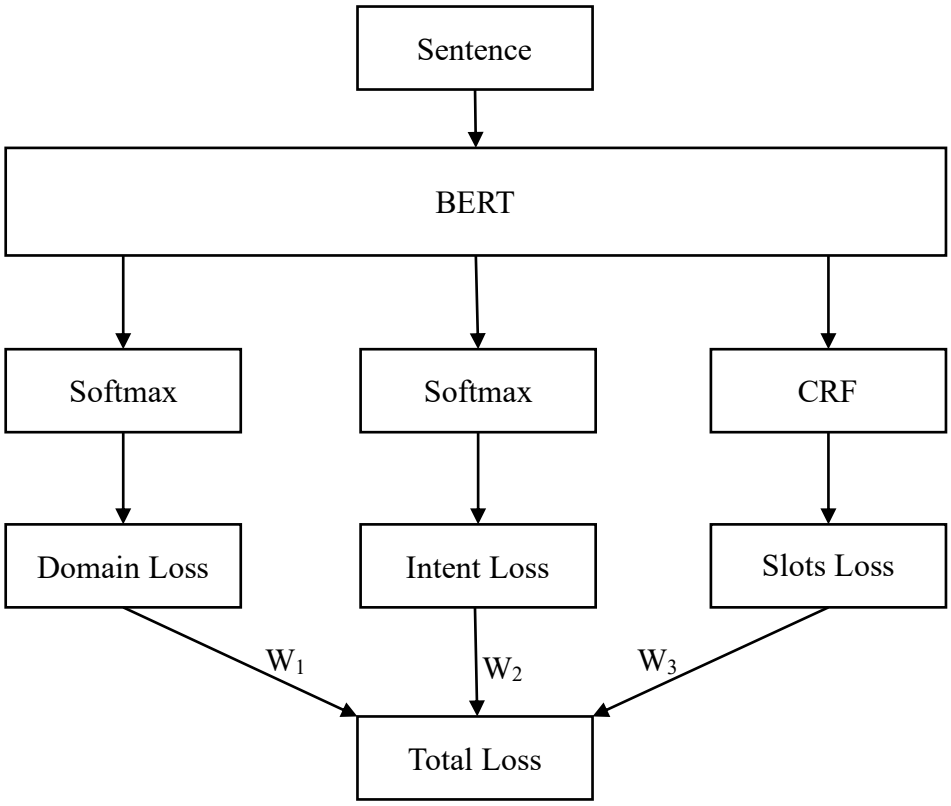


图 2 联合训练模型 II

2.4 加入额外特征的 BERT 模型

BERT 本身的特征提取的能力和语义表达能力已经非常优秀，但是加入一些额外的特征依然会有一定的提升效果，我们在 BERT 的输出层之后，又加了一个额外的特征层，与 BERT 的输出层结果 concat 起来作为下一层的输入，使模型可以获取到更加丰富的特征。这儿我们加入的特征为关键词特征，采用 onehot 的

方式接入。

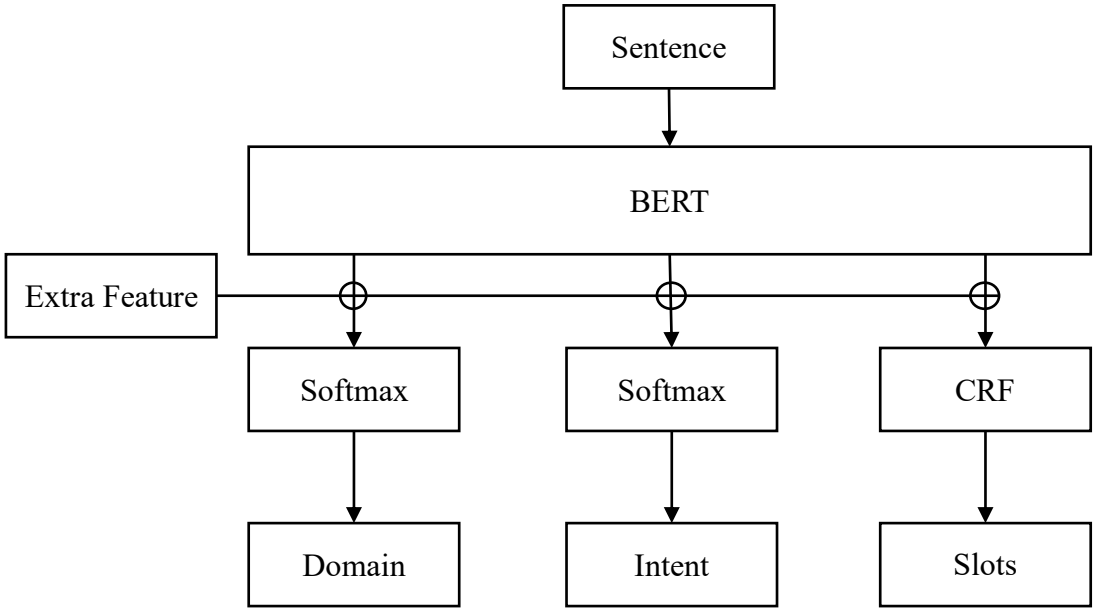


图 3 加入额外特征的 BERT 模型

3.实验结果及分析

3.1 实验数据

SMP2019 自然语言理解评测的数据总体情况如表 1 所示。领域包含“app”、“websit”、“health”等 29 类，意图包括“LAUNCH”、“QUERY”，“PLAY”等 24 类，语义槽包括“artist”、“category”、“datetime”等 60 类。

表 1 训练集语料各任务数量统计

领域种类	意图种类	语义槽种类
29	24	60

特别的，主办方提供数据面临严重的数据不平衡问题，最多的领域“cookbook”有 428 条，而最少的“story”与“joke”只有 1 条。故而需要对数据进行平衡化处理，对数据量过多的删除部分句子，而过少的数据进行了扩充。

本文采用两种方法对文本进行扩充：

1、将训练集语料输入百度知道，获取相似问，这样可以获取到同一问题的不同问法。

2、根据组委会提供的 29 个 domain，人工进行造句，对数据量少的 domain 进行了数据补充。为了增强模型的鲁棒性，特意造了一些易混淆 domain 的例子。例如，在领域“music”下造句“我想听首英文歌。”这时不应该把 domain 根据关键词“英文”而归到领域“translation”下。

补充数据前后的训练集数量表 2 所示：

表 2 扩充后训练集数量统计

原始训练集	扩充百度知道	扩充自主造句
2579	8892	9889

3.2 评价指标

对于领域分类、意图识别，我们采用准确率（acc）来评价，对于语义槽填充，我们采用 F 值来评价，为了综合考虑模型的能力，我们最终采用句准确率（sentence acc）来衡量一句话领域分类、意图识别和语义槽填充的综合能力。 以下所有测试结果数据均为在我们自行构造的测试集上的结果，并非竞赛结果。

3.3 实验结果与分析

由于 BERT 单模型在 NLP 的多项任务上超越其他深度模型，包括本次评测的文本分类及实体识别。故而本文未使用其他模型作为比较，直接对 BERT 模型进行优化。首先，数据对结果的影响是最大的，影响如表 3 所示：

表 3 BERT 原始模型加入数据效果比较

数据集	Domain 准确率	Intent 准确率	Slots F 值
原始训练集	0.7428	0.6941	0.3845
扩充百度知道	0.8574	0.8109	0.4621
扩展自主造句	0.9023	0.8212	0.5425
自主+百度知道	**0.9140**	**0.8337**	**0.6824**

可见，扩展了百度知道后数据有部分提升，但效果并不是很明显，通过自主造句扩展之后，Domain，intent 的准确率和 slots 的 F 值有大幅的上涨，达到了最优的效果。因为 BERT 对文本的表示能力很强，在加入多个类似文本并不能提升

算法的效果。

采用联合训练模型的结果如表 4 所示，联合模型 I 训练 2 个循环之后达到最优；联合模型 II 在 Slots 的效果表现上较差；联合模型 I+II 的方式为先用模型 II 训练基础模型，然后在模型 II 的 BERT 参数基础上，再采用模型 I 的训练方式分别对模型进行优化，最终的效果达到最优

表 4 联合训练模型效果比较

模型	Domain 准确率	Intent 准确率	Slots F 值
BERT 基础模型	0.9140	0.8337	0.6824
联合模型 I	0.9228	0.8497	0.6910
联合模型 II	0.9210	0.8501	0.6774
联合模型 I+II	**0.9274**	**0.8541**	**0.7104**

加入词特征对模型的影响如表 5 所示。

表 5：词特征对模型的影响

模型	Domain 准确率	Intent 准确率	Slot F 值
联合模型 I+II	0.9274	0.8541	0.7104
加入词特征	**0.9288**	**0.8557**	**0.7345**

然后本文分别对三个任务单独抽取关键词表，并对结果进行约束，例如，出现“红烧”、“清蒸”等关键词则约束其为“cookbook”，出现“歌曲”约束其为“music”，出现“到北京”约束其为“end_city”等。并加入规则对其进行约束，例如“播放一首英文歌曲”，同时出现两个及以上的关键词则取最后一位的关键词所在 domain。加入关键词表及规则约束后的最终的结果如表 6 所示：

表 6 加入关键词及规则效果比较

数据集	Domain 准确率	Intent 准确率	Slots F 值
未加入	0.9288	0.8557	0.7345
加入关键词	0.9374	0.8639	0.7421
加入规则	0.9340	0.8607	0.7364
规则与关键词	**0.9450**	**0.8717**	**0.7597**

4.总结

本文介绍了本次参加 SMP2019 中文人机对话技术评测(ECDT)中的自然语言理解任务的技术方案和模型分析。本文采用的基于 BERT 的联合训练模型在该评测任务的综合指标下取得第二名。

由于比赛时间有限，我们重点针对 Domain 任务作了数据扩充、优化及平衡性处理，并未对全部任务做处理。其次，对数据的分析不够深入，在人工标注数据进行扩充的时候会有一些疑惑不解的地方。再次，针对测试集中任务的 badcase 分析深度不够，且未找到特别有效的解决办法，最后，本文只使用了 BERT 的单模型，并未根据其他模型的特点，来融合其他模型进行预测。在未来的工作中需要仔细分析 badcase，并尝试模型融合。

5.参考文献

[1] Tur G, Deng L, Hakkani-Tür D, et al. Towards deeper understanding: Deep convex networks for semantic utterance classification[C]// Proceedings of the 37th IEEE International Conference on Acoustics, Speech,

[2] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv: Computation and Language, 2018.

[3] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding

by generative pre-training[J], 2018.

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need[J]. neural information processing systems, 2017: 5998-6008.