鲁棒性加入音素特征

计划实验内容list：

1. 分析当前的训练集和测试集，统计trs文本和asr文本相似发音的字的比例。（发音相同；发音相近）
2. 阅读归纳有关build ASR-robust representations的相关论文，结合音素特征，尝试应用到我们的业务数据集上。包括：

    - Learning Spoken Language Representations with Neural Lattice Language Modeling
    - LEARNING ASR-ROBUST CONTEXTUALIZED EMBEDDINGS FOR SPOKEN LANGUAGE UNDERSTANDING
    - Towards an ASR error robust Spoken Language Understanding System
    - Are Neural Open-Domain Dialog Systems Robust to Speech Recognition Errors in the Dialog History? An Empirical Study
    - Simplify the Usage of Lexicon in Chinese NER

论文阅读：《Using Phoneme Representations to Build Predictive Models Robust to ASR Errors》

- 本文出发点：ASR识别通常会产生同音异形字，相似音素的发音会让模型将音素预测成不同的字，进一步将错误引向更高维度的空间，增加了NLU识别的难度，如果在低维度的音素空间中（音素相比字符错误更少），对音素进行嵌入表征，作为辅助特征加入到字符级别的嵌入表征会带来更好的预测效果。
- 对于ASR识别错误，评估常用WER （word error rate），该论文研究字符级别和音素级别的embedding表征，相应的CER（charactor error rate），PER（phoneme error rate）较word级别的错词率更低。
- 实验对音素的嵌入表征进行了四种探索：

    1. p2vc: phoneme2vec on surrounding phonemes
    2. p2vm: phoneme2vec on mixed REF and ASR utterances
    3. p2va: phoneme2vec on aligned REF and ASR utterances.
    4. Phoneme Embeddings from Seq2Seq - s2s

- 最终实验在CNN和LSTM模型下进行，输入特征包括word_token, charactor_token, phoneme_token

ASR错误示例：

Table 1: Examples of ASR errors. WER, CER and PER are word, character and phoneme error rates, respectively.

Reference text (followed by its phonemes)	Transcribed text (followed by its phonemes)	WER	CER	PER
What Canadian city has the largest population w-ah-t k-ax-n-ey-d-iy-ax-n s-ih-t-iy hh-ae-z dh-ax l-aa-r-jh-ax-s-t p-aa-p-y-ax-l-ey-sh-ax-n	*Well, comedian city has the largest population* w-eh-l k-ax-m-iy-d-iy-ax-n s-ih-t-iy hh-ae-z dh-ax l-aa-r-jh-ax-s-t p-aa-p-y-ax-l-ey-sh-ax-n	0.285	0.205	0.142
What is amitriptyline w-ah-t ih-z ae-m-iy-t-r-ih-p-t-ax-l-ay-n	*One is amateur delete* w-ah-n ih-z ae-m-ax-t-er d-ax-l-iy-t	1.000	0.631	0.529
Remarkably accessible and affecting r-ax-m-aa-r-k-ax-b-l-iy ax-k-s-eh-s-ax-b-ax-l ae-n-d ax-f-eh-k-t-ax-ng	*Remarkably accessible on defecting* r-ax-m-aa-r-k-ax-b-l-iy ax-k-s-eh-s-ax-b-ax-l ax-n d-ax-f-eh-k-t-ax-ng	0.500	0.093	0.074

实验对音素的嵌入表征进行了四种探索(音素共40个，嵌入维度为20)：

1. p2vc:

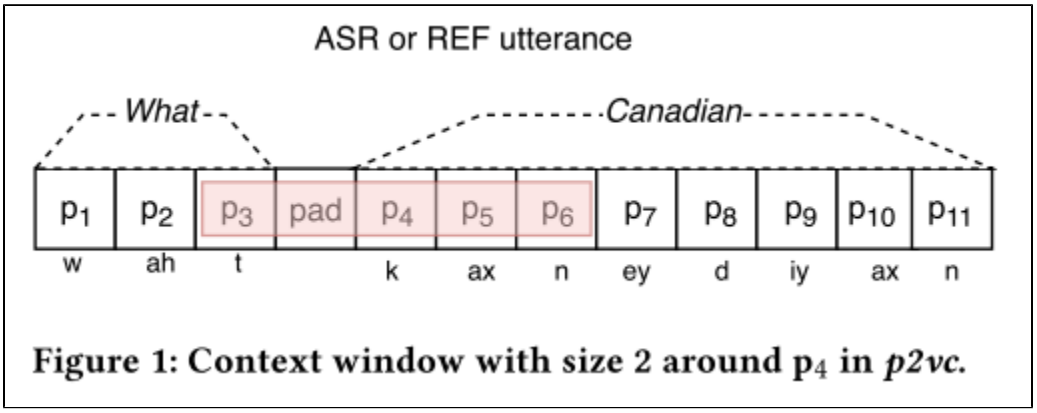

Figure 1: Context window with size 2 around p_4 in *p2vc*.

2. p2vm

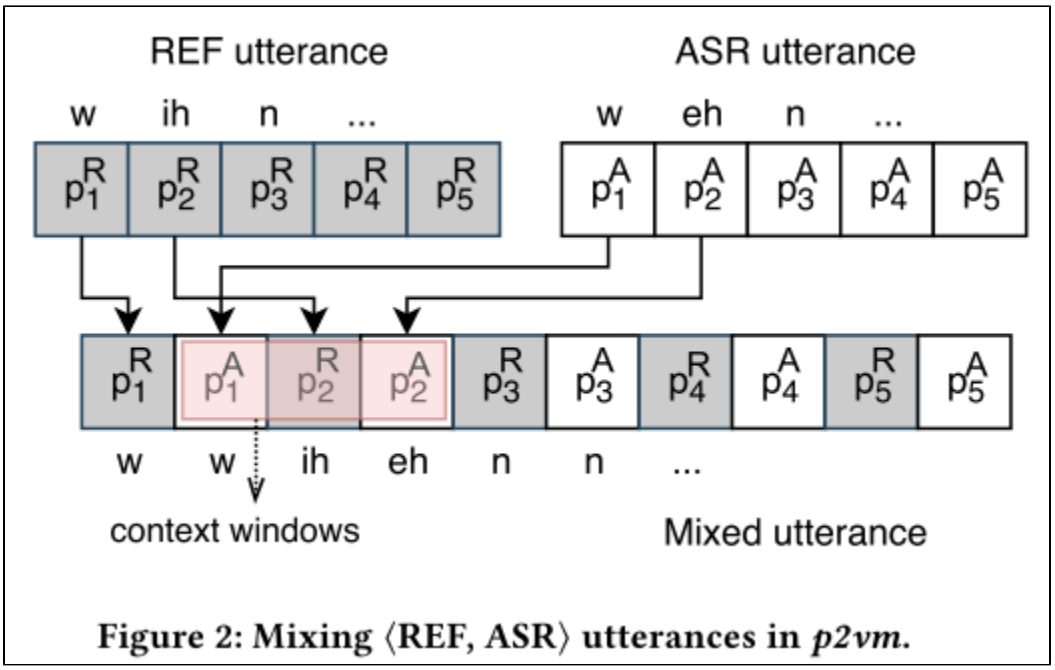

Figure 2: Mixing ⟨REF, ASR⟩ utterances in *p2vm*.

3. p2va

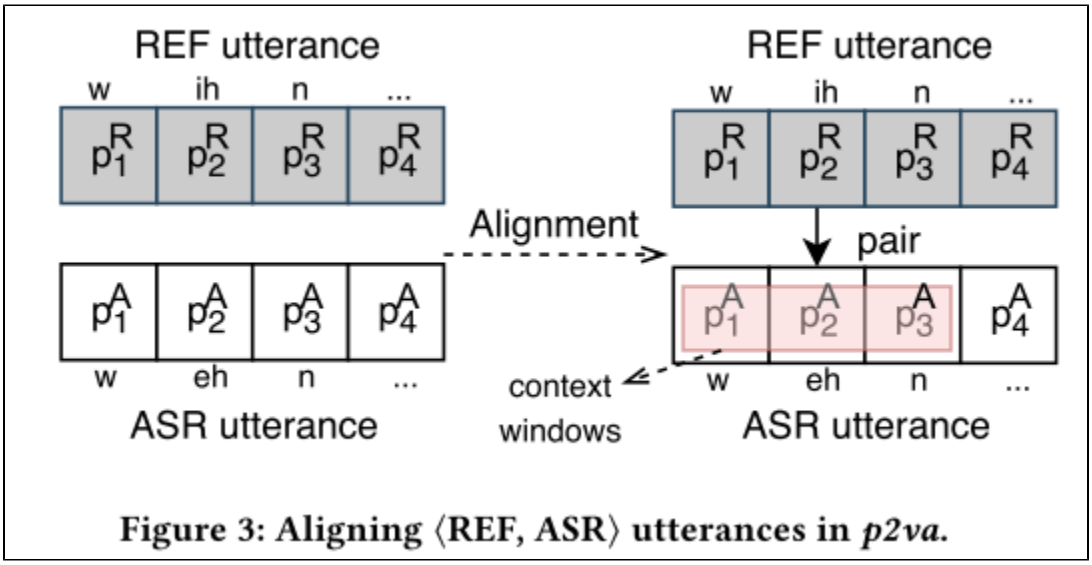

Figure 3: Aligning ⟨REF, ASR⟩ utterances in *p2va*.

4. Phoneme Embeddings from Seq2Seq - s2s

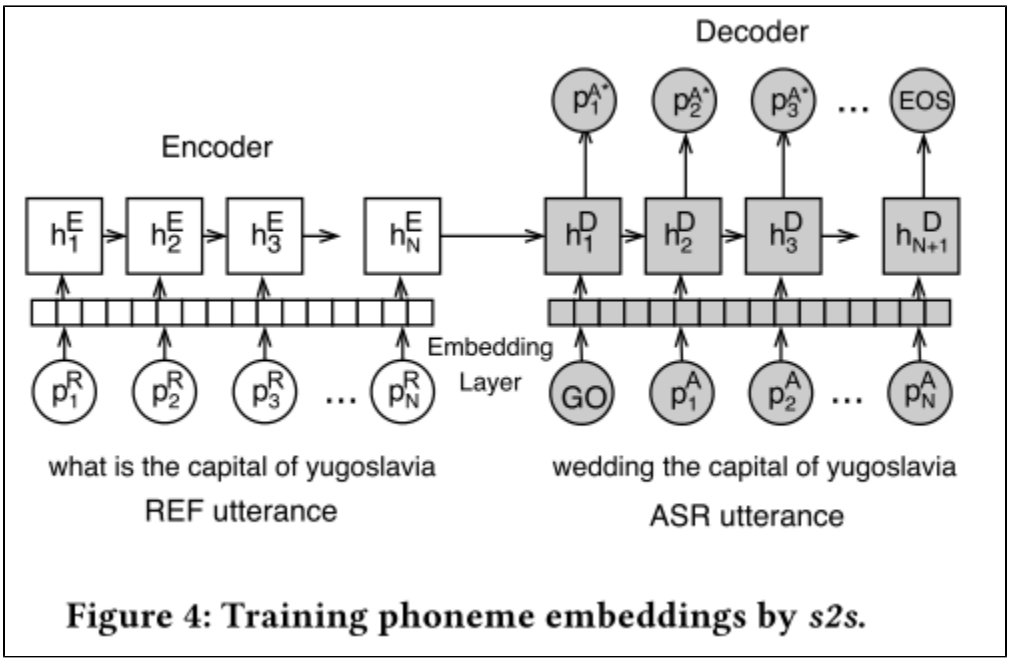

Figure 4: Training phoneme embeddings by *s2s*.

NLU模型结构：

1. CNN的模型结构：

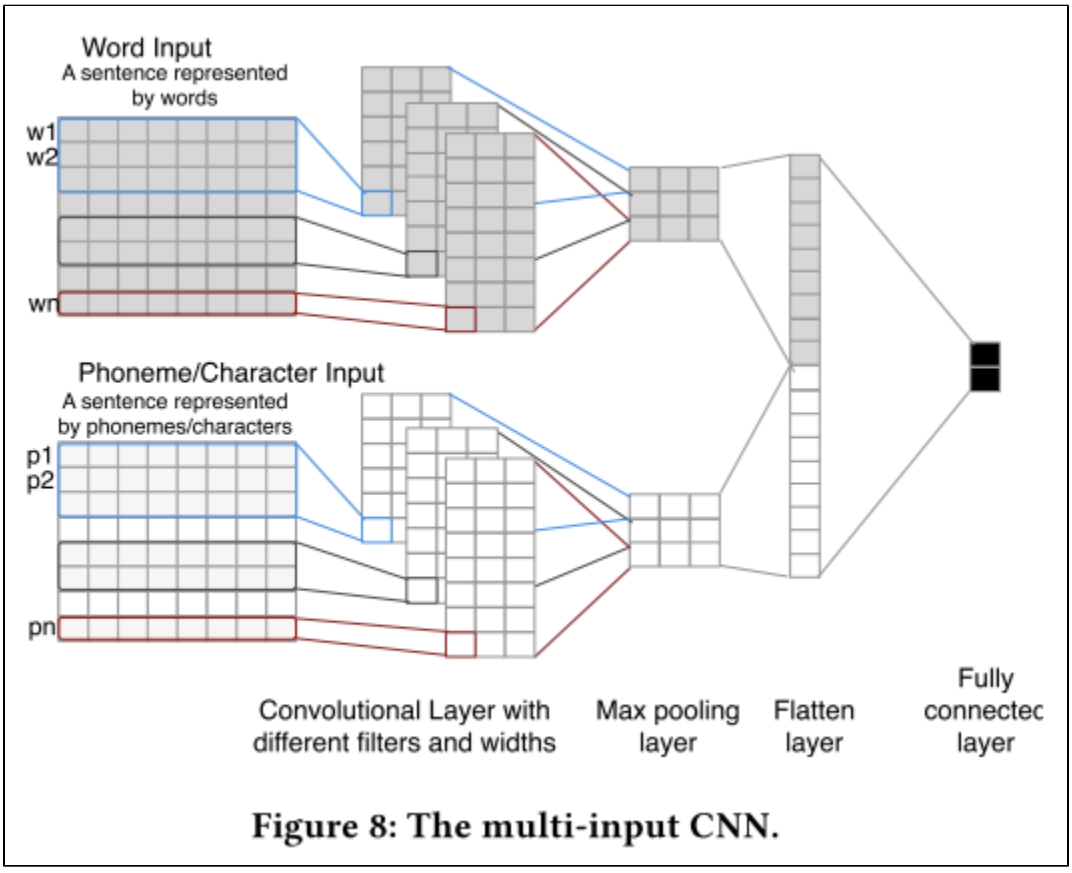

Figure 8: The multi-input CNN.

2. LSTM模型结构：

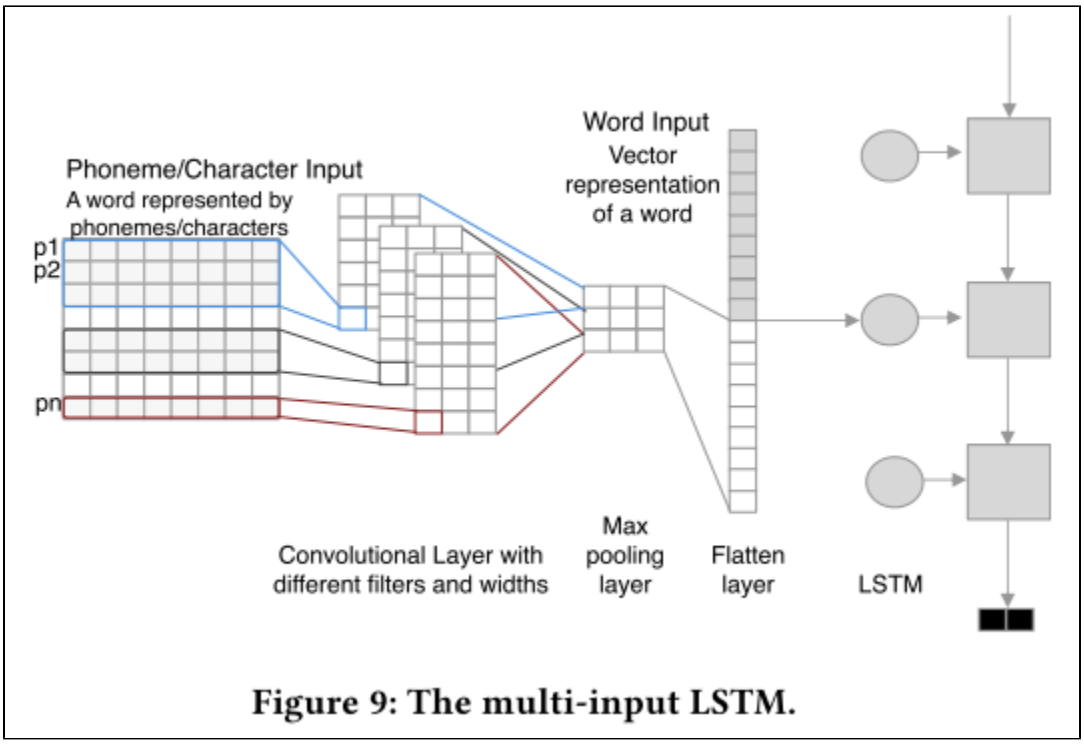

Figure 9: The multi-input LSTM.

NLU意图预测结果（use p2va0；w,c,p 分别是词，字符，音素，拼接代表组合）：

Table 5: Accuracy of classification models using different inputs. The best scores are bold.

	SST	TQ-50	TQ-6
w	0.414	0.627	0.743
c	0.293	0.549	0.678
p	0.318	0.570	0.702
wc	0.408	0.648	0.756
wp	**0.419**	0.646	**0.758**
wcp	0.405	**0.650**	0.754

可以看到，加粗的最好的结果总是包含了phoneme embedding

数据分析：

1.1 分析结论（在165条意图改变的数据中）：

1. 通过加入音素特征来增强文字的robust表征在大件外呼场景不可行，原因：ASR转写错误的文字，转写结果中音素也是错误的；（音素正确，文字转写的错误占非常低，且大多是结尾发音字，如：啊，了）
2. 相同发音造成的意图识别错误（发音一致，转写字与正确字不一致，导致意图不一致）出现的比例：0.6%（1/165）165个意图识别错误；解决此问题在当前数据中的收益不大。
    a. 例子：

	ASR结果	人工标注结果

文本结果	了解我会帮你传达还有别的**是**吗	了解我会帮你传达还有别的**事**吗
音素信息	l iao3 j ie3 uu uo3 h ui4 b ang1 n i3 ch uan2 d a2 h ai2 ii iu3 b ie2 d e5 sh ix4 m a2	l e5 j ie3 uu uo3 h ui4 b ang1 n i3 ch uan2 d a2 h ai2 ii iu3 b ie2 d e5 sh ix4 m a5
意图结果	ask	confirm

3. 相近发音造成的意图识别错误（发音相近，转写字与正确字不一致，导致意图不一致）在训练集中出现比例：40.60%（67/165）165个意图识别错误；在测试集中出现的频率解决此问题为首要问题。
    a. 例子：

ASR结果	人工标注结果
那个（other）	**哪**个（ask_identity）
啊你**就吗**（other）	啊你讲么（deafness）
现在不方便的我现在太**慢**了（deny）	现在不方便的我现在太**忙**了（busy）

4. 由于标注问题导致的意图识别错误占比也有20.60%，后续需要注意语音转写和人工转写意图的校验工作。

1.2 相近发音产生的错误解决方案：

1. 针对表格中的类似错误，可以通过构建句子pair对，拉近句子语义表示来解决。
2. 像（喂-对），（对-喂）这类错误，暂时无很好的解决方法（单个字就是对应一个意图）。

2.1 数据统计：

训练数据6389条
语音转写和人工转写长度相等5542条，其中二者句子pair中，不同字有1426个，总字数为15664个，错字比例为9.1%
长度相等数据中，(语音转写句子，人工转写句子)pair，存在字不同934条
存在不同字的样本(934条)中有165条意图发生改变。即意图改变样本占比17.67%
在意图发生改变的数据(165条),不同字的pair对共350组
不同字pair对(350组)，181条声母相同占比51.71%，167条韵母相同占比47.71%

2.2相同发音，不同字

的仅7条数据，其中仅3条数据意图不同，2条为标注错误，1条为发音相同造成的错误（数据顺序：asr,trs）：

	0	1	2	3	4	5	6	7	8
0	3a58309d-dc5b-4a27-87a5-4831ece61ffb	我是什么是是的	我是什么事是的	uu uo3 sh ix4 sh en2 m e5 sh ix4 sh ix4 d e5	uu uo3 sh ix4 sh en2 m e5 sh ix4 sh ix4 d e5	confirm	confirm	uuuo3 shix4 shen2 me5 shix4 shix4 de5	uuuo3 shix4 shen2 me5 shix4 shix4 de5
1	id1	你他妈说话呢	你他妈说话你	n i3 t a1 m a1 sh uo1 h ua4 n i3	n i3 t a1 m a1 sh uo1 h ua4 n i3	deafness	other	ni3 ta1 ma1 shuo1 hua4 ni3	ni3 ta1 ma1 shuo1 hua4 ni3
2	id1	哎是的	唉是的	aa ai1 sh ix4 d e5	aa ai1 sh ix4 d e5	confirm	confirm	aaai1 shix4 de5	aaai1 shix4 de5
3	id1	哎对	唉对	aa ai1 d ui4	aa ai1 d ui4	confirm	confirm	aaai1 dui4	aaai1 dui4
4	id1	嗯这个先不确定到时候我会通知在通知你们吧	嗯这个先不确定到时候我会通知再通知你们吧	ee en2 zh e4 g e4 x ian1 b u4 q ve4 d ing4 d a...	ee en2 zh e4 g e4 x ian1 b u4 q ve4 d ing4 d a...	busy	busy	eeen2 zhe4 ge4 xian1 bu4 qve4 ding4 dao4 shix2...	eeen2 zhe4 ge4 xian1 bu4 qve4 ding4 dao4 shix2...
5	b34dccf3-2541-41a0-b2b4-828dd4999f25	醉酒可以推迟多久	最久可以推迟多久	z ui4 j iu3 k e3 ii i3 t ui1 ch ix2 d uo1 j iu3	z ui4 j iu3 k e3 ii i3 t ui1 ch ix2 d uo1 j iu3	change_time	ask	zui4 jiu3 ke3 iii3 tui1 chix2 duo1 jiu3	zui4 jiu3 ke3 iii3 tui1 chix2 duo1 jiu3

了解我会帮你传达还有别的是吗	了解我会帮你传达还有别的事吗	l iao3 j ie3 uu uo3 h ui4 b ang1 n i3 ch uan2 d a2 h ai2 ii iu3 b ie2 d e5 sh ix4 m a2	l e5 j ie3 uu uo3 h ui4 b ang1 n i3 ch uan2 d a2 h ai2 ii iu3 b ie2 d e5 sh ix4 m a5	ask	confirm	是-事,

2.3 在165条意图不同的数据中：

trs文本和asr文本标注错误34条，占比20.60%（语音转写文本意图标注和asr转写意图标注间隔时间长，标注人不一样）。

语音识别导致差异大而产生的错误64条，占比38.79%。

相似发音造成意图错误67条，占比40.60%。其中，单字改变引起的错误如：（喂，对），(那，哪)等占了50个，比例为74.63%。以及多个字产生的错误如：（三-啥,四-事）占18个，比例为25.37%。

一、背景

意图鲁棒性的实验中，只考虑了句子维度的特征，如AM/LM，音素的特征未考虑。因此，实验音素特征对意图鲁棒性的影响。

实验模型采用TextCNN 和 Bert。

二、实验1：embedding层面的拼接方式：字和音素的embedding进行concat 和 add操作

实验结论

目前的实验将音素当做字来进行表征（音素上下文表征音素向量），缺乏了音素应有的特性（发音相近）。CNN实验表明，音素embedding的concat和add操作对意图识别准确率未带来提升

实验数据

2020.10.24—2020.10.31人工标注语音数据中，共6389条，与之前6404条数据大致对齐，测试集完全对齐共1492条。其中11月2日音素数据未获得，采用10月31日数据补充

CNN实验

（1）实验结果

实验编号	训练集特征	embedding方式	拼接方式	训练集标签	准确率（三次实验）	中间值	平均值
1	query	TextCNN	无	Intent_trs	83.59% 84.73% 84.06%	84.06%	84.12%
2	音素	TextCNN	无	Intent_trs	83.32% 83.52% 83.93%	83.53%	83.59%
3	query+音素	TextCNN	concat	Intent_trs	83.98% 83.98% 83.71%	83.98%	83.89%
4	query+音素	TextCNN	add	Intent_trs	84.65% 83.65% 83.91%	83.91%	84.07%
5	query	word2vec	无	Intent_trs	82.65% 82.38% 82.31%	82.38%	82.45%
6	音素	word2vec	无	Intent_trs	83.72% 83.12% 83.32%	83.32%	83.39%
7	query+音素	word2vec	concat	Intent_trs	83.24% 82.91% 81.90%	82.91%	82.68%
8	query+音素	word2vec	add	Intent_trs	82.51% 82.98% 83.11%	82.98%	82.87%

（2）实验方法

TextCNN embedding方式：随机初始化embedding向量，使用TextCNN模型经过任务finetune后的embedding

- 构建字和音素的vocab表，先使用字进行训练得到字的embedding矩阵，在此embedding基础上，进行音素训练，更新该embedding的音素向量。
- 加载上述embedding, 分别对字和音素各自进行三次实验，得到相应的结果。
- 为了实现embedding层面的拼接，增加了两个placeholder, 分别对应字的声母和韵母音素特征。
- 加载上述embedding，分别以concat方式 和add方式进行三次实验，得到相应的结果。

word2vec embedding方式：

在训练集的6389条数据下使用gensim word2vec模型， 滑动窗口设置成5，min_count设置为1，训练30个epoch，得到仅有字的embedding和仅有音素的embedding，按照vocab顺序，将二者拼接成一个词表。

在此embedding下，进行1,2,3,4实验的对应实验。

（3）实验分析

在当前数据集下，统计分析音素与字的对应联系，汉字训练集字表共669个字，其中仅有65个字的音素产生了变化，意味着90%左右的字与音素特征一一对应，10%左右的字对应的音素特征发生了变化。如果仅靠这10%的字对应的音素变化，来影响意图分类的结果，难以体现。

在当前6339条训练数据下的word2vec模型训练得到的embeding，字的表现不佳，而音素的表象较字的效果更好，原因是音素vocab仅177个，较字数669个，其空间更小。

（4）后续todo:

参考chen论文：《LEARNING ASR-ROBUST CONTEXTUALIZED EMBEDDINGS FOR SPOKEN LANGUAGE UNDERSTANDING》，将trs文本映射到相关音素，和ASR音素对齐，实现发音相近音素的相似表征。

三、实验2： 采用文本拼接的方式： 字+音素的交叉拼接 & 句子+音素的前后拼接

实验结论

BERT实验和CNN实验表明，音素的两种拼接方式对意图识别准确率未带来提升。

实验数据

2020.10.24—2020.10.31人工标注语音数据中，共6389条，与之前6404条数据大致对齐，测试集完全对齐共1492条。其中11月2日音素数据未获得，采用10月31日数据补充。

Bert实验

（1）实验结果

实验编号	训练集特征	拼接方式	训练集标签	准确率(三次均值)
1	query+音素	字+音素	Intent_trs	85.65%
2	query+音素	字+音素	Intent_hyp	86.28%
3	query+音素	句子+音素	Intent_trs	86.33%
4	query+音素	句子+ 音素	Intent_hyp	87.22%
5	query	无	Intent_trs	87.43%
6	query	无	Intent_hyp	87.89%

（2）实验方法

- 将所有文本的音素去重后加入到vocabulary中。具体的，根据Bert作者描述（https://github.com/google-research/bert/issues/62），##字 在训练中文的时候几乎没有用到，所以替换vocabulary中的部分##字 为识别到的音素。这里不直接加入到字典中，是因为直接加入导致字典扩大，预训练模型不支持，并且直接加入的数据也没有预训练embedding。
- 将每个字的音素拼接到字之后，并修改Bert tokenization的切分方式， 使音素token不被wordpiece切开。得到的数据格式为：<s> 喂 uu_B ui4_E 是 sh_B ix4_E 的 d_B e5_E </s>、<s> 喂 是 的 <SEP> uu_B ui4_E sh_B ix4_E d_B e5_E </s>两种

（3）实验分析

将低频使用的汉字直接替换成音素特征，一方面引入了原来汉字的意思产生部分误差，另一方面音素未经过预训练，音素之间没有关联。导致实验结果并不理想

CNN实验

（1）实验结果

实验编号	训练集特征	拼接方式	训练集标签	准确率
1	query+音素	字+音素	Intent_trs	83.25%
2	query+音素	字+音素	Intent_hyp	85.14%
3	query+音素	句子+音素	Intent_trs	83.24%
4	query+音素	句子+音素	Intent_hyp	84.58%
5	query	无	Intent_trs	83.66%
6	query	无	Intent_hyp	85.68%

（2）实验方法

- 考虑到CNN是以滑动窗口形式捕获特征，窗口大小通常取2,3,4,5，拼接音素特征后长度扩大了原来的三倍，语义单元也应该相应扩大三倍，所有取拼接音素后的窗口为6,9,12,15。
- 数据格式为：<s> 喂 uu_B ui4_E 是 sh_B ix4_E 的 d_B e5_E </s>、<s> 喂 是 的 <SEP> uu_B ui4_E sh_B ix4_E d_B e5_E </s>两种

（3）实验分析

音素和字之间的联系未学习到，不能纠正识别错的字，所以对意图识别没有提升