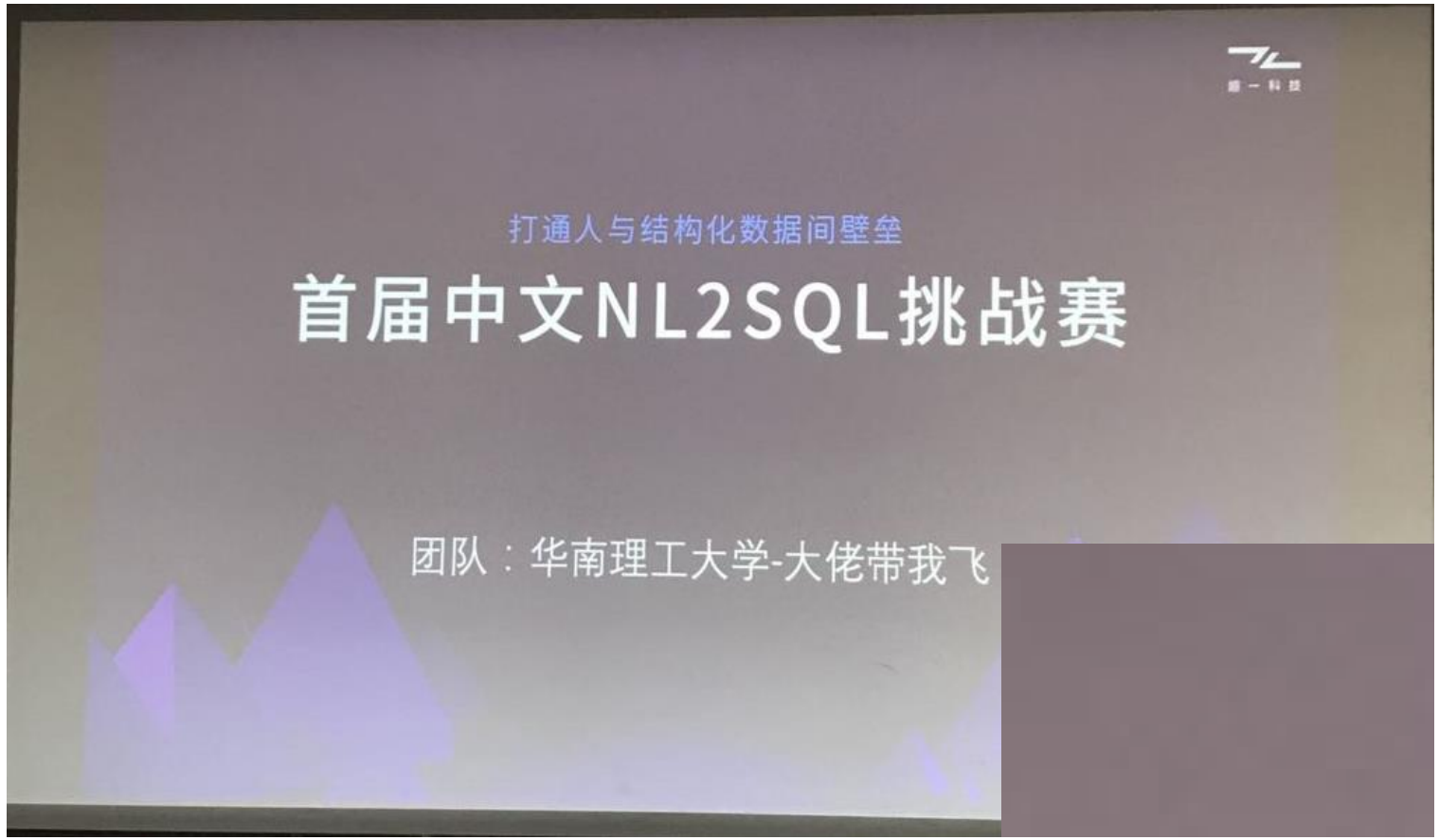

打通人与结构化数据间壁垒
首届中文NL2SQL挑战赛
团队：华南理工大学-大佬带我飞

首届中文NL2SQL挑战赛

问题描述：

对于给定通用领域的表格数据和自然语言，生成与之相对应的SQL (Structured Query Language)语句。

例子：

Question: 问一下风华企业周涨跌幅不超过10%，它月涨跌幅和年涨跌是多少

Table:

股票代码	公司简称	周涨跌	月涨跌	年涨跌
000636.sz	风华高科	-0.68	22.98	-0.85
600303.sh	曙光股份	-2.06	-1.28	-52.64
002454.sz	松芝股份	0.46	-8.86	-42.11
002048.sz	宁波华翔	-0.79	-0.04	-52.17
002214.sz	大立科技	3.49	4.02	-28.92

SQL: Select 月涨跌,年涨跌 from Table where 公司简称 = 风华高科 and 周涨跌 < 10

Pipeline：

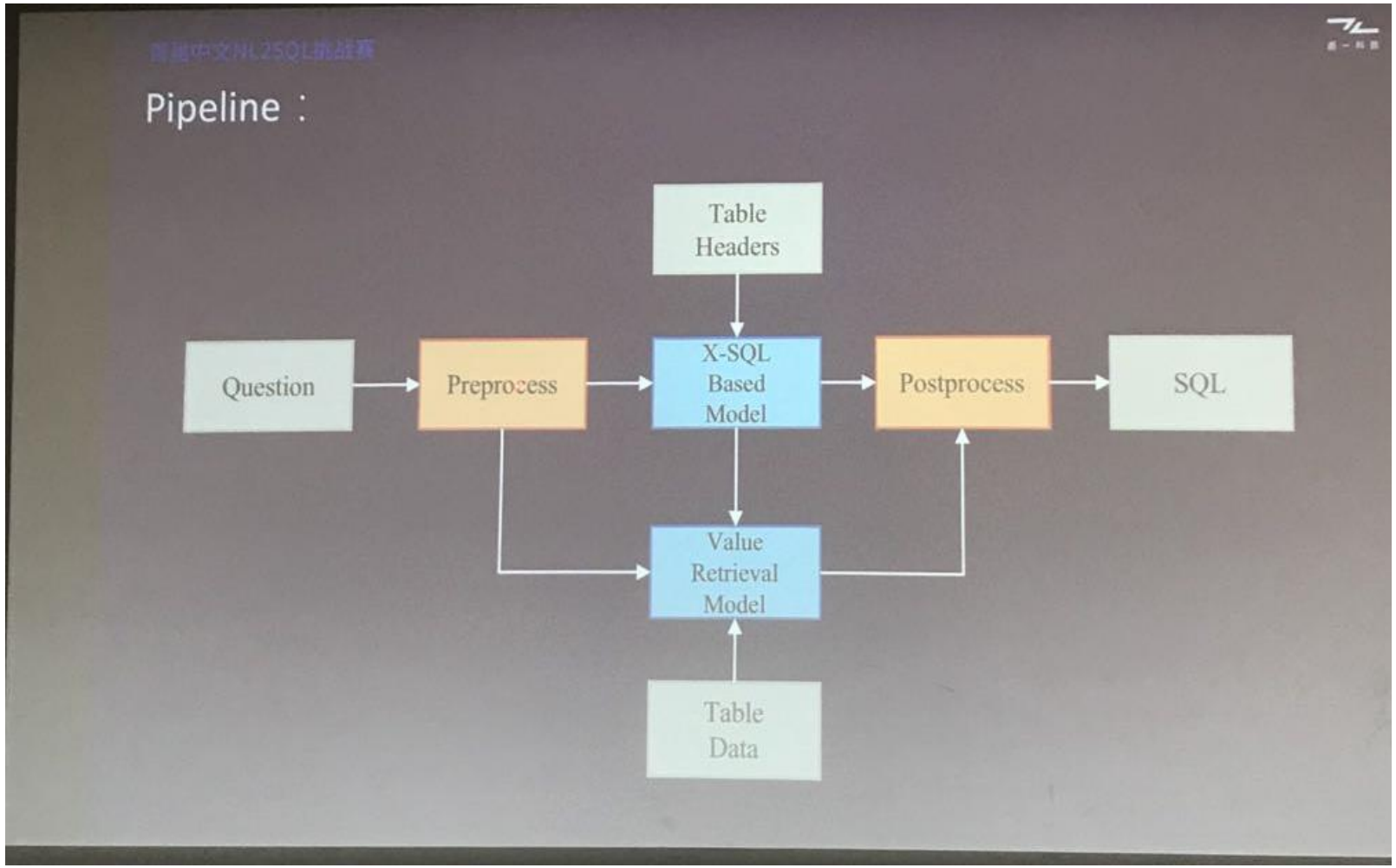
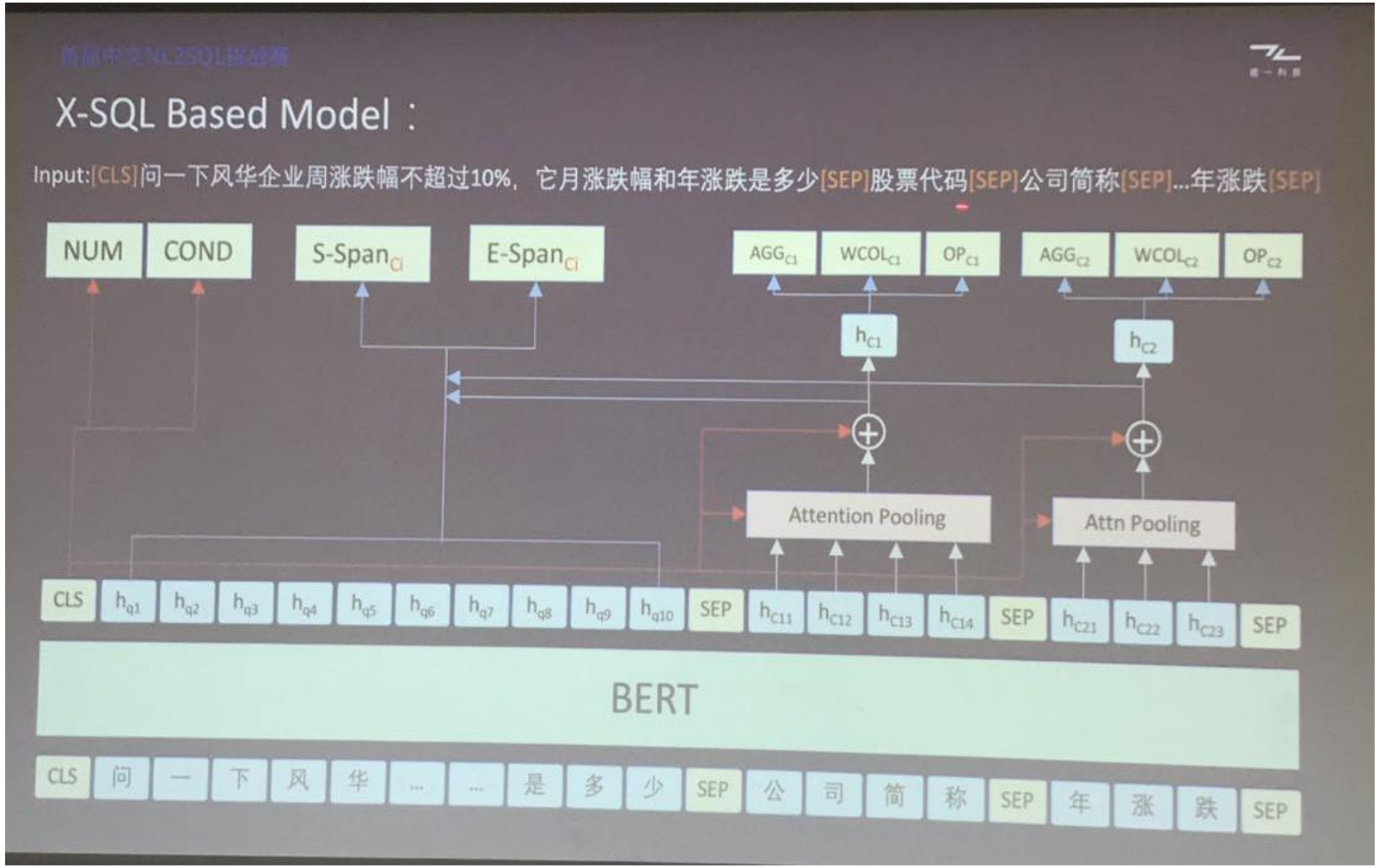
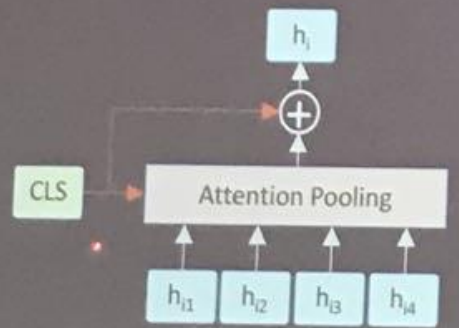


Table Headers
Question
Preprocess
X-SQL Based Model
Postprocess
SQL
Value Retrieval Model
Table Data

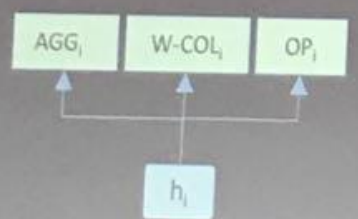

首届中文NL2SQL挑战赛
X-SQL Based Model：
Input:[CLS]问一下风华企业周涨跌幅不超过10%，它月涨跌幅和年涨跌是多少[SEP]股票代码[SEP]公司简称[SEP]...年涨跌[SEP]
NUM
COND
S-Span$_{Ci}$
E-Span$_{Ci}$
AGG$_{C1}$
WCOL$_{C1}$
OP$_{C1}$
AGG$_{C2}$
WCOL$_{C2}$
OP$_{C2}$
h$_{C1}$
h$_{C2}$
Attention Pooling
Attn Pooling
CLS
h$_{q1}$ h$_{q2}$ h$_{q3}$ h$_{q4}$ h$_{q5}$ h$_{q6}$ h$_{q7}$ h$_{q8}$ h$_{q9}$ h$_{q10}$
SEP
h$_{C11}$ h$_{C12}$ h$_{C13}$ h$_{C14}$
SEP
h$_{C21}$ h$_{C22}$ h$_{C23}$
SEP
BERT
CLS 问 一 下 风 华 ... ... 是 多 少 SEP 公 司 简 称 SEP 年 涨 跌 SEP

Attention Pooling：



$$\alpha_{it} = softmax\left(W_q h_{cls} \cdot (W_k h_{it})^T\right)$$

$$h_i = h_{cls} + \sum_t \alpha_{it}\, h_{it}$$



for OP predictions, using independent logistic classifiers instead of softmax.

Span Prediction：

Original X-SQL: $P_{start}^{val}(q_j|C_i) = softmax\ g(Uh_{q_j} + Vh_{C_i})$

$$\text{argmax}_j\ g\left(Uh_{q_j} + Vh_{C_i}\right)$$
$$= \text{argmax}_j\ g\left(Uh_{q_j}\right) + g\left(Vh_{C_i}\right)$$
$$= \text{argmax}_j\ g\left(Uh_{q_j}\right)$$

In the original X-SQL, we will get the same span prediction, even for different columns.

周涨跌幅不超过10，年涨跌不超过5

Logit(h_q): 1, 2, 1, 1, 1, 2, 1, 6, 0, 1, 2, 1, 1, 1, 2, 4

首届中文NL2SQL挑战赛

Span Prediction：

Modification:

$$P_{start}^{val}(q_j|C_i)$$

$$= softmax\ g(Uh_{q_j} + Attn\left(h_{q_j}, h_{C_i}\right)h_{q_j})$$

or

$$P_{start}^{val}(q_j|C_i) = softmax\ g(Uh_{q_j} + V[h_{q_j}; h_{C_i}])$$

Span Prediction：

Modification:

$$P_{start}^{val}(q_j|C_i)$$

$$= softmax\ g(Uh_{q_j} + Attn\left(h_{q_j}, h_{c_i}\right)h_{q_j})$$

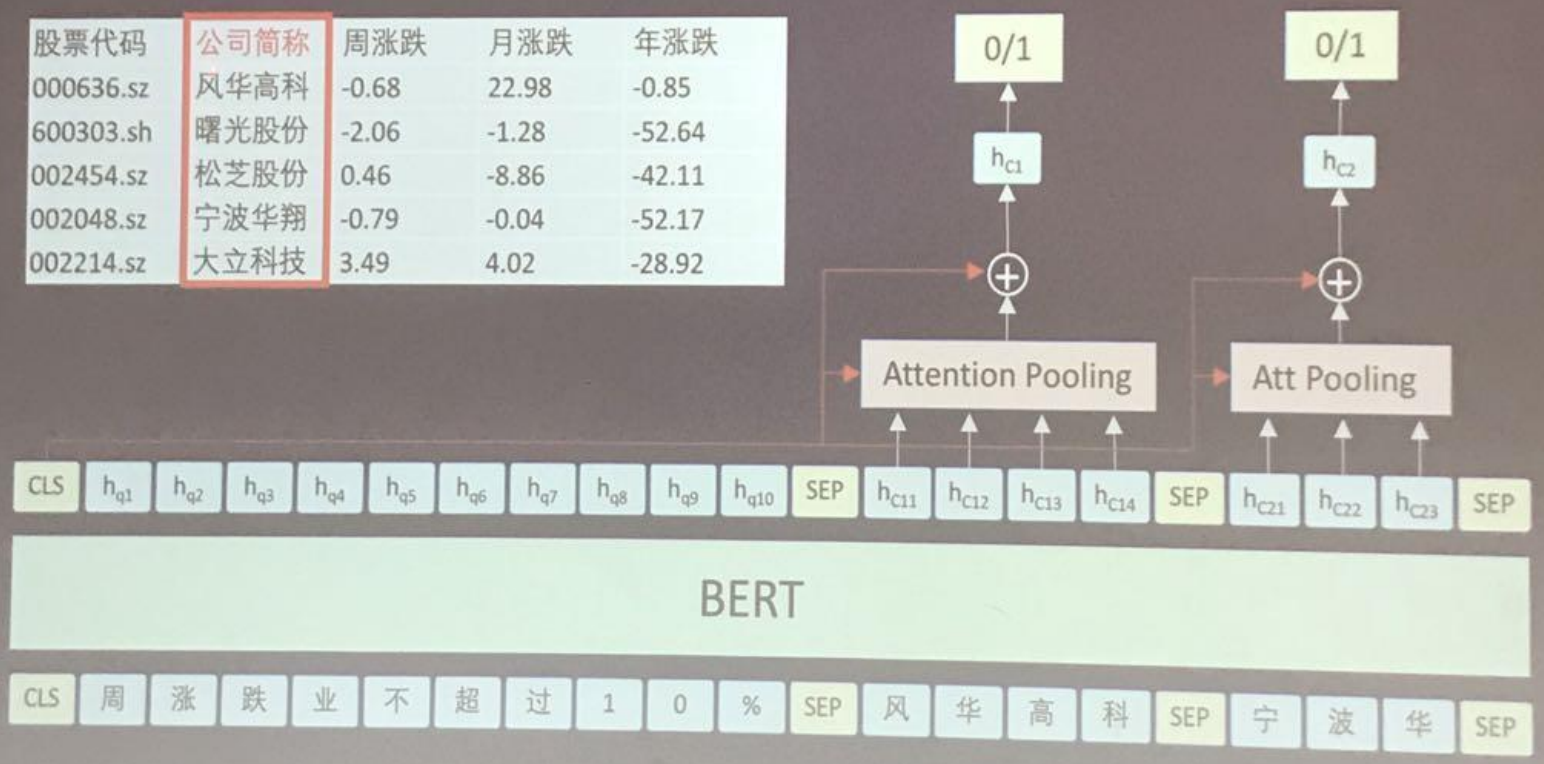
or

$$P_{start}^{val}(q_j|C_i) = softmax\ g(Uh_{q_j} + V[h_{q_j}; h_{C_i}])$$

Value Retrieval Model：

Input:[CLS]问一下风华企业周涨跌幅不超过10%，它月涨跌幅和年涨跌是多少[SEP]风华高科[SEP]宁波华翔……大立科技[SEP]

股票代码	公司简称	周涨跌	月涨跌	年涨跌
000636.sz	风华高科	-0.68	22.98	-0.85
600303.sh	曙光股份	-2.06	-1.28	-52.64
002454.sz	松芝股份	0.46	-8.86	-42.11
002048.sz	宁波华翔	-0.79	-0.04	-52.17
002214.sz	大立科技	3.49	4.02	-28.92



Postprocess（Execution-guided）：

W-Value:

Real: span prediction

Text: value retrieval

Aggregation:

Real: " ", COUNT ⟷ MASK [1,0,0,0,1,0]

Text: " ", AVG, MAX, MIN, COUNT, SUM

Future Work：

- 数值单位的统一。（如：五角=0.5元 500米=0.5公里）
- 一些先验知识的关联和引用。（如：情人节=2月14日）
- 引入一些NER的模型和预处理，提高Value Retrieval Model的性能。