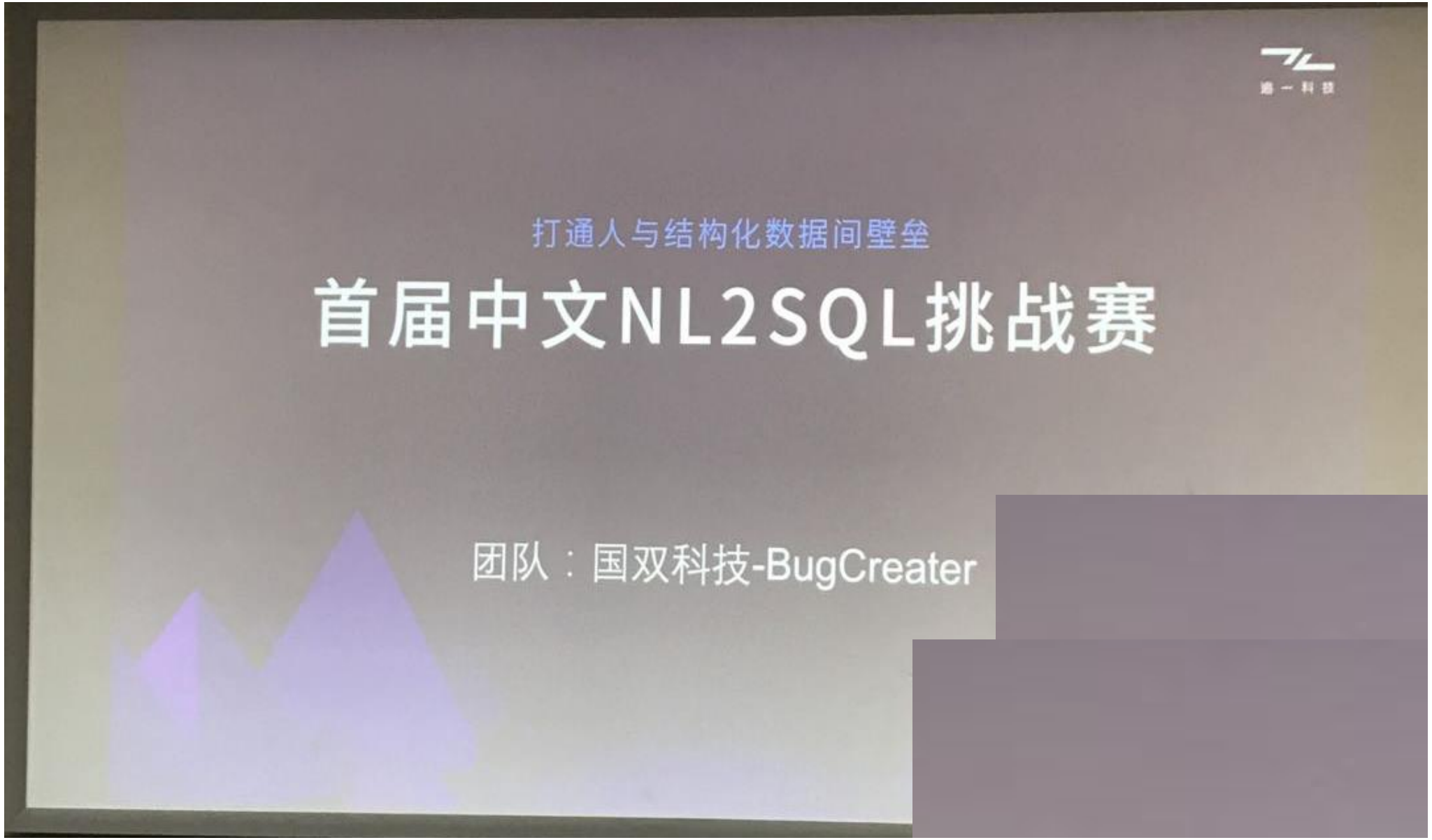

打通人与结构化数据间壁垒
首届中文NL2SQL挑战赛
团队：国双科技-BugCreater

任务介绍

```
{
  "question": "世茂茂悦府新盘容积率大于1，请问它的套均面积是多少",

  "sql": {
        "sel": [7],                  #SQL选择的列
        "agg": [0],                  #聚合函数
        "cond_conn_op": 0,   #条件之间的关系
        "conds": [
            [6, 0, 1],        #条件(列、类型、值)
            [1, 2, "世茂茂悦府"]
        ]
  }
}
```

- 输入：question、表格信息（列名、列类型、内容）
- 目标：预测需要的SQL语句多个字段对应元素的内容
  - sel：与agg结合后做多分类
  - agg：同上
  - cond_conn_op：多分类
  - conds：抽取问题

辅助预测内容：

- sel_num：多分类
- conds_num: 多分类

数据分析

- 评价指标：$Score_{lf} + Score_{ex}$（样本粒度的指标）
- 与WikiSQL四点不同：可以利用表的内容信息，存在conds value不能从question提取的样本；sel、agg存在多个；无conds为空的样本
- 同上train中有1w左右样本的conds value不能直接从question中匹配得到（不确定数目的样本存在错误匹配 例如哪个股票2019年价格在二十块 cond value为20）

	训练集	验证集	测试集
Num	41522	4396	4000(~5000)
Question length(max)	112	88	103
Header num(max)	23	22	24
Header length(0.99)	123	136	123
Sel num(max) Cond num(max)	3 4	2 3	

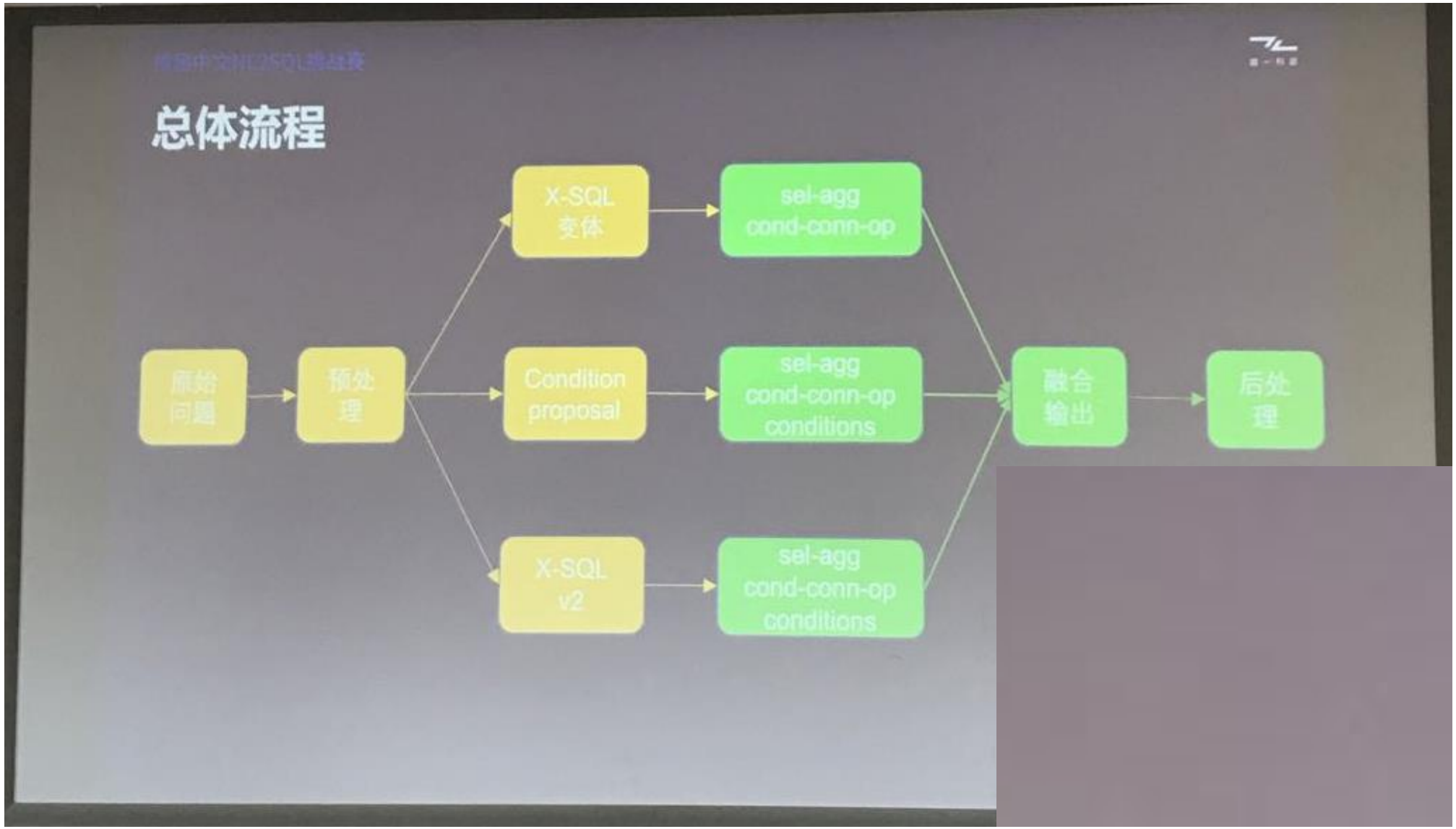
相关工作

- WikiSQL数据集：相关领域代表数据集
- 对问句和header联合建模方法（BERT & MTDNN + Multi-task Prediction）：
  - SQLova （Hwang et. al. 2019）
  - X-SQL （He et. al. 2019）
- 候选集合通过SQL检索结果校验：
  - Execution guided（Wang et.al 2018）

https://github.com/salesforce/WikiSQL

难点分析

- 测试集存在训练集完全未见过的表格(10%->20~30%)
  - 模型需要具备header向量表达的泛化能力（header embedding）
- conds待抽取词语重叠、同一column的opt可能不一致（例如：大于5且小于10）
  - 需要针对每个column的每个条件分别建模
- Label中词语有25%不在问句中出现
  - 需要在训练和测试中建立抽取短语向Column短语的映射
- 训练集部分不匹配以及完全不匹配的样本如何利用
  - 类似lic2019信息抽取（克服一定漏标注问题、label匹配可能出现错误）


总体流程
原始问题
预处理
X-SQL变体
Condition proposal
X-SQL v2
sel-agg cond-conn-op
sel-agg cond-conn-op conditions
sel-agg cond-conn-op conditions
融合输出
后处理

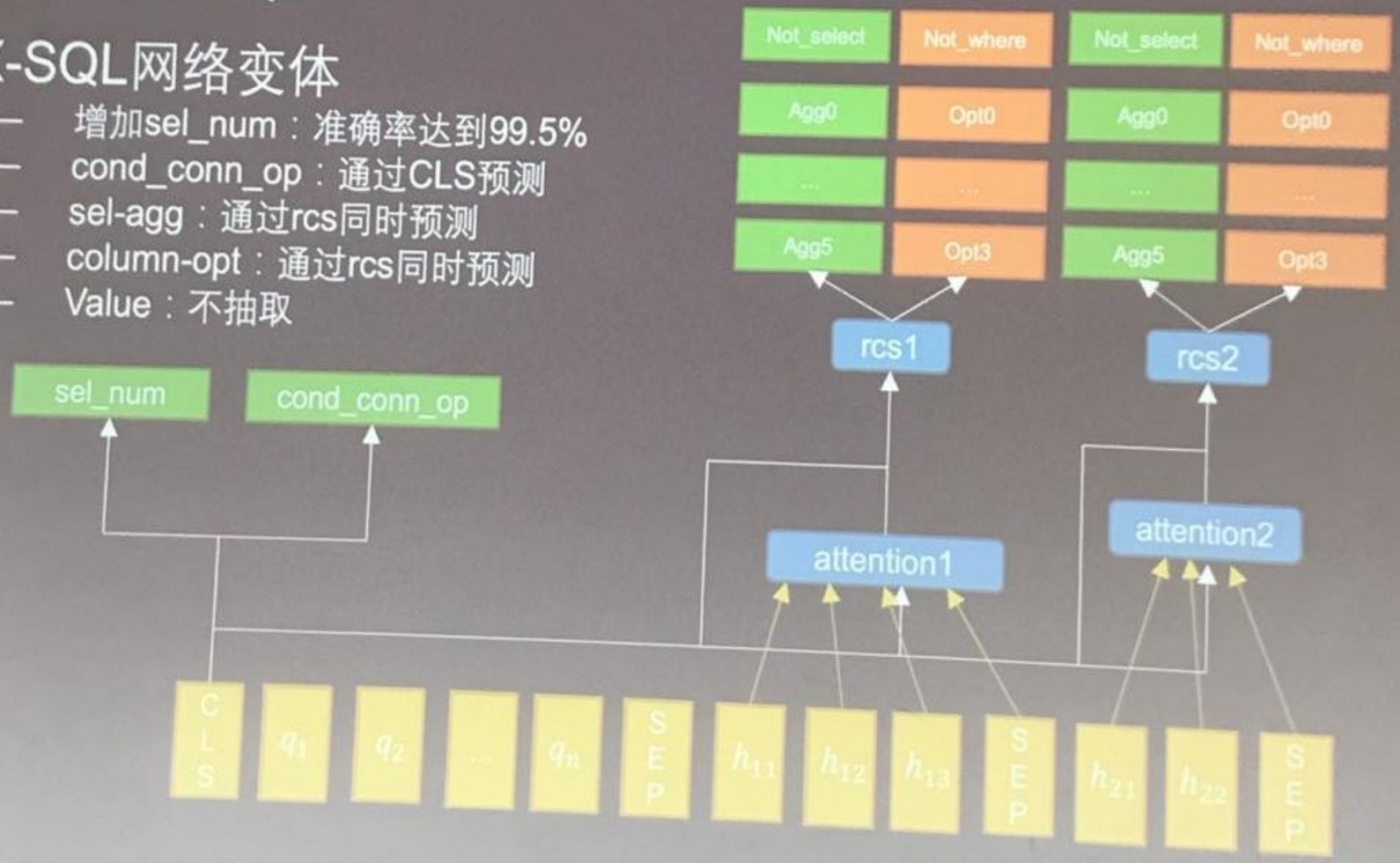
预处理

- 汉字转数字
  - 按照特定规则，将Question中的年份、价格等汉字表达转换为数字
  - 处理总样本量5%
- 标注校正（蒸馏）
  - 训练数据中，有许多条件的值不在Question中。通过半监督学习的方法，对这数据进行校正
  - 处理总样本量的20%

模型1：X-SQL

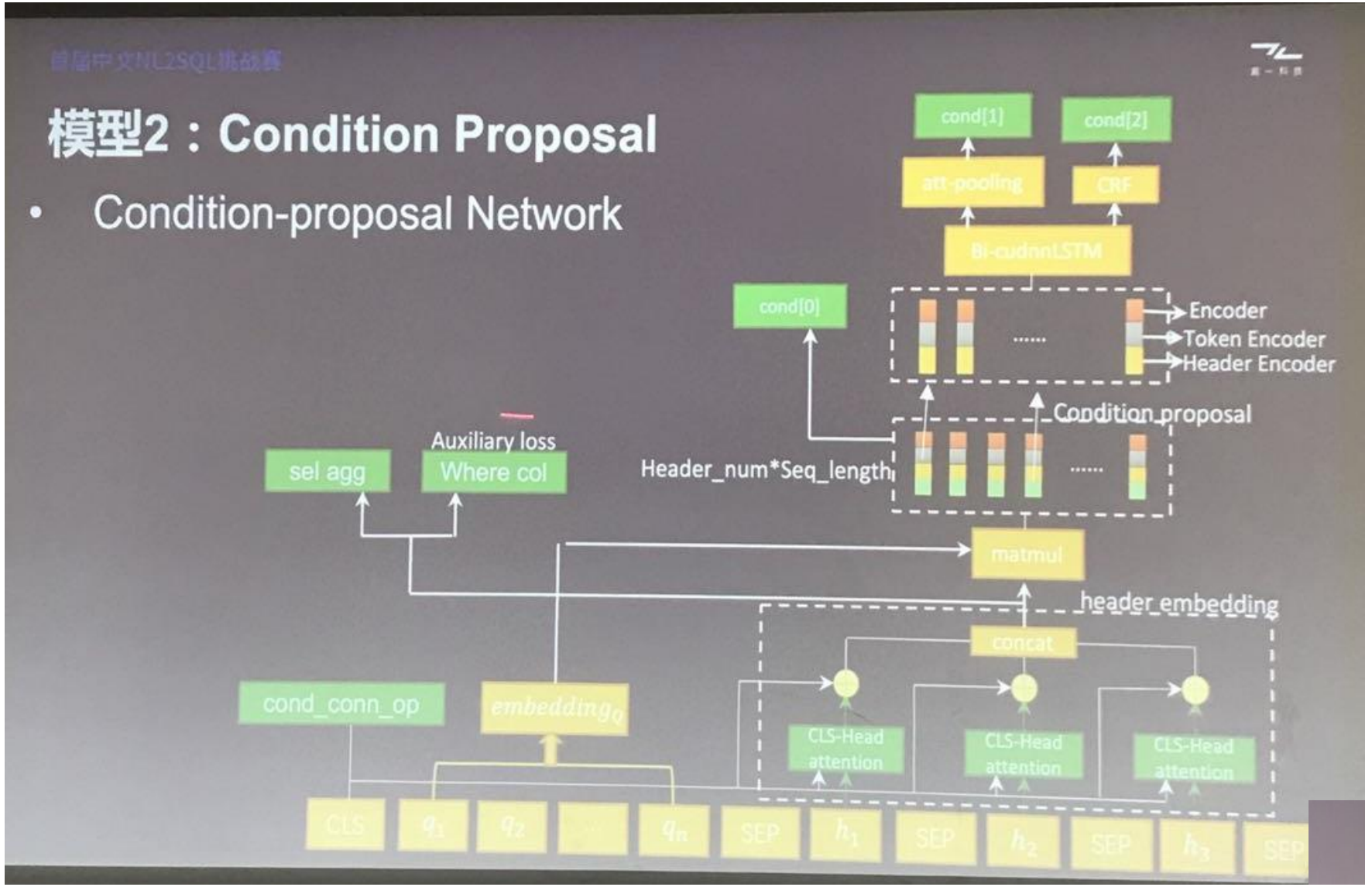
X-SQL网络变体

- 增加sel_num：准确率达到99.5%
- cond_conn_op：通过CLS预测
- sel-agg：通过rcs同时预测
- column-opt：通过rcs同时预测
- Value：不抽取



模型2：Condition Proposal

- Condition-proposal Network



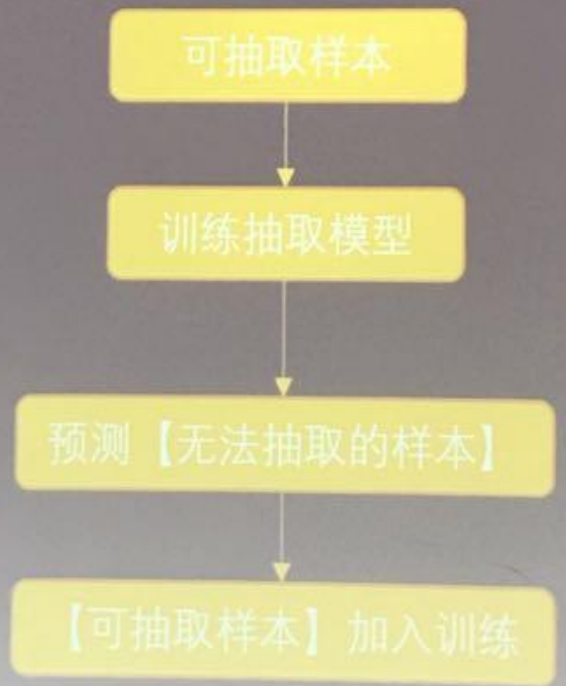
补充数据

- 标注校正（蒸馏）

10000+样本存在这样的问题

【问句】
MF381航班10点起飞，11:20降落，该航班飞的是哪条线路

【原始标注】
[2, 2, "1000-1120"] ×
[0, 2, "MF381"] √

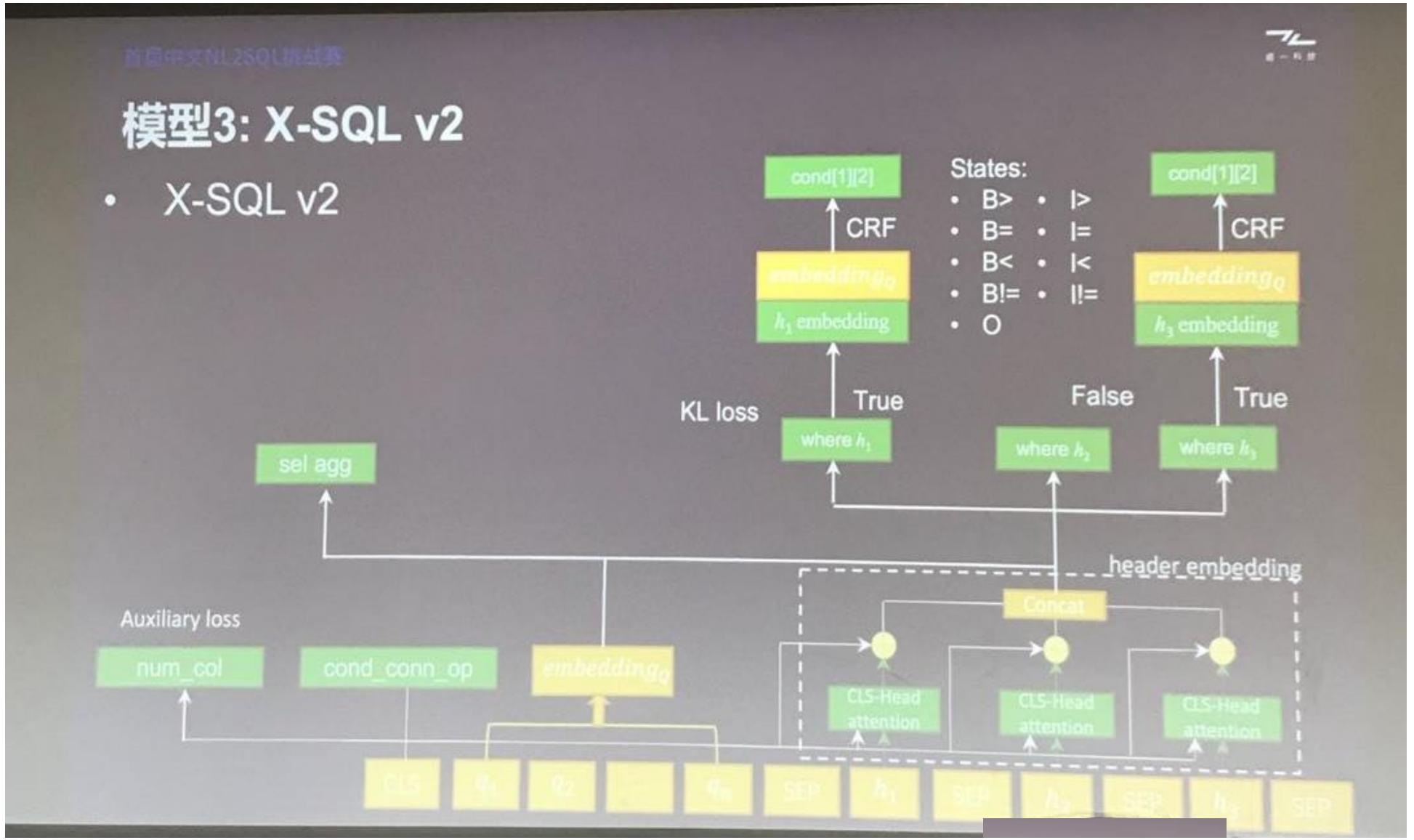


【问句】
MF381航班10点起飞，11:20降落，该航班飞的是哪条线路

【处理后】
[2, 2, " 10点起飞，11:20降落"] √
[0, 2, "MF381"] √

模型3: X-SQL v2

- X-SQL v2



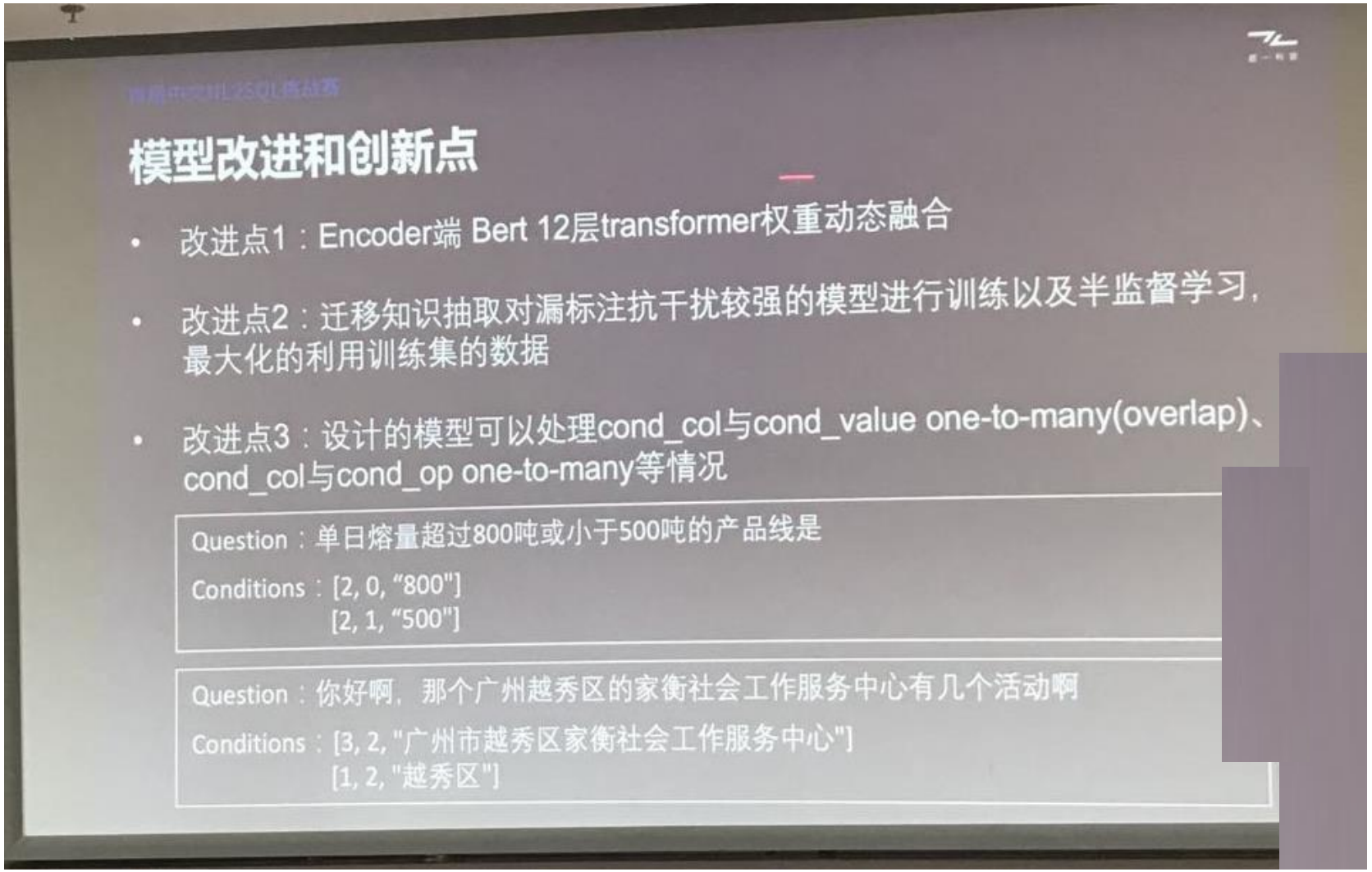
模型融合

- 基于概率平均和投票的模型融合算法
  - sel_num：x-SQL概率平均
  - select-agg：三个模型按sel-agg组合投票（选票数最多的select_num个）
  - cond_conn_op：三个模型的投票
  - conds：两个模型按column-opt-column组合投票

模型	sel-agg	sel_num	cond_conn_op	conds
X-SQL	6	6	6	0
Condition-proposal	4	0	4	4
X-SQL v2	1	0	1	1
阈值	6	（概率融合）	6	2
score	sel=0.9795 agg=0.9838	0.9956	0.9741	column=0.9472 opt=0.9561 value=0.9231

后处理

- 对cond的处理
  - 根据列的类型，过滤抽取得到的值；
  - 剔除相互冲突的条件；
  - 针对cond[1]等于2的条件，从table找到最接近的值
    - real类型：在数值上做比较
    - text类型：使用Levenshtein Distance度量序列之间差异
  - 当cond[1]等于2时，修正条件值不同行的情况
- 根据cond的情况，对冲突的cond_conn_op的重置
- 对cond数目大于2的情况，进行execute guide


模型改进和创新点
• 改进点1：Encoder端 Bert 12层transformer权重动态融合
• 改进点2：迁移知识抽取对漏标注抗干扰较强的模型进行训练以及半监督学习，最大化的利用训练集的数据
• 改进点3：设计的模型可以处理cond_col与cond_value one-to-many(overlap)、cond_col与cond_op one-to-many等情况
Question：单日熔量超过800吨或小于500吨的产品线是
Conditions：[2, 0, “800"]
[2, 1, “500"]
Question：你好啊，那个广州越秀区的家衡社会工作服务中心有几个活动啊
Conditions：[3, 2, "广州市越秀区家衡社会工作服务中心"]
[1, 2, "越秀区"]

模型的泛化性

- 模型对于训练集未出现的表格的识别能力

ALL VALID	UNSEEN VALID
Execution：0.9313 Logic Form: 0.8953 Score: 0.9133	Execution：0.9339 Logic Form: 0.8869 Score: 0.9104

Question：铁旗门这部电影一共有多少集

Label: {'agg': [0], 'cond_conn_op': 0, 'sel': [4], 'conds': [[1, 2, '铁旗门']]}

Pred: {'sel': [4], 'agg': [0], 'cond_conn_op': 1, 'conds': [[1, 2, '铁旗门'], [3, 2, '电影']]}

Milestone

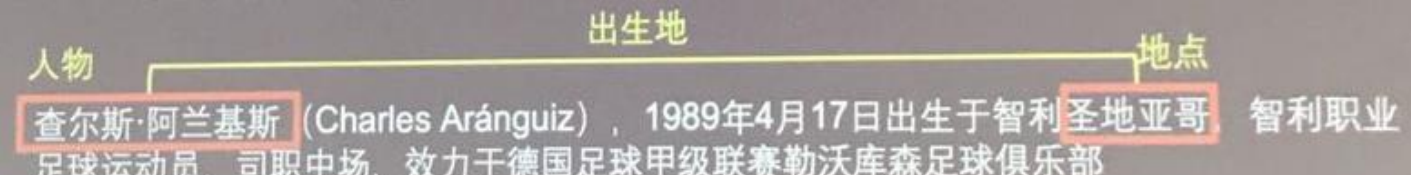
模型	Valid分数	TestA分数	TestB分数
Baseline	0.8030	0.8228	
12层transformer动态权重融合	0.8241	0.8365	
针对不能匹配的数据，采用预测数据作为label	0.8546	0.8582	
使用X-SQL的rcs与hidden states点乘作为proposal	0.8703	0.8802	
shuffle columns	0.8872	0.8952	
模型融合	0.9027		0.9078
融合model1（X-SQL）	0.9119		0.9132
融合model3（X-SQL v2）	0.9133		0.9143

方案优势

- 模型创新性
  - 12层transformer动态权重融合；
  - 通过列的shuffle实现数据增强；
  - 迁移知识抽取领域基于proposal的模型，模型能够较强的抵抗漏标注的噪声，并且利用半监督学习，召回10000+条数据。提高cond值的抽取效果 。
  - 在X-SQL的基础上，增加辅助任务，构建多任务联合学习的方案

方案优势

- 模型通用性：适用于多种信息抽取任务
  - 知识图谱的抽取：



型通用性：适用于多种信息抽取任务

知识图谱的抽取：

出生地

地点

斯·阿兰基斯 (C

运动员，司职中

圣地亚哥

部

2019语言与智能技术竞赛

Rank	Model	Precision	Recall	F1
1	[知识工场] BERT(ensemble) gdm 复旦大学	0.8975	0.8886	0.893
2	[variant bert+multi head selection] (ensemble) littlebert 个人	0.8962	0.8886	0.8924
3	[ERNIE CTagging + MultiSub Reviewer] (ensemble) Kill_Thread Ecole X	0.8976	0.8852	0.8914
4	good luck(ensemble) 格物致知 国双科技	0.8948	0.8858	0.8903

方案优势

- 模型通用性：适用于多种信息抽取任务
  - 电商评论观点的抽取：

这家店快递很快，就是大小不太合适，勉强用吧。

AspectTerms	OpinionTerms	Polarities	Categories
快递	很快	正面评价	物流
大小	不太合适	负面	尺寸

方案优势

- 模型创新性
  - 提出condition proposal和X-SQL v2模型实现nl2sql端到端解析
- 模型通用性好
  - 适用于多种信息抽取任务（如nl2sql、三元组抽取、属性观点词联合抽取）
  - 模型获得之江大赛“电商评论挖掘”复赛第一名
- 模型实用性
  - 单模型为端到端的pipeline，10min时间内完成11个模型预测