鲁棒性NLU预研

Snips数据上的效果

Snips数据（13084-train;700-test;7-intents）					
Model				Trans	ASR
Baseline	TextCNN	(a)	Train on Trans	97.40%	78.40%
		(b)	Train on ASR	95.40%	92.71%
	BiLSTM	(c)	Train on Trans	96.44%	74.73%
		(d)	Train on ASR	95.57%	92.00%
数据增强	TextCNN	(e)	Train on Trans and ASR	96.94%	93.66%
基于预训练	BiLSTM	(f)	Pre-trained ELMo	96.00%	80.42%
		(g)	(f)+Fine-Tuning	96.44%	87.56%
		(h)	(f)+CA Fine-Tuning(sup-conf)	96.70%	87.13%
		(i)	(f)+CA Fine-Tuning(unsup-conf)	96.57%	89.71%
基于对抗训练	TextCNN	(j)	Add KL loss	97.11%	93.54%
		(k)	Add KL loss（9 times datas）	97.47%	93.51%
基于N-Best	TextCNN	(l)	Combined 10-Best Sentence	95.40%	94.77%
		(m)	AvgPooling 10-Best embedding	93.31%	94.51%
		(n)	MaxPooling 10-Best embedding	90.40%	94.20%

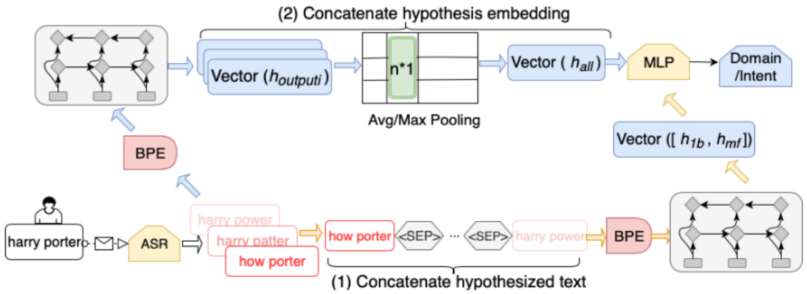
一. 背景

在传统的SLU（Spoken Language Understanding）中，主要包含ASR和NLU两个模块。ASR模块首先将用户语音转化成文本，NLU模块则对转化后的文本进行意图识别。在之前的做法中，ASR输出和用户表达最接近的文本给NLU进行意图识别。但是，ASR系统会产生一定的错误，输出的文本并不一定是最优的，这样的文本在NLU进行意图识别时会造成较大误差，引起识别结果的下降。因此，在NLU端，需要进一步融合ASR的错误，提升意图鲁棒性，以正确识别用户意图。

二. 基于N-Best的鲁棒性NLU

1. 论文：《Improving Spoken Language Understanding by Exploiting Asr N-Best Hypotheses 》Mingda Li, Weitong Ruan et. al. (ICASSP 2020)

核心思想：用ASR输出的N-Best代替ASR top1，利用更多样的文本信息。并且提出多种方法利用ASR的N-Best，包含：（1）在输入端拼接N-Best文本，模型embedding后进行预测；（2）输入端还是单条文本，每个N-Best进行embedding后，采用Pooling的方式融合成一个embedding，然后进行预测。



2. 复现结果（Snips数据集）

结论：复现结果基本符合论文结果

Snips数据（13084-train;700-test;7-intents）				
Model			Trans	ASR
1-best	(a)	Baseline	97.40% (0)	78.40% (0)
	(b)	Train on ASR	95.40% (-2%)	92.71% (+14.31%)
10-best	(c)	Combined Sentence	95.40% (-2%)	94.77% (+16.37%)
	(d)	PoolingAvg	93.31% (-4.09%)	94.51% (+16.11%)
	(e)	PoolingMax	90.40% (-7%)	94.20% (+15.80%)

3. 实验分析

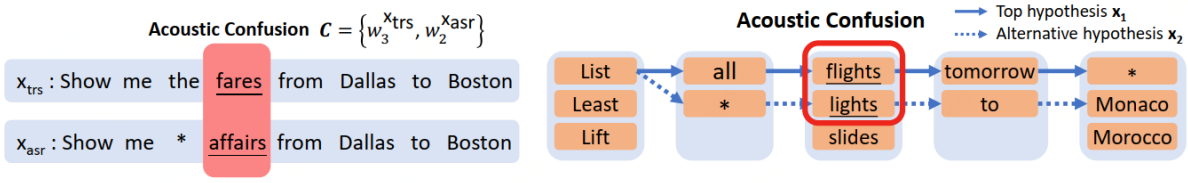
1. Snips数据集上，直接拼接N-Best文本能够带来更大的收益。其能够保证在ASR文本上有很大提升，并且在Trans数据上不会带来更大的下降

4. 外呼数据实验（大件外呼N-Best实验）

三. 基于预训练的鲁棒性NLU

1. 论文：《Learning Asr-Robust Contextualized Embeddings For Spoken Language Understanding》Huang, Chao Wei , and Chen, Yun-Nung. (ICASSP 2020)

核心思想：构建真实文本 和 ASR识别文本的（字-字）pair对，增加了对混淆字的loss（cosine_embedding_loss），即两句话中对应位置出现了不同的字，计算两个不同字的CosineEmbeddingLoss，优化的时候会拉近混淆字的embedding表征，从而使分类任务具有健壮性。



(a) Supervised confusion extraction with aligned utterances

(b) Unsupervised confusion extraction using a WCN generated by ASR

Fig. 2: Illustration of different extraction approaches. $*$ denotes a blank symbol for alignment purpose.

$$\mathcal{L}_{\text{conf}} = \frac{1}{|C|} \sum_{c \in C} \sum_{i=0}^{1} 1 - \frac{h_{t_1,i}^{x_1} \cdot h_{t_2,i}^{x_2}}{\left\| h_{t_1,i}^{x_1} \right\| \left\| h_{t_2,i}^{x_2} \right\|},$$

2. 复现结果（Snips数据集）

结论：复现结果大致符合论文结果

Model		论文结果	复现结果
(b)	Context-independent	72.70	74.73
(c)	Pre-trained ELMo	77.86	80.42
(d)	(c)+fine-tune, L_LM only	87.74	87.56

(e)	(c)+fine-tune, L_FT(sup-conf)	88.52	87.55
(f)	(c)+fine-tune, L_FT(unsup-conf)	89.55	89.71

注：论文里面是5次平均值，复现是跑了一次的结果

Snips数据（13084-train;700-test;7-intents）			
Model		Trans	ASR
(a)	Train on Trans	96.44% (0)	74.73% (0)
(b)	Train on ASR	95.57% (-0.87)	92.00% (+17.27%)
(c)	Pre-trained ELMo	96.00% (-0.44%)	80.42% (+5.69%)
(d)	(c)+fine-tune, $\mathcal{L}_{LM}$ only	96.44% (0)	87.56% (+12.83%)
(e)	(c)+fine-tune, $\mathcal{L}_{FT}$(sup-conf)	96.70% (+0.26%)	87.13% (+12.4%)
(f)	(c)+fine-tune, $\mathcal{L}_{FT}$(unsup-conf)	96.57% (+0.13%)	89.71% (+14.98%)

3.实验分析

分析复现结果，发现了一系列问题：

1. d实验效果优于或等于e实验，ca_finetune带来的效果不明显（e实验单独跑了3次，其最高值为当前值，未超过d实验结果）。
2. f实验结果最好，但查看其finetune数据量是d,e实验的9倍，实验结果的提升是由ca_finetune带来的，还是通过增加数据提升的，原因不明确。
3. d实验仅用真实文本的训练数据就能到达87.56的实验结果，如果将e实验的9倍数据进行finetune，效果可能超过e实验。

4.备注（由于论文是基于allennlp的elmo模型进行训练的，该模型无开源中文预训练，中文数据无法进行实验。预训练后续可尝试用拼音或文字+拼音的形式，同英文处理方式，做预训练）

四. 基于对抗训练的鲁棒性NLU

1. 论文：《Towards an ASR error robust Spoken Language Understanding System》Ruan Weitong et. al. (INTERSPEECH 2020)

核心思想：对(语音识别文本，真实文本)二者的预测结果分布加入KL loss度量两句话的预测分布差异性，通过拉近ASR文本和正确文本预测分类的概率分布，优化参数。

提出了SLU鲁棒性评估准则：模型对ASR误差的鲁棒性在有ASR错误的数据上预测结果提高，并且在没有ASR错误的数据上预测性能不能退化。

$$
\begin{aligned}
Loss\left(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{A}_i, \boldsymbol{\theta}\right) = \quad & \epsilon_1 * CE\left(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{\theta}\right) \\
& + \epsilon_2 * CE\left(\boldsymbol{y}_i; \boldsymbol{A}_i, \boldsymbol{\theta}\right) \\
& + \epsilon_3 * KL\left(p\left(\boldsymbol{y}_i|\boldsymbol{A}_i\right), p\left(\boldsymbol{y}_i|\boldsymbol{X}_i\right)\right)
\end{aligned}
$$

论文实验结果：

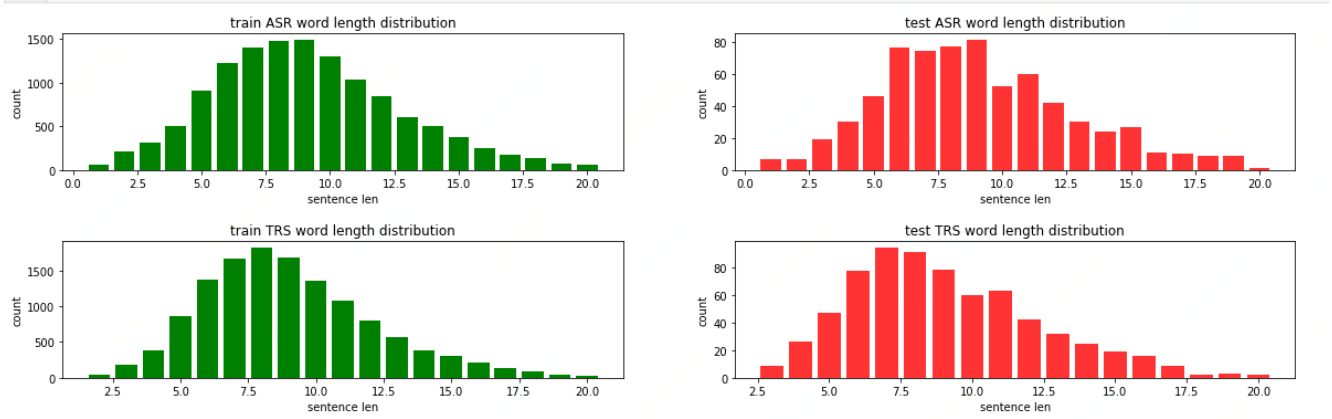
Table 2: *Relative classification performance (F1) w.r.t. baseline model performance on Alexa dataset (negative means performance degradation).*

model	parameters ϵ_1	ϵ_2	ϵ_3	Trans.(%)	ASR(%)
baseline	1.0	0.0	0.0	0	0
data augmentation	1.0	1.0	0.0	0	1.76
train on ASR	0.0	1.0	0.0	-3.76	1.41
proposed model	1.0	0.0	40.0	**0.88**	**4.41**

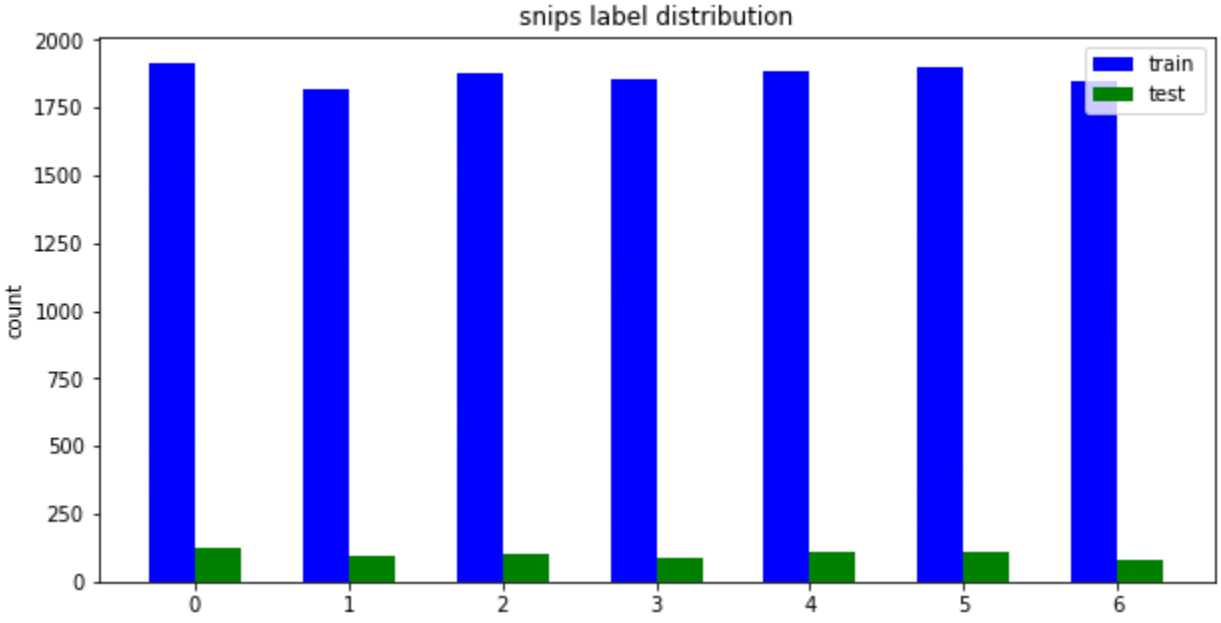
2. 复现结果（Snips数据集）

结论：复现结果大致符合论文结果

snips数据长度分布图：



snips数据 label 分布图：



论文中的AITS数据语音数据难以获取，采用已有的snips数据，使用TextCNN模型进行实验。

Snips数据（13084-train;700-test）													
Model		parameters			Trans						ASR					
		(trs, asr, kl)			test1	test2	test3	test4	test5	mean	test1	test2	test3	test4	test5	mean
(a)	Baseline	1	0	0	96.71%	97.43%	97.71%	98.00%	97.14%	**97.40%**	77.86%	77.86%	78.14%	79.43%	78.71%	78.40%
(b)	Data augmentation	1	1	0	97.14%	97.00%	96.71%	97.00%	96.86%	96.94%	93.43%	93.86%	93.57%	93.29%	94.14%	**93.66%**
(c)	Train on ASR	0	1	0	95.14%	95.71%	95.43%	95.14%	95.57%	95.40%	93.00%	92.86%	92.43%	92.57%	92.71%	92.71%
(d)	Add KL loss	1	0	1	96.71%	97.14%	97.00%	97.29%	97.43%	97.11%	93.29%	93.71%	93.57%	92.86%	94.29%	93.54%

统计分析结果(ε1, ε2, ε3分别代表真实文本loss、ASR文本loss、KL loss的权重大小)：

Snips数据（13084-train;700-test;7-intents）						
Model		parameters			Trans	ASR
		ε_1	ε_2	ε_3		
(a)	Baseline	1.0	0.0	0.0	97.40% (0)	78.40% (0)
(b)	Data augentation	1.0	1.0	0.0	96.94% (-0.46%)	93.66% (+15.2
(c)	Train on ASR	0.0	1.0	0.0	95.40% (-2%)	92.71% (+14.3
(d)	Add KL loss	1.0	0.0	1.0	97.11% (-0.29%)	93.54% (+15.1

论文2实验，并未提供模型的超参数，在增加KL loss计算时，loss不稳定，超参数调整的时候出现过loss为nan的情况。实验在30个epoch和learning rate设置为0.003的时候得到当前实验结果。

实验d在取原文参数1,0,40的时候，极不稳定，loss出现nan，取ε3为1进行实验。

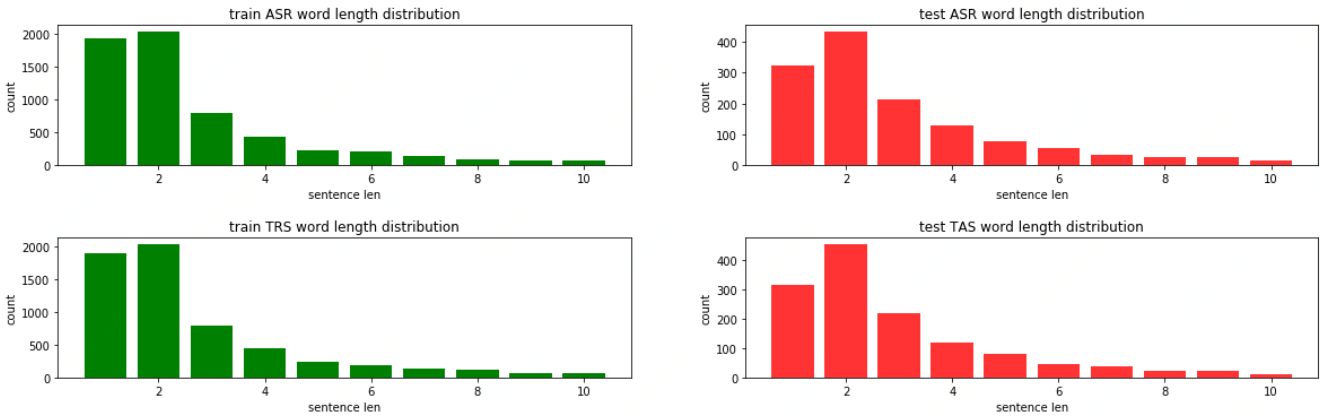
实验结果表明：

- **相比于baseline，增加KL loss的对抗训练能够在Trans数据上保持基本一致，并提高在ASR识别数据上的表现（实验(a) vs 实验(d)），符合论文提出结果。**
- 数据增强提升鲁棒性实验与增加KL loss的对抗训练实验效果基本一致，都能够保持Trans的效果并提升ASR数据的表现（实验(b) vs 实验(d)）。
- 仅在ASR数据上进行模型训练和预测，在提升ASR数据表现的同时也降低了在Trans数据上的效果，符合论文提出结果（实验(a) va 实验(c)）。

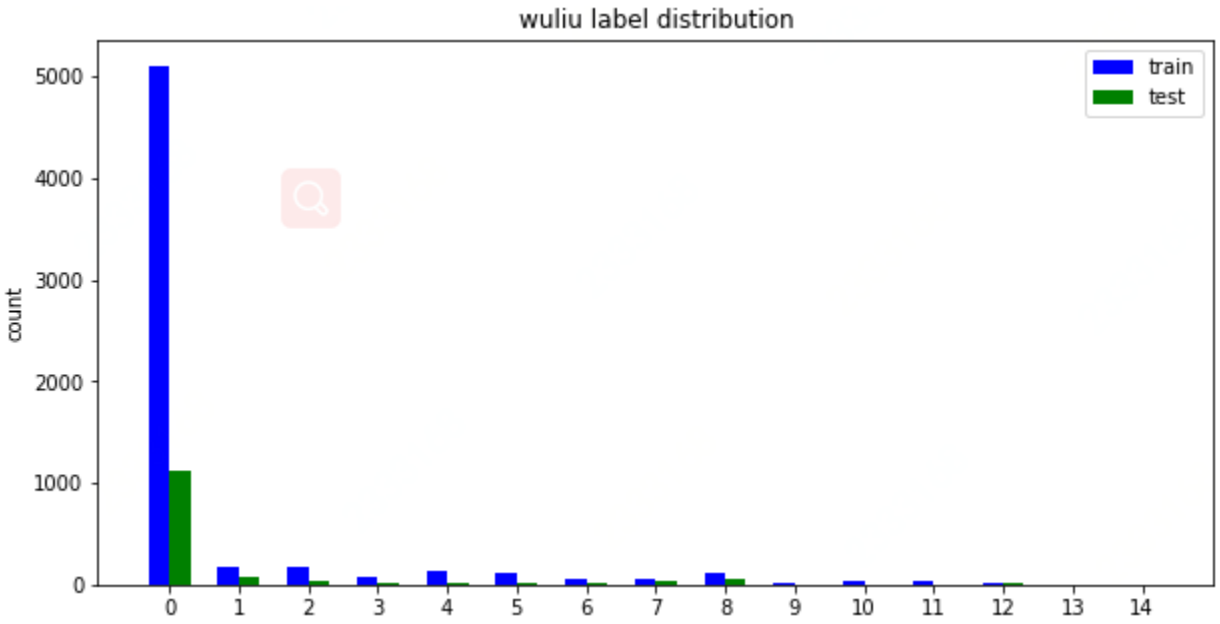
3. 外呼数据实验(训练集去重后 (asr: 1567, trs: 1666) ，测试集去重后(asr: 529, trs: 539))

结论：复现结果不太符合论文结果，主要原因是外呼数据的分布和Snips数据差别较大。对测试集进行长短切分后，发现在长文本上KL loss有所效果，在短文本上未带来效果。进一步扩充数据中

外呼数据长度分布图：



外呼数据 标签 分布图：



该部分分三个实验进行：

- 实验一，在train,test数量分别为6257条（去除特殊标记[SD]等）和1419条进行实验。
- 实验二，去除了实验一中训练集里面的无意义badcase数据（喂-对，对-不对） 等，测试集保持不变。
- 实验三，训练集采用实验二的数据，将测试集划分为短文本（低于4个字）和长文本（高于4个字），实验在不同测试集上的效果。

实验一：

采用6257条训练数据（去除特殊标记[SD]等），1419条测试数据进行实验。

外呼数据（6257-train;1419-test）																		
Model		parameters			Trans							ASR						
					test1	test2	test3	test4	test5	mean		test1	test2	test3	test4	test5	mean	
(a)	Baseline	1	0	0	92.04%	91.26%	91.05%	92.04%	90.91%	91.46%	0.00%	88.23%	87.74%	87.46%	88.65%	87.53%	87.92%	0.00%
(b)	Data augmentation	1	1	0	90.77%	90.27%	90.63%	90.42%	89.99%	90.42%	-1.04%	87.17%	87.32%	86.75%	87.32%	86.75%	87.06%	-0.86%
(c)	Train on ASR	0	1	0	89.92%	90.70%	89.15%	89.29%	90.06%	89.82%	-1.64%	86.12%	86.96%	86.19%	86.12%	86.75%	86.43%	-1.49%
(d)	Add KL loss	1	0	1	90.27%	90.56%	89.78%	90.35%	89.64%	90.12%	-1.34%	86.82%	87.39%	86.68%	87.10%	86.61%	86.92%	-1.00%

外呼数据（6257-train;1419-test）						
Model		parameters			Trans	ASR
		ε_1	ε_2	ε_3		
(a)	Baseline	1.0	0.0	0.0	91.46% (0)	87.92% (0)
(b)	Data augmentation	1.0	1.0	0.0	90.42% (-1.04)	87.06% (-0.86)
(c)	Train on ASR	0.0	1.0	0.0	89.82% (-1.64)	86.43% (-1.49)
(d)	Add KL loss	1.0	0.0	1.0	90.12% (-1.34)	86.92% (-1.00)

实验一结果分析：

- 论文提出的KL loss实验在外呼ASR数据上未达到预期结果（实验(a) vs 实验(d)）。

实验二：

采用6120条训练数据（去除特殊标记[SD]，去除（喂-对）这一类的无法纠正的badcase数据），在1419条测试数据上进行实验。

外呼数据（6120-train;1419-test）															
Model		parameters			Trans							ASR						
					test1	test2	test3	test4	test5	mean		test1	test2	test3	test4	test5	mean	
(a)	Baseline	1	0	0	90.35%	91.19%	91.12%	91.61%	91.26%	91.11%	0.00%	86.89%	87.46%	87.67%	88.02%	87.74%	87.56%	0.00%
(b)	Data augmentation	1	1	0	90.84%	90.20%	90.35%	90.77%	90.20%	90.47%	-0.63%	87.32%	87.32%	87.39%	87.81%	87.32%	87.43%	-0.12%
(c)	Train on ASR	0	1	0	89.92%	90.49%	90.49%	90.20%	90.41%	90.30%	-0.80%	86.54%	87.46%	87.32%	86.82%	87.32%	87.09%	-0.46%
(d)	Add KL loss	1	0	1	89.50%	90.49%	89.92%	90.35%	90.98%	90.25%	-0.86%	86.82%	87.66%	86.96%	87.03%	87.32%	87.16%	-0.40%

外呼数据（6120-train;1419-test）						
Model		parameters			Trans	ASR
		ε_1	ε_2	ε_3		
(a)	Baseline	1.0	0.0	0.0	91.11% (0)	87.56% (0)
(b)	Data augmentation	1.0	1.0	0.0	90.47% (-0.64)	87.43% (-0.12)
(c)	Train on ASR	0.0	1.0	0.0	90.30% (-0.81)	87.09% (-0.46)
(d)	Add KL loss	1.0	0.0	1.0	90.25% (-0.86)	87.16% (-0.40)

实验二结果分析：

- 论文提出的KL loss实验在外呼ASR数据上未达到预期结果（实验(a) vs 实验(d)）。

实验三：

训练集6120训练集， 测试集1419条，按数据长度大于4，切分测试集1084条短文本和335条长文本进行实验。

- 短文本1084条数据实验结果

外呼数据（6120-train;1084-short_test）																		
Model		parameters			Trans							ASR						
					test1	test2	test3	test4	test5	mean		test1	test2	test3	test4	test5	mean	
(a)	Baseline	1	0	0	97.51%	97.60%	97.97%	97.97%	97.79%	97.77%	0.00%	94.10%	94.19%	94.00%	94.19%	94.10%	94.12%	0.00%
(d)	Add KL los	1	0	1	97.42%	97.23%	96.86%	97.42%	96.86%	97.16%	-0.61%	93.91%	94.10%	93.63%	94.19%	93.91%	93.95%	-0.17%

- 长文本335条数据实验结果

外呼数据（6120-train;335-long_test）																		
Model		parameters			Trans							ASR						
					test1	test2	test3	test4	test5	mean		test1	test2	test3	test4	test5	mean	
(a)	Baseline	1	0	0	69.25%	70.45%	71.04%	71.64%	70.15%	70.51%	0.00%	64.78%	66.57%	65.37%	67.16%	65.07%	65.79%	0.00%
(d)	Add KL los	1	0	1	71.04%	69.85%	68.36%	70.15%	70.15%	69.91%	-0.60%	67.46%	66.87%	63.58%	65.67%	68.06%	66.33%	0.54%

- 测试集1419条总体实验结果

外呼数据（6120-train;1419-test）																		
Model		parameters			Trans							ASR						
					test1	test2	test3	test4	test5	mean		test1	test2	test3	test4	test5	mean	
(a)	Baseline	1	0	0	90.84%	91.19%	91.61%	91.75%	91.26%	91.33%	0.00%	87.18%	87.67%	87.24%	87.81%	87.25%	87.43%	0.00%
(d)	Add KL los	1	0	1	91.19%	90.77%	90.13%	90.98%	90.55%	90.73%	-0.61%	87.67%	87.67%	86.54%	87.46%	87.81%	87.43%	0.00%

实验结果表明：

- KL loss在长文本上能够带来效果提升，在短文本上没有效果，主要是因为外呼数据中，短文本上下文参考信息较少，且大部分错误无法正确预测（ASR识别错误，如 喂-对）。

实验四：

在6120条训练数据上增加了1624条数据，测试集保持不变，得到以下实验结果：

- 短文本1084条数据实验结果

外呼数据（去重 (6120+1624)-train;1084-short_test）																		
Model		parameters			Trans							ASR						
					test1	test2	test3	test4	test5	mean		test1	test2	test3	test4	test5	mean	
(a)	Baseline	1	0	0	97.29%	97.02%	97.48%	97.20%	97.20%	97.24%	0.00%	93.78%	93.33%	94.14%	93.42%	93.24%	93.58%	0.00%
(d)	Add KL loss	1	0	1	97.20%	97.20%	96.57%	96.84%	97.02%	96.97%	-0.27%	93.87%	94.14%	93.42%	94.05%	94.14%	93.92%	0.34%

- 长文本335条数据实验结果

外呼数据（去重 (6120+1624)-train;335-long_test）																		
Model		parameters			Trans							ASR						
					test1	test2	test3	test4	test5	mean		test1	test2	test3	test4	test5	mean	
(a)	Baseline	1	0	0	64.84%	65.48%	67.10%	67.10%	70.97%	67.10%	0.00%	61.94%	63.87%	65.16%	63.55%	66.77%	64.26%	0.00%
(d)	Add KL loss	1	0	1	68.71%	70.00%	66.45%	69.03%	69.35%	68.71%	1.61%	64.84%	63.87%	63.23%	65.48%	64.52%	64.39%	0.13%

- 测试集1419条总体实验结果

外呼数据（去重 (6120+1624)-train;1419-test）																		
Model		parameters			Trans							ASR						
					test1	test2	test3	test4	test5	mean		test1	test2	test3	test4	test5	mean	
(a)	Baseline	1	0	0	89.63%	89.58%	90.30%	90.10%	91.01%	90.12%	0.00%	86.26%	86.37%	87.30%	86.37%	86.99%	86.66%	0.00%
(d)	Add KL loss	1	0	1	90.48%	90.78%	89.46%	90.28%	90.49%	90.30%	0.17%	87.01%	86.99%	86.29%	87.31%	87.15%	86.95%	0.29%