项目描述

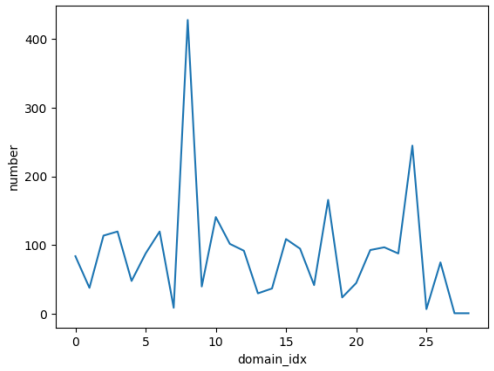
smp2019自然语言理解比赛主要包括三个任务：领域识别、意图识别和槽位填充
评测的指标包括三个：领域识别准确率、意图识别准确率、槽位填充F1值、以及整体准确率（领域、意图、槽位的key-value都正确）

数据集

数据格式：是一个包含若干个字典的列表， 每个字典是一个样例， 主要包括'text'、“domain”、“intent”、“slots”4个字段
[{
"text": "请帮我打开uc",
"domain": "app",
"intent": "LAUNCH",
"slots": {
"name": "uc" }
}
]
下面的表格是对数据集的一些标签类别以及分布的统计信息

domain	intent	slot	max_len_text	num_instance
29	24	125	27	2579

标签分布图

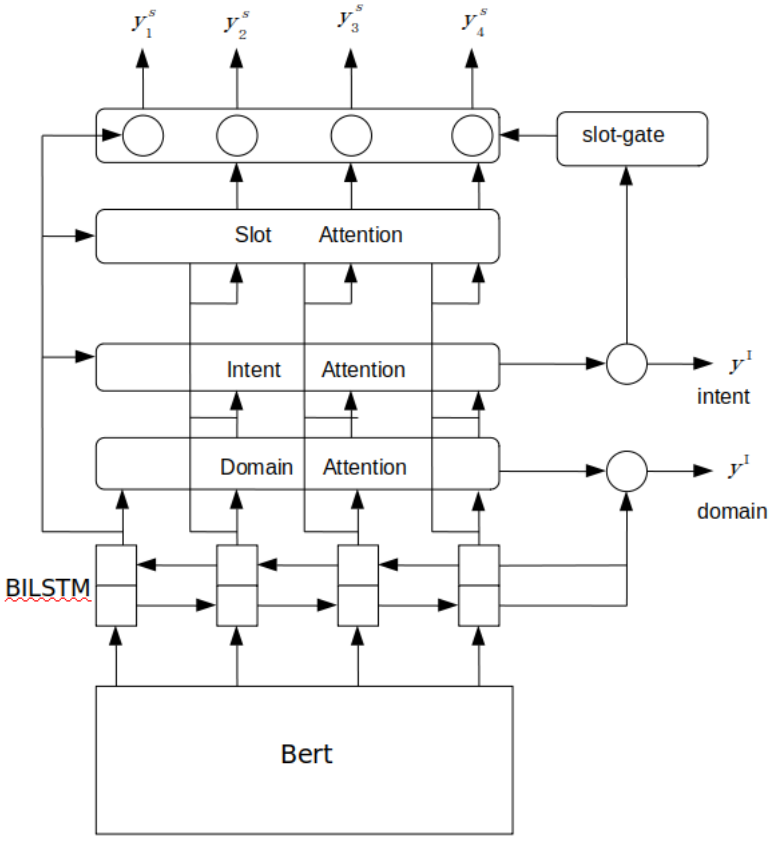


doma

我们可以看到标签的分布是很不平衡的，可以考虑使用一些处理标签不平衡的方法，例如下采样、数据增强、加权损失函数的方式

模型

主要是基于槽位门机制以及attention机制的bert模型，模型的结构如下：



领域识别：不妨设bert模型的输出状态为x_1, x_2, ..., x_{n-1}, x_n
双向lstm的输出结果为h_1, h_2, ..., h_{n-1}, h_n
计算slot attention：

$$c_i^S = \sum_{j=1}^{T} \alpha_{ij}^S h_j$$

$$\alpha_{ij}^D = \frac{exp(e_{ij})}{\sum_{k=1}^{T} exp(e_{i,k})}$$

$$e_{ij} = \sigma(W_{he}^S h_k)$$

$$y_i^S = softmax(W_{hy}^S(h_i + c_i^S))$$

计算domain attention和intent attention：

$$y^I = softmax(W_{hy}^I(h_T + c^I))$$

加入槽位门机制：

$$g = \sum v \cdot tanh(c_i^S + W \cdot c^I)$$

$$y_i^S = softmax(W_{hy}^S(h_i + c_i^S \cdot g))$$

实验结果

本文使用的超参数

batch_size	max_epochs	max_seq_len	learning_rate
5	10	50	2e-5

实验结果

	domain_acc	intent_acc	slot_f1	frame_acc
bert+bisltm+crf	0.9229	0.8197	0.6935	0.5776