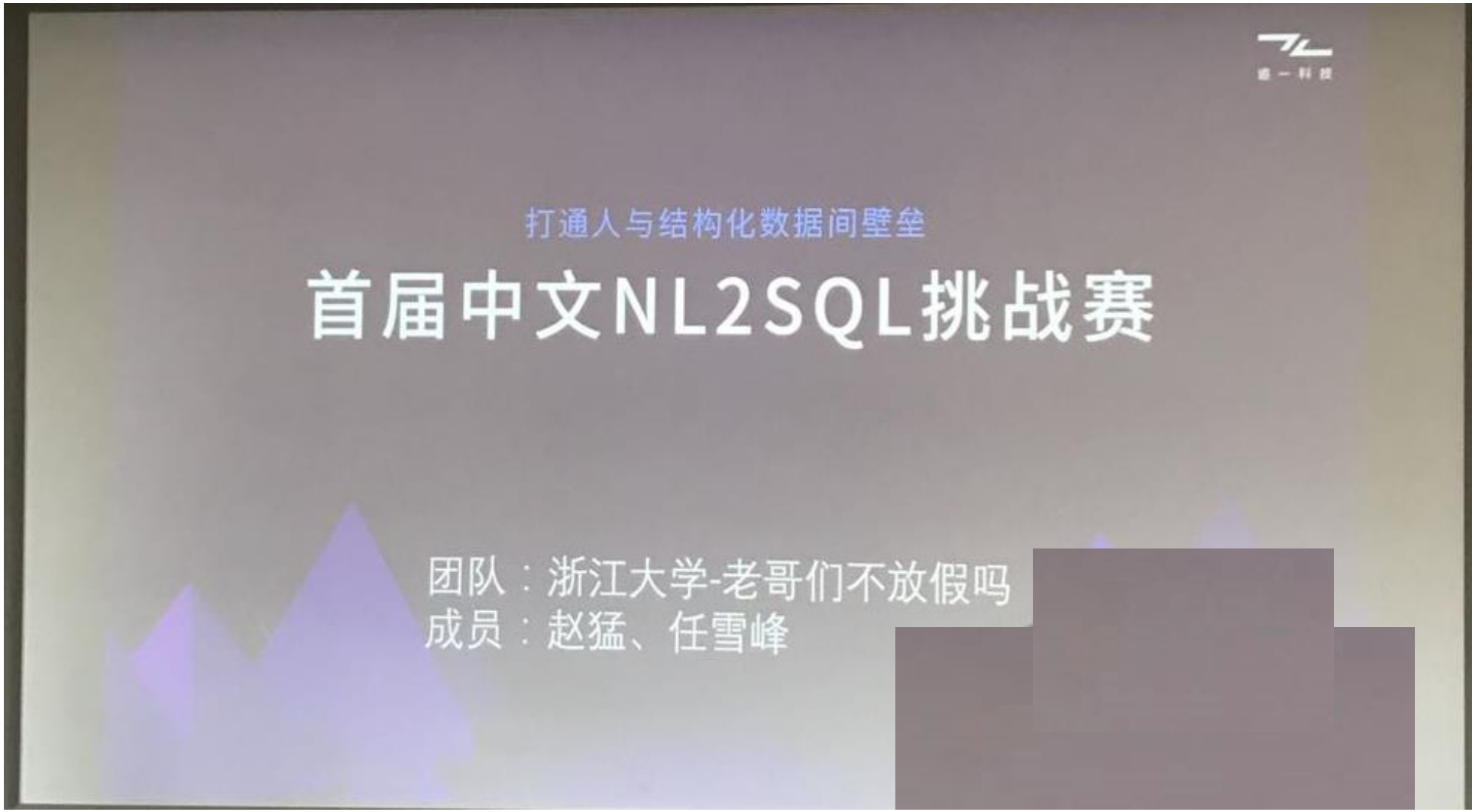

打通人与结构化数据间壁垒
首届中文NL2SQL挑战赛
团队：浙江大学-老哥们不放假吗
成员：赵猛、任雪峰

数据集分析

- 与Spider数据集相比：
  - 是指定数据表的单表查询，不含JOIN操作
  - 是相对简单的查询，不含GroupBy，Having，嵌套查询

Complex question	What are the name and budget of the departments with average instructor salary greater than the overall average?

Complex SQL	`SELECT T2.name, T2.budget`
	`FROM instructor as T1 JOIN department as T2 ON T1.department_id = T2.id`
	`GROUP BY T1.department_id`
	`HAVING avg(T1.salary) >`
	`(SELECT avg(salary) FROM instructor)`

数据集分析

- 与Wikisql数据集相比：
  - select 的col以及对应的agg操作是不定数目的
  - 增加了where condition之间的连接符(and,or)
  - where col可能会出现多次选择同一列
  - where value可能在问题中无法直接匹配
  - 生成的sql执行结果可能为空，无法使用Execution-Guided Decoding，而这个往往在wikisql能提升模型几个百分点

相比于Spider的复杂性和Wikisql的过于规范性，nl2sql数据集更能满足大多实际场景的需求

模型思考

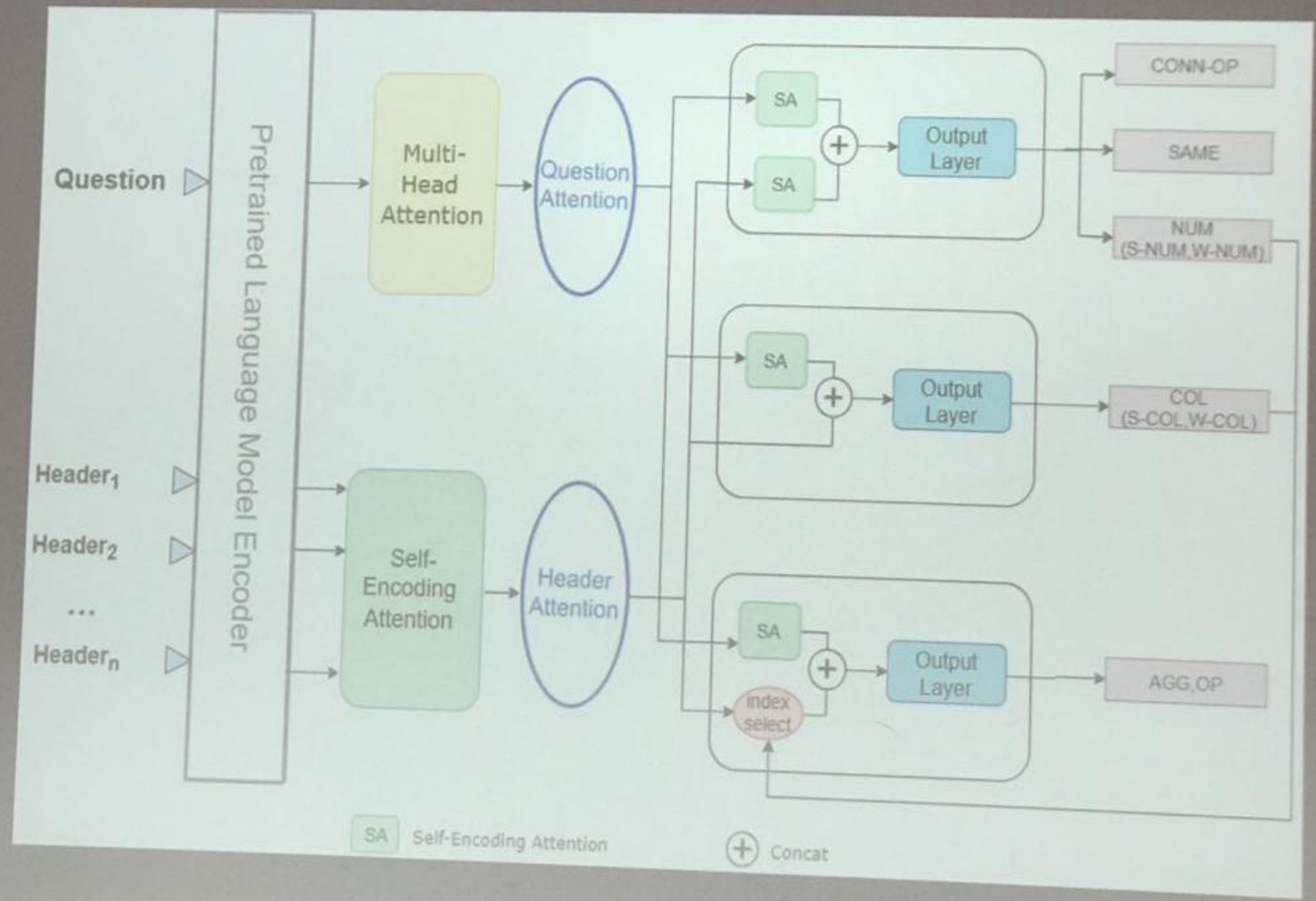
- 尝试过SQLNet，SQLova，X-SQL，表现不太令人满意
- 从QA的角度来思考NL2SQL
  - NL2SQL核心任务是列COL的抽取以及条件值W-VAL的抽取

    给定问题Q和表列名$H_1,H_2...,H_n$，期望输出中核心部分S-COL,W-COL都是得到候选列集合$\{H_j,...,H_k\}$,这个问题可以看做是不定项多项选择问题，

    候选列对应的操作S-AGG,W-OP的预测,可以看做是单项选择，也可以看做是普通的多分类问题，CONN-OP也是多分类问题

    对于W-VAL的抽取，即使标签条件值可能不在Q中完全出现，但是对于Q中与其相似的部分，可以看做是抽取式问答任务

型概览


Question
Header$_1$
Header$_2$
...
Header$_n$
Pretrained Language Model Encoder
Multi-Head Attention
Question Attention
Self-Encoding Attention
Header Attention
SA
SA
Output Layer
SA
SA
Output Layer
SA
SA
index select
Output Layer
CONN-OP
SAME
NUM (S-NUM,W-NUM)
COL (S-COL,W-COL)
AGG,OP
SA Self-Encoding Attention
Concat

Main Moudle

- Self-Encoding Attention

$$Attention(H) = \mathbf{softmax}(HW)^{\mathrm{T}}H$$

自编码注意力用于将字符级别的特征编码为词级别的特征，用于提取列特征

- Multi-Head Attention

单独对问题部分的特征进行再编码来更好利用问题语义间的关联信息

- Output Layer

输出层采用Bert采用的BertOutput结构，Dense+Dropout+LayerNorm

Cls SubMoudle

- 进行分类任务，预测CONN-OP(3)，S-NUM(2)，W-NUM(3)，SAME(2)
  - 统计数据S-COL个数，最大为3，训练集为3/41522,验证集0/4396
    统计数据W-COL个数，最大为4，训练集为39/41522,验证集0/4396

    训练过程丢弃S-COL个数为3和W-COL个数为4的数据

  - 加入了一个新的SAME预测，用来判断问题Q是否在W-COL中包含相同的列

Tag SubMoudle

- 进行序列标注任务，预测S-COL，W-COL
  - 经过Self-Encoding Attention编码后的列名向量$H \in R^{hlen \times d}$，对H进行序列标注判断是否是候选列
  - 在此基础上，由于面临存在选择同列的问题，当前流行的模型不能解决这个问题，因此在标注时候采用多输出的方法，比如对于W-COL，模型对每一列输出3个Score，然后只选择前W-NUM个Score取获得最大分数的列作为结果

    训练时候，对标签中的W-COL按在Q中出现位置进行排序，来使模型3个输出层按顺序分别去关注对应位置的W-COL
  - 损失函数部分采用交叉熵，只关注前NUM个输出层的损失，其他为0

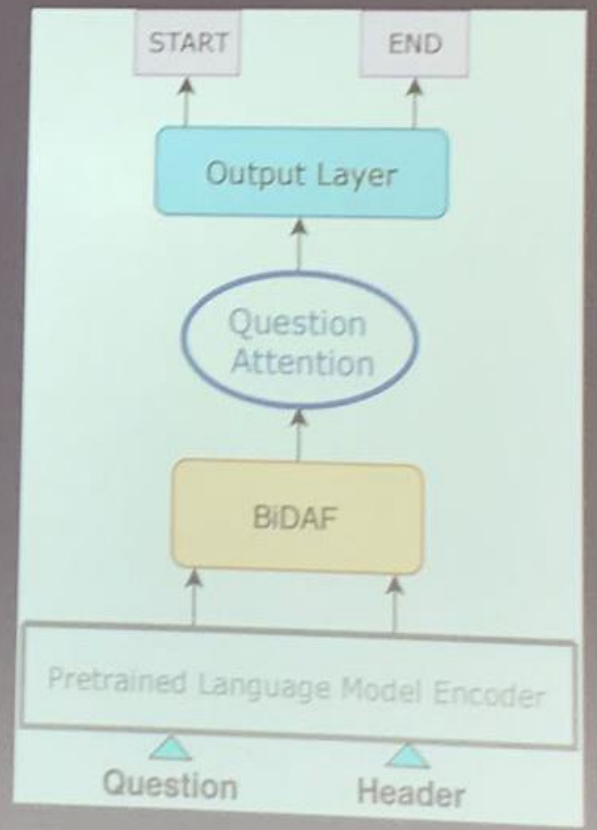
Tag SubMoudle(with index_select)

- 根据Cls子模块预测的NUM和Tag子模块预测的COL，进行index select，得到候选列向量$H' \in R^{NUM \times d}$，类似于Tag模块的 $H \in R^{hlen \times d}$，来预测S-AGG(6)，W-OP(4)

- 对于Tag子模块，进行过许多模型的尝试，比如BiDAF，QANet，DCMN等，在该数据集上表现均不如最终模型

W-VAL Moudle

- 把主模块抽取到的条件列单独和问题拼接作为模型输入，预测SQuAD风格的Answer Span的起始位置和结束位置



W-VAL 标签自动标注

- 标签中W-VAL有很多无法在问题Q中直接匹配到
  - 数据预处理，通过正则表达式，定义词典等操作，进行中文数字转阿拉伯数字，日期规范化，股票等专有名词缩写转换等清洗操作
  - W-VAL起始位置标注，通过先对Q分词，再用编辑距离，最长公共子串，相似度匹配相结合的模糊匹配方法尽可能标注出Q中与标签VAL相似的部分

q_o	wv
什么人大的项目属于哲学这个学科?	[中国人民大学, 哲学]
帮我算一下达标的饮品数量是多少	[合格]
最新股价不足十四块六而且最新股票总价值超过四十四亿港币的是什么股票	[44, 14.6]

q	wv_a
什么人大的项目属于哲学这个学科?	[人大, 哲学]
帮我算一下达标的饮品数量是多少	[达标]
最新股价不足14.6而且最新股票总价值超过44亿港币的是什么股票	[44, 14]

W-VAL 结果匹配

- 最初方案，任意类型的列都通过模型抽取value，然后text类型的列抽取到的value再通过类似标注时采用的模糊匹配方法去和数据表中该列的所有列值匹配，得到最相似的值作为结果
- 改进
  - 对于real类型的列，仍通过抽取模型来抽取，再通过单位转换等后处理得到W-VAL
  - 对于text类型的列，由于有些标注的词语区分度不够，去掉模型抽取这一操作，直接拿该列所有列值与问题Q做模糊匹配

结果检验修正

- SAME修正：SAME预测的准确率很高(99.8%)，对于W-COL，如果预测的结果包含同列，SAME为0，那么就会选择其他概率大的列来替换，反之，通过复制替换操作保证结果包含同列
- CONN-OP修正：W-NUM为1，CONN-OP一定为空，反之，一定不为空，同时，在SAME为1情况下，CONN-OP保证为or
- W-OP修正：对于text类型列，保证W-OP不为>,<，在预处理阶段也是对列类型做了重新判断，纠正那些应该是real类型但是包含了一些None，na这样干扰数据变成text类型的列

- 预训练语言模型：前期使用的中文BERT-wwm预训练模型，后来中文XLNet-mid预训练模型发布，换成XLNet大约比BERT高了1个百分点
- 优化器：Adam + Warmup+Lookahead,weight decay:1e-2,lr:2e-5,batch size:12
- 模型集成：由于复赛的时间限制，最终使用了三个模型进行集成，主模型a，模型b替换输出层，模型c在Tag模块使用了BiDAF模型，模型a，b，c进行预测分数加权。

  开始时使用1:1:1进行加权，后来观察到模型c对于select部分预测效果较好，改进权重，修改S-NUM，S-COL，S-AGG的分数权重为1:1:1.8，在复试测试集提高了0.5个百分点

模型结果

	初赛测试集	复赛测试集
single model	89.3(88.8)	90.3
ensemble model	89.6(89.4)	91.1

	conn_op	sel	agg	wc	wo	wv	lx
验证集	0.981	0.989	0.987	0.966	0.976	0.915	0.891