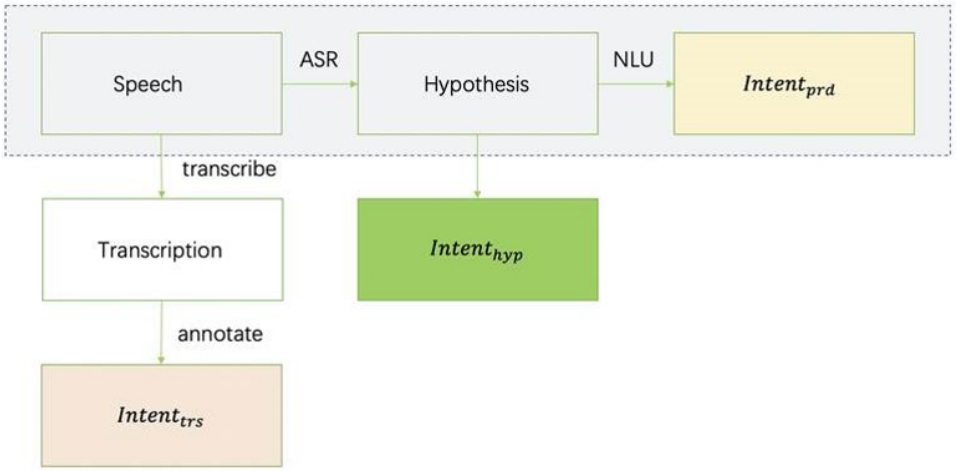
大件外呼N-Best实验

背景

在外呼的意图识别中，基本采用的是Speech——Hypothesis——Intent_prd这样的架构。在此架构下，训练数据主要来自于Hypothesis和Intent_hyp，预测的时候，对比意图也是Intent_prd和Intent_hyp。在这样的评测下，模型已经能够达到较高的意图准确率，在评测端到端的真实效果时（即对比Intent_prd和Intent_trs），模型的效果仍然不够理想。这是因为模型的意图识别鲁棒性不高，NLU拿到的是ASR top1结果进行意图识别，如果top1结果不准确，很容易造成意图错误。为了提升意图识别对ASR的容错，也即提高端到端的真实效果，我们实验了端到端的意图识别模型。



计划实验内容list：

1. 训练集意图标签的作用
2. 训练模型的作用
3. 训练集（不同时期的日志）的影响
4. ASR返回N-Best的作用
5. N-Best中N的影响效果
6. 不同BERT预训练模型的影响
7. N-Best中加入句子特征的作用：LM(语言模型)，AM(声学模型)得分等
8. N-Best中加入词特征的作用：词得分，词音素，词时长等
9. ASR去噪声对N-Best端到端意图的影响
10. N-Best拼接方式调优，如对文本的embedding进行maxpooling等

序号	算法（单轮模型）	训练集数量来源	训练集特征	训练集标签	测试集（端到端意图）		
					固定种子3次中间值（BERT）	去除纯特殊标记（背景音等）	固定种子3次平均值（BERT）
1	LR	训练集1（8.8k线上日志），ASR识别的Hypothesis	单条文本tf-idf向量	Intent_hyp	84.72%	87.87%	
2.1	BERT_base	训练集1（8.8k线上日志），ASR识别的Hypothesis	单条文本bert向量	Intent_hyp	86.13%	88.98%	86.10%
2.2	BERT_(商城+物流)	训练集1（8.8k线上日志），ASR识别的Hypothesis	单条文本bert向量	Intent_hyp	86.66%	89.54%	86.62%
3.1	LR	训练集2（6.4k线上日志），ASR识别的Hypothesis	单条文本tf-idf向量	Intent_hyp	84.38%	87.17%	
3.2	LR	训练集2（6.4k线上日志），ASR识别的Hypothesis	单条文本tf-idf向量	Intent_trs	83.31%	86.12%	
5.1	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	单条文本bert向量	Intent_hyp	87.00%	89.82%	87.02%
5.2	BERT_base	训练集2（6.4k线上日志），ASR识别的Hypothesis	单条文本bert向量	Intent_trs	85.59%	88.42%	85.66%
5.3	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	单条文本bert向量	Intent_trs	87.13%	89.54%	87.09%
5.4	BERT_(商城+物流)	训练集2（8.8k+6.4k线上日志），ASR识别的Hypothesis	单条文本bert向量	Intent_hyp	87.53%	90.24%	87.6%
6.1	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	10条拼接文本bert向量	Intent_hyp	87.40%	90.03%	87.40%
6.2	BERT_base	训练集2（6.4k线上日志），ASR识别的Hypothesis	10条拼接文本bert向量	Intent_trs	87.47%	89.68%	87.53%

6.3	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	10条拼接文本bert向量	Intent_trs	88.34%	90.66%	88.36%
6.4	BERT_(商城+物流)	训练集2（6.2k线上日志，6.4k去除杂质），ASR识别的Hypothesis	10条拼接文本bert向量	Intent_trs	88.47%	90.86%	88.67%
7.1	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	10条拼接文本bert向量及其LM，AM得分(将得分进行分桶表示)	Intent_trs	88.81%	90.86%	88.81%
7.2	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	10条拼接文本bert向量及其LM，AM得分(用归一化得分进行句子向量加权)	Intent_trs	88.61%	90.66%	88.70%

注：

（1）随机种子和固定种子是指训练集在进行打乱时的操作，随机种子表示随机打乱训练集；固定种子表示训练集打乱顺序相同，3次实验是由于模型内部还有一些随机扰动

（2）N-Best中的杂质表示N条数据中没有一条和transcription意图相同

数据说明

- 测试集：10.24+10.26+10.27标注数据，共1492条（取这三天是因为音频留存率较高，更符合真实线上数据）
- 训练集1（8.8k线上日志）：线上日志标注ASR结果的单句意图，共8876条，这部分数据属于之前积累，没有对应音频文件，即没有N-Best结果和Intent_trs结果
- 训练集2（6.4k线上日志）：09.07—09.11+10.25+10.28+10.29+10.30+11.05标注数据，共6404条

日期	消息量(mid)	有音频的消息量	音频留存率	未标注消息量占比
10.24	508	485	95.47%	4.95%
10.25	557	489	87.79%	3.68%
10.26	541	492	90.94%	3.05%
10.27	539	515	95.55%	3.69%
10.28	527	341	64.71%	0
10.29	534	354	66.29%	1.13%
10.30	476	269	56.51%	3.72%
11.05	523	305	58.32%	4.59%

注：

数据来源于每天随机抽取160通日志进行了人工转写和ASR N-Best输出

音频留存率=有音频的消息量/总消息量

未标注表示用户确实没说话或背景音太嘈杂，标注不出来用户说的话

一、训练集意图标签的作用

实验目的：相同数据下，用不同的意图标签进行模型训练，评判意图标签对结果的影响

实验结论：直接使用Intent_trs会带来负向效果

实验分析：由于ASR识别的Hypothesis和Intent_trs并不统一，比如用户说的是“对”，转写后Intent_trs是“confirm”，而ASR的top1结果是“喂”，其本身意图是“deafness”，在实验4下，会将“喂”—"confirm"作为训练数据，引入了噪音数据，造成准确率的负向效果

序号	算法（单轮模型）	训练集数量来源	训练集特征	训练集标签	测试集（端到端意图）	去除纯特殊标记（背景音等）
3	LR	训练集2（6.4k线上日志），ASR识别的Hypothesis	单条文本tf-idf向量	Intent_hyp	84.38%	87.17%
4	LR	训练集2（6.4k线上日志），ASR识别的Hypothesis	单条文本tf-idf向量	Intent_trs	83.31%	86.12%

二、训练模型的作用

实验目的：相同数据下，用不同的模型训练，评判模型对结果的影响

实验结论：相同数据情况下，BERT模型相比LR模型效果更好

实验分析：BERT预训练模型泛化性能更强

序号	算法（单轮模型）	训练集数量来源	训练集特征	训练集标签	测试集（端到端意图）	去除纯特殊标记（背景音等）
1	LR	训练集1（8.8k线上日志），ASR识别的Hypothesis	单条文本tf-idf向量	Intent_hyp	84.72%	87.87%
2.1	BERT	训练集1（8.8k线上日志），ASR识别的Hypothesis	单条文本bert向量	Intent_hyp	86.13%	88.98%

4	LR	训练集2（6.4k线上日志），ASR识别的Hypothesis	单条文本tf-idf向量	Intent_trs	83.31%	86.12%
5.2	BERT	训练集2（6.4k线上日志），ASR识别的Hypothesis	单条文本bert向量	Intent_trs	85.59%	88.42%

三、训练集（不同时期的日志）的影响

实验目的：相同模型和标签的情况下，用不同时期的日志数据作为训练集，评判是否出现概念漂移的现象

实验结论：用不同时期的日志作为训练集，对端到端的意图影响不大

实验分析：在数据量够大和标注标准统一的情况下，不同时期的训练集带来的效果类似

序号	算法（单轮模型）	训练集数量来源	训练集特征	训练集标签	测试集（端到端意图）	去除纯特殊标记（背景音等）
1	LR	训练集1（8.8k线上日志），ASR识别的Hypothesis	单条文本tf-idf向量	Intent_hyp	84.72%	87.87%
3	LR	训练集2（6.4k线上日志），ASR识别的Hypothesis	单条文本tf-idf向量	Intent_hyp	84.38%	87.17%

四、ASR返回N-Best的作用

实验目的：相同数据标签下，用ASR返回的top1和top10，评判N-Best的影响

实验结论：ASR N-Best对意图鲁棒性会带来正向提升

实验分析：外呼的ASR top1数据可能不准确，采用N-Best数据拼接多条结果，更能包容ASR top1数据的错误结果，鲁棒性更强

序号	算法（单轮模型）	训练集数量来源	训练集特征	训练集标签	测试集（端到端意图）		
					固定种子3次中间值（BERT）	去除纯特殊标记（背景音等）	固定种子3次平均值（BERT）
5.1	BERT_base	训练集2（6.4k线上日志），ASR识别的Hypothesis	单条文本bert向量	Intent_hyp	87.00%	89.82%	87.02%
5.2	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	单条文本bert向量	Intent_trs	85.59%	88.42%	85.66%
5.3	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	单条文本bert向量	Intent_trs	87.13%	89.54%	87.09%
6.1	BERT_base	训练集2（6.4k线上日志），ASR识别的Hypothesis	10条拼接文本bert向量	Intent_hyp	87.40%	90.03%	87.40%
6.2	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	10条拼接文本bert向量	Intent_trs	87.47%	89.68%	87.53%
6.3	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	10条拼接文本bert向量	Intent_trs	88.34%	90.66%	88.36%

用“[SEP]”拼接ASR top10结果，拼接后文本长度（去除[SEP]）限制在100以内（95.61%的数据拼接后文本长度低于100），如实验方法：

1. 这样吧啊过几天都那个我那个我那个准备好了没直接我安装是吗[SEP]这样吧啊过几天都那个我那个我那个准备好了没了直接我安装是吗[SEP]这样吧啊过几天都那个我那个我那个准备好了没了直接给我安装是吗
2. 啊对[SEP]嗯对[SEP]呃对[SEP]昂对[SEP]哎对

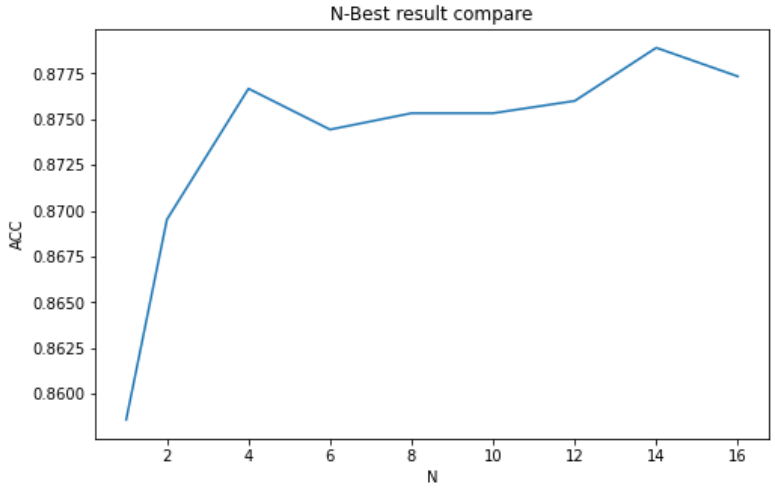
五、N-Best中N的影响效果

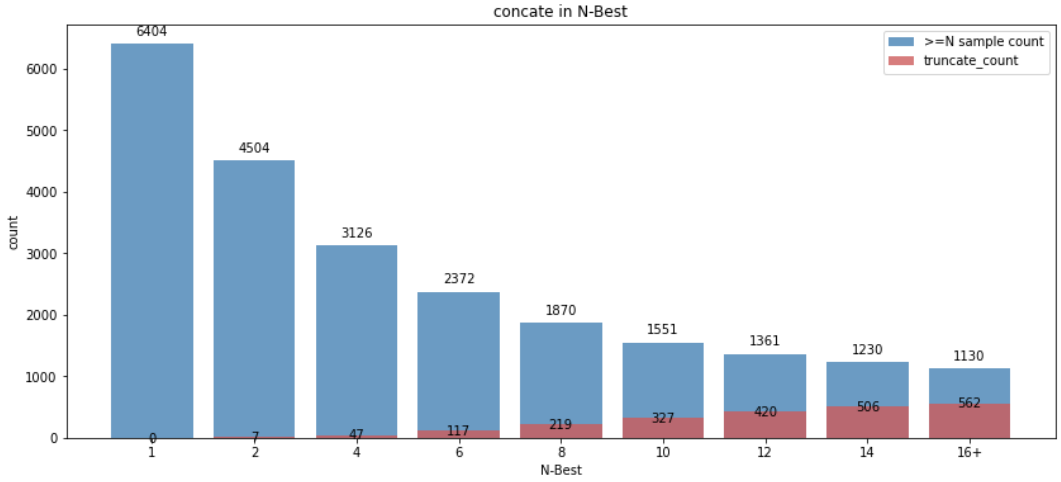
实验目的：评测不同的N对意图的影响

实验结论：N取14的准确率最好，但N取4更符合上线需求（效果和N=14差别不大，但是预测速度会提升）

实验分析：N太小获取不到更丰富的特征（N=1，N=2）,N=4达到第一个峰值，往后N增加引入了杂质，导致准确率下降，然后到N=14达到最大值

试验方法：固定数据输入顺序，即固定训练集在进行打乱时的随机种子，从N=2取到N=16，由于模型内部有随机扰动，取3次试验的平均值





注：左边图表示ACC随N的变化曲线；右边图表示达到ASR输出结果中达到N的数据量，红色部分表示被100的长度限制所截断的数据量

具体数据：

N	第一次实验	第二次实验	第三次实验	均值
2	86.93%	86.46%	87.47%	86.95%
4	87.60%	88.20%	87.20%	87.67%
6	87.73%	86.93%	87.67%	87.44%
8	87.87%	88.14%	86.60%	87.53%
10	87.87%	87.47%	87.26%	87.53%
12	87.73%	87.13%	87.93%	87.60%
14	87.80%	88.14%	87.74%	87.89%
16	87.53%	87.73%	87.94%	87.73%

六、不同BERT预训练模型的影响

实验目的：实验外呼预训练模型，证实了外呼领域、商城领域的BERT预训练模型在外呼意图识别上能够带来提升，还需要进一步证实在意图鲁棒性N-Best实验上，预训练是否会带来提升

实验结论：采用商城+物流领域数据训练的BERT模型在N-Best上会给外呼意图识别带来0.87%的提升

实验分析：相比于基础的bert预训练模型，商城+物流数据训练的模型更符合业务场景

试验方法：替换基础的bert预训练模型，在商城+物流的预训练模型上进行模型的fine-tuning

序号	算法（单轮模型）	训练集数量来源	训练集特征	训练集标签	测试集（端到端意图）—固定种子3次中间值	去除纯特殊标记（背景音等）	固定种子3次平均值
2.1	BERT_base	训练集1（8.8k线上日志），ASR识别的Hypothesis	单条文本bert向量	Intent_hyp	86.13%	88.98%	86.10%

2.2	BERT_(商城+物流)	训练集1（8.8k线上日志），ASR识别的Hypothesis	单条文本bert向量	Intent_hyp	86.66%	89.54%	86.62%
5.2	BERT_base	训练集2（6.4k线上日志），ASR识别的Hypothesis	单条文本bert向量	Intent_trs	85.59%	88.42%	85.66%
5.3	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	单条文本bert向量	Intent_trs	87.13%	89.54%	87.09%
6.2	BERT_base	训练集2（6.4k线上日志），ASR识别的Hypothesis	10条拼接文本bert向量	Intent_trs	87.47%	89.68%	87.53%
6.3	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	10条拼接文本bert向量	Intent_trs	88.34%	90.66%	88.36%

同样，对实验6.3进行了不同N的实验，结果如下：

N	第一次实验	第二次实验	第三次实验	均值
2	88.27%	88.07%	87.80%	88.05%
4	88.61%	88.34%	88.34%	88.43%
6	88.74%	88.20%	88.00%	88.32%
8	88.34%	88.87%	88.94%	88.72%
10	88.34%	88.94%	87.80%	88.36%
12	87.87%	88.67%	88.34%	88.29%
14	88.07%	87.94%	88.20%	88.07%
16	88.61%	88.40%	88.61%	88.54%

七、N-Best中加入句子特征的作用

实验目的： 相同数据下，都用ASRtop10，引入更多的句子维度特征，评判句子维度特征对N-Best的影响

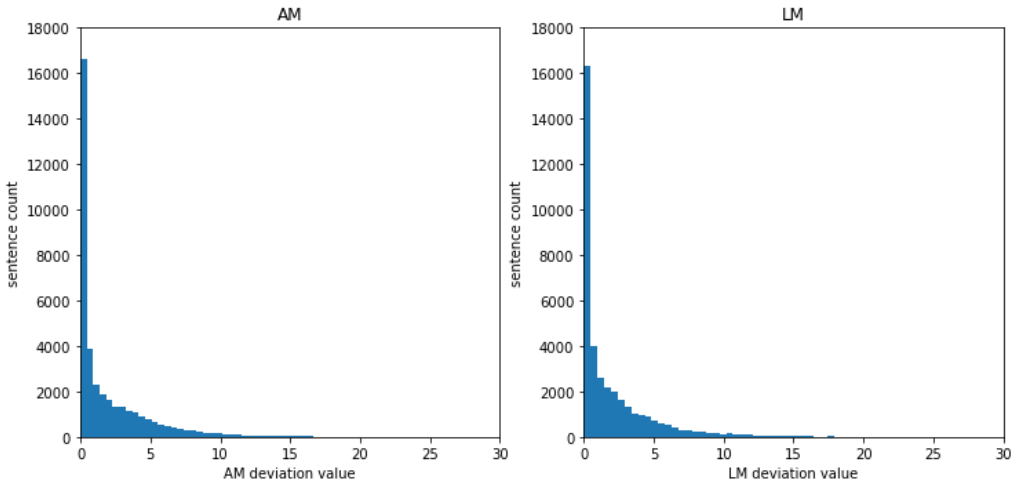
（1）方案1：将AM、LM以分桶方式表示，作为输入特征

实验结论：以分桶方式表示每句话的AM、LM得分，并作为输入进行交互，可以提升意图识别准确率0.45%

实验分析：每句话的AM、LM得分一定程度上可以表示这句话的可信程度，而每句话是否可信对最终意图有决定性的影响

实验方法：

- 分桶方式：以top1的得分为基准，计算topN的AM、LM得分与top1的差值，以差值进行分桶，差值越大表示后面句子与第一句相差越远。
- 桶的区分：在对差值进行统计后，发现大部分是差值小于10，因此将桶的区分设置为：<1，1-3，3-5，5-7，7-10，10-20，20-30，>30



序号	算法（单轮模型）	训练集数量来源	训练集特征	训练集标签	测试集（端到端意图）—固定种子3次中间值	去除纯特殊标记（背景音等）	固定种子3次平均值
6.3	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	10条拼接文本bert向量	Intent_trs	88.34%	90.66%	88.36%
7.1	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	10条拼接文本bert向量及其LM，AM得分(将得分进行分桶表示)	Intent_trs	88.81%	90.86%	88.81%

实验7.1对于不同的N效果如下：

N	第一次实验	第二次实验	第三次实验	均值

2	88.00%	87.53%	87.27%	87.60%
4	88.74%	88.54%	88.14%	88.47%
6	88.34%	88.07%	88.47%	88.29%
8	88.61%	88.81%	88.74%	88.72%
10	88.81%	88.94%	88.67%	88.81%
12	88.40%	88.40%	88.40%	88.40%
14	88.67%	88.74%	88.40%	88.61%
16	89.21%	88.27%	88.94%	88.81%

（2）方案2：用AM、LM归一化得分给每个句子的表征向量加权

实验结论：以AM、LM归一化得分给每个句子的表征向量加权，可以提升意图准确率0.34%

实验分析：相对于单纯的N-Best，将每个句子的表征向量进行加权可以获取不同句子的更多信息

实验方法：

- 归一化得分：以LM为例，真实LM绝对值数值越小，表示句子越正确。如果采用LM_i/sum_j(LM_j)的方式来归一化，LM绝对值数值越小，则归一化后的数值就越小，与真实结果相反。所以采用[MAX(LM)+1-LM_i]/sum_j(MAX(LM)+1-LM_j)来归一化，这样数值越大，说明句子越好，与真实结果相同。
- 文本拼接方式：在ASR topN的文本之前加“[unusedN]”表示第几个句子，然后再用“[SEP]”拼接，拼接后文本长度（去除[SEP]和[unusedN]）限制在100以内，如：[unused1] 对 [SEP] [unused2] 呃对 [SEP] [unused3] 嗯对 [SEP] [unused4] 啊对 [SEP] [unused5] 不对
- 加权方式：取出[unusedN]的向量作为每个句子的向量表示，然后用每个句子的AM/LM归一化得分进行加权，即文本表征*归一化得分，最后将N个结果求和表示整个拼接后的句子表征

序号	算法（单轮模型）	训练集数量来源	训练集特征	训练集标签	测试集（端到端意图）—固定种子3次中间值	去除纯特殊标记（背景音等）	固定种子3次平均值
6.2	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	10条拼接文本bert向量	Intent_trs	88.34%	90.66%	88.36%
7.2	BERT_(商城+物流)	训练集2（6.4k线上日志），ASR识别的Hypothesis	10条拼接文本bert向量及其LM，AM得分(用归一化得分进行句子向量加权)	Intent_trs	88.61%	90.66%	88.70%

实验7.2对于不同的N效果如下：

N	第一次实验	第二次实验	第三次实验	均值
2	87.94%	87.94%	88.20%	88.03%
4	88.81%	88.87%	88.94%	88.87%
6	88.67%	88.81%	89.21%	88.90%
8	88.81%	88.47%	88.07%	88.45%
10	88.47%	89.01%	88.61%	88.70%
12	88.20%	88.40%	88.74%	88.45%
14	88.67%	88.34%	88.74%	88.58%
16	88.27%	88.67%	88.54%	88.49%

对比实验五、实验六和实验七，不同N的影响如下（准确率是取用的均值）：