MCA-Reader

北京信息科技大学智能信息处理实验室

目 录

CONTENTS

数据处理

01

数据处理

1、分词及答案标注

（1）先分词，再标注

原始信息：

文章：安卡拉战役，是塞琉古帝国内战-兄弟战争的一场决定性战役。

答案：塞琉古帝国

分词结果：

文章：[安卡拉 战役 ， 是 塞 琉古 帝国 内战 - 兄弟 战争 的 一场 决定性 战役 。]

答案：[塞琉古 帝国]

答案的分词结果与文章中答案的分词结果不一致

01

数据处理

1、分词及答案标注

（2）先标注，再分词

原始信息+标注：

文章：安卡拉战役，是##塞琉古帝国##内战-兄弟战争的一场决定性战役。

答案：塞琉古帝国

分词结果：

文章：[安卡拉 战役，是 ## 塞琉古 帝国 ## 内战 - 兄弟 战争 的 一场 决定性 战役 。]

答案：[塞琉古 帝国]

保证可以准确地标注到答案，但是局限于训练集和验证集

2、未登录词处理

（1）分词

原始文章：莱昂德罗·内托是巴西足球运动员，司职前锋。

分词：[莱昂 德罗 · 内托 是 巴西 足球 运动员 ， 司职 前锋 。]

ltp命名实体识别：莱昂德罗·内托/Nh 巴西/Ns

原始文章+NER标注： ##莱昂德罗·内托##是##巴西##足球运动员，司职前锋。

分词：[## 莱昂德罗·内托 ## 是 ## 巴西 ## 足球 运动员 ， 司职 前锋 。]

进一步提升了分词的准确性

2、未登录词处理

（2）表示

假设：相同实体类别词的词向量是相似的，因此可以为每种实体类别训练一个词向量。在具体实现中，我们先将文本做如下，再利用w2v做词向量训练。

原始文章：莱昂德罗·内托是巴西足球运动员，司职前锋。

ltp命名实体识别：莱昂德罗·内托/Nh 巴西/Ns

命名实体替换： **Nh**是**Ns**足球运动员，司职前锋。

分词：**Nh** 是 **Ns** 足球 运动员 ， 司职 前锋 。

一定程度上解决了未登录词表示问题

3、手工特征提取

词级别共现与字级别共现。若某个词在文章和对应的问题中同时出现，则将对应位置的词级别特征值置为1，否则置为0；字级别特征更加精细，计算方式为某一词中的字共现的次数除以词的长度。

词级别特征：

文章：['《', '铁拳', '男人', '》', '是', '一部', '由', '朗·霍华德', '执导', '的', '影片', '。']

[1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0]

问题：['《', '铁拳', '男人', '》', '的', '导演', '是', '谁', '？']

[1, 1, 1, 1, 1, 0, 1, 0, 0]

字级别特征：

文章：['《', '铁拳', '男人', '》', '是', '一部', '由', '朗·霍华德', '执导', '的', '影片', '。']

[1, 1, 1, 1, 1, 0, 0, 0, 0.5, 1, 0, 0]

问题：['《', '铁拳', '男人', '》', '的', '导演', '是', '谁', '？']

[1, 1, 1, 1, 1, 0.5, 1, 0, 0]

模型介绍

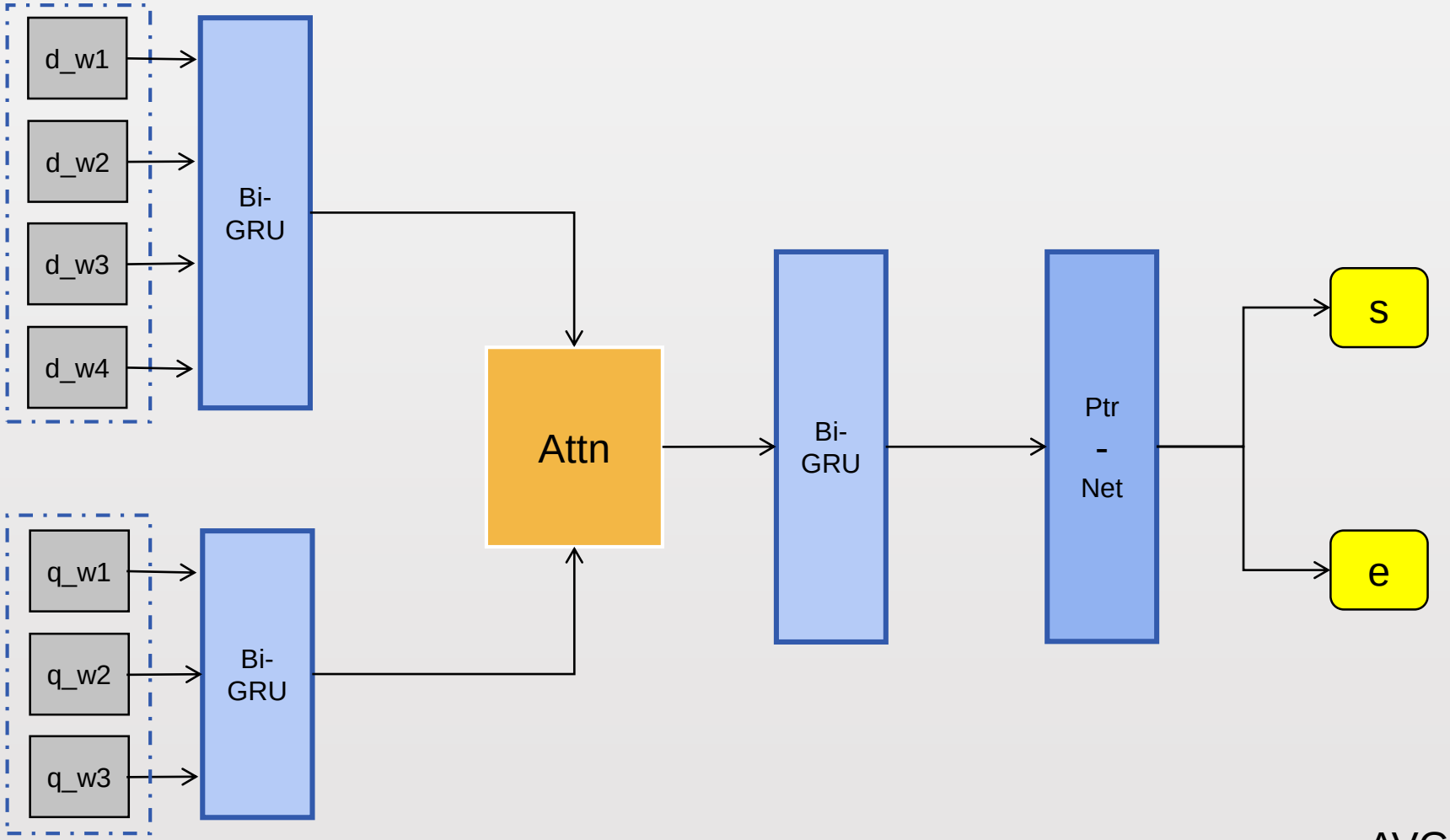
02

模型进化史

Model	AVG (Dev)	△
Base-Model	55.875	-
+ Layer1 (Attn and Encode)	56.423	+ 0.548
+ Dense connect	59.935	+ 3.512
+ Word level feature	62.420	+ 2.485
+ Data amplification (CMRC2017) + UNK word replace	64.265	+ 1.845
+ Layer2 (Attn and Encode)	65.015	+ 0.750
+ Char level feature	67.059	+ 2.044
+ Post processing	68.397	+ 1.338
+ Data amplification (DRCD) + word segmentation with NER	72.711	+ 4.314
+ ensemble (5 models)	74.697	+ 1.986

02

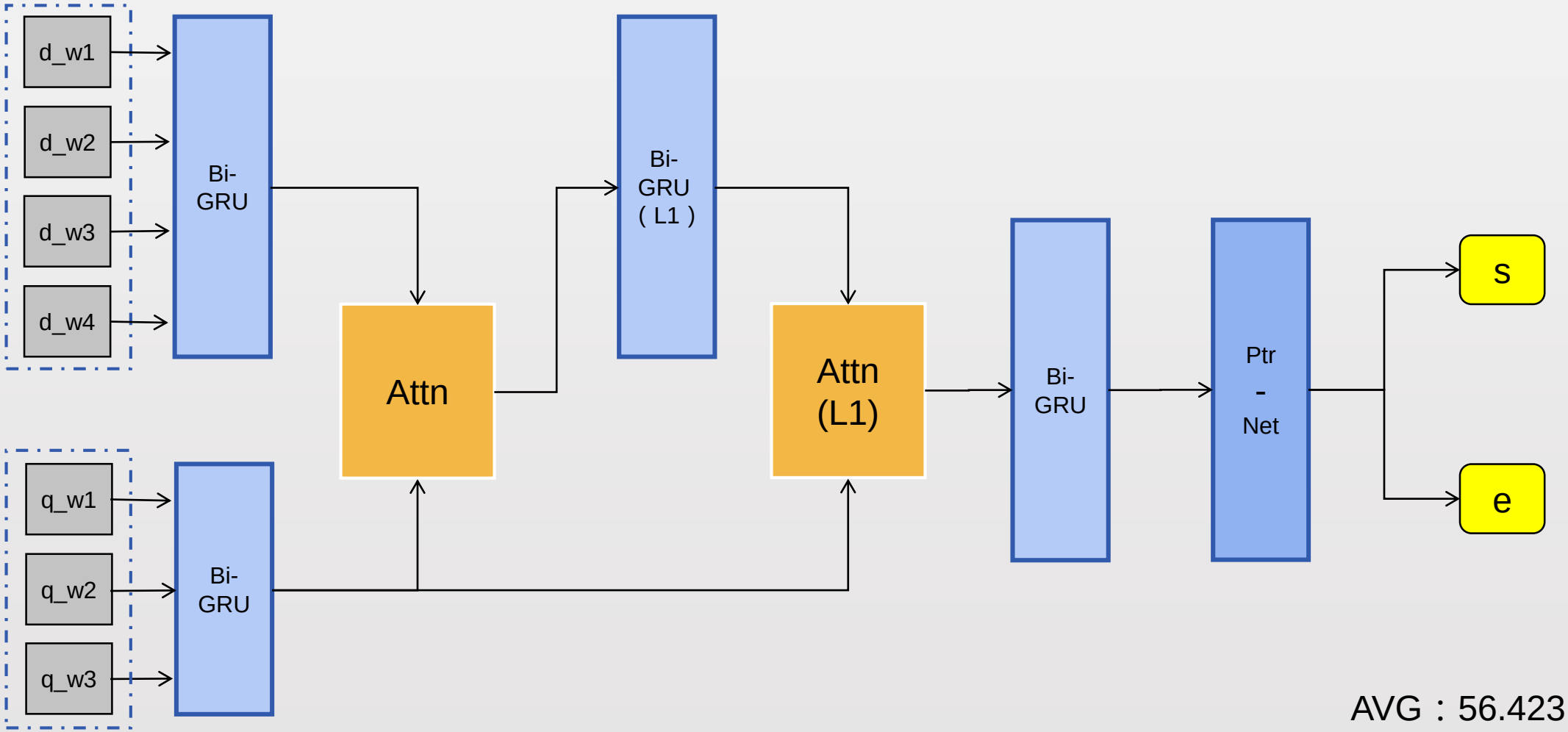
Base Model



AVG：55.875

02

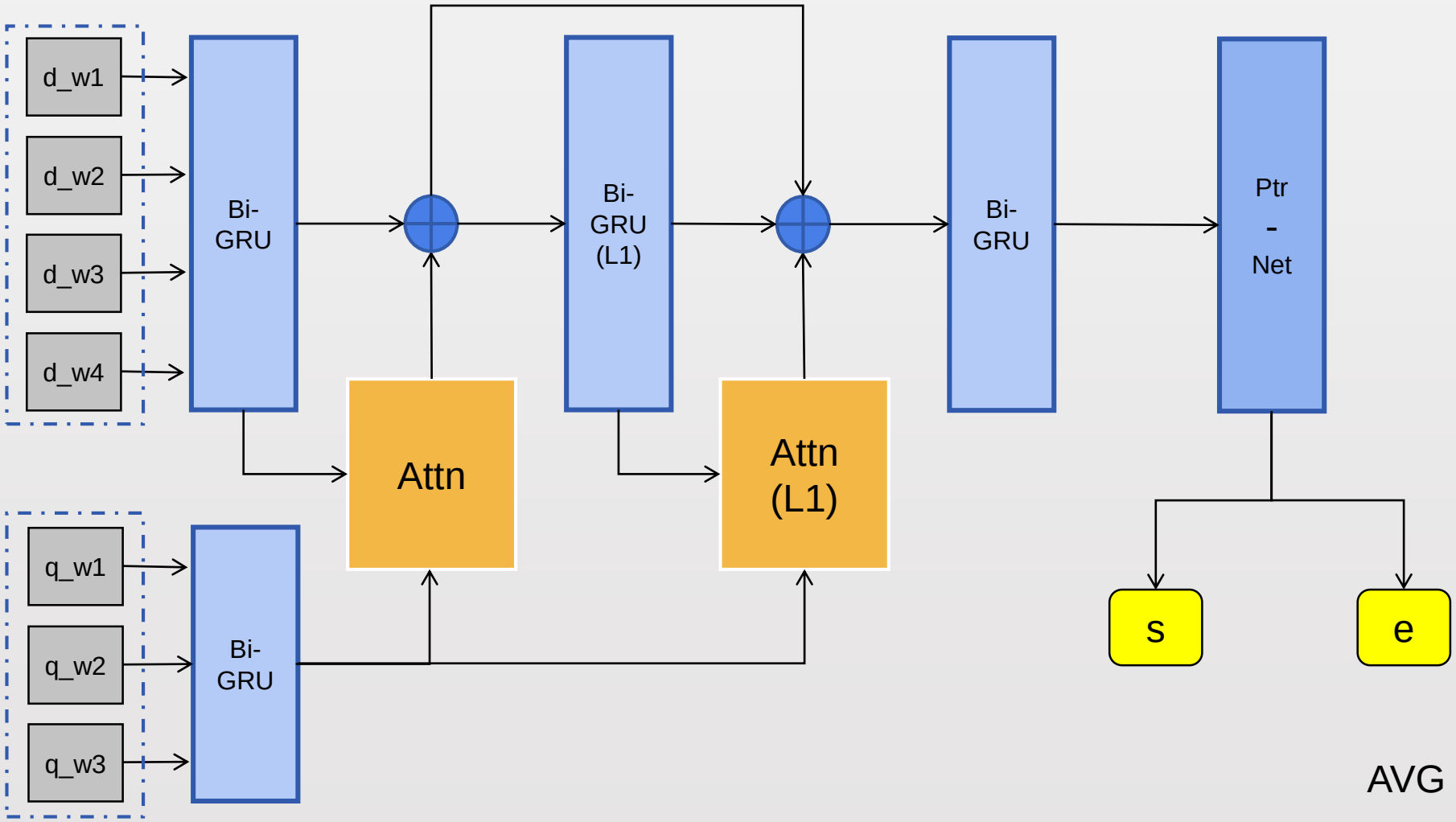
+Layer1



AVG：56.423

02

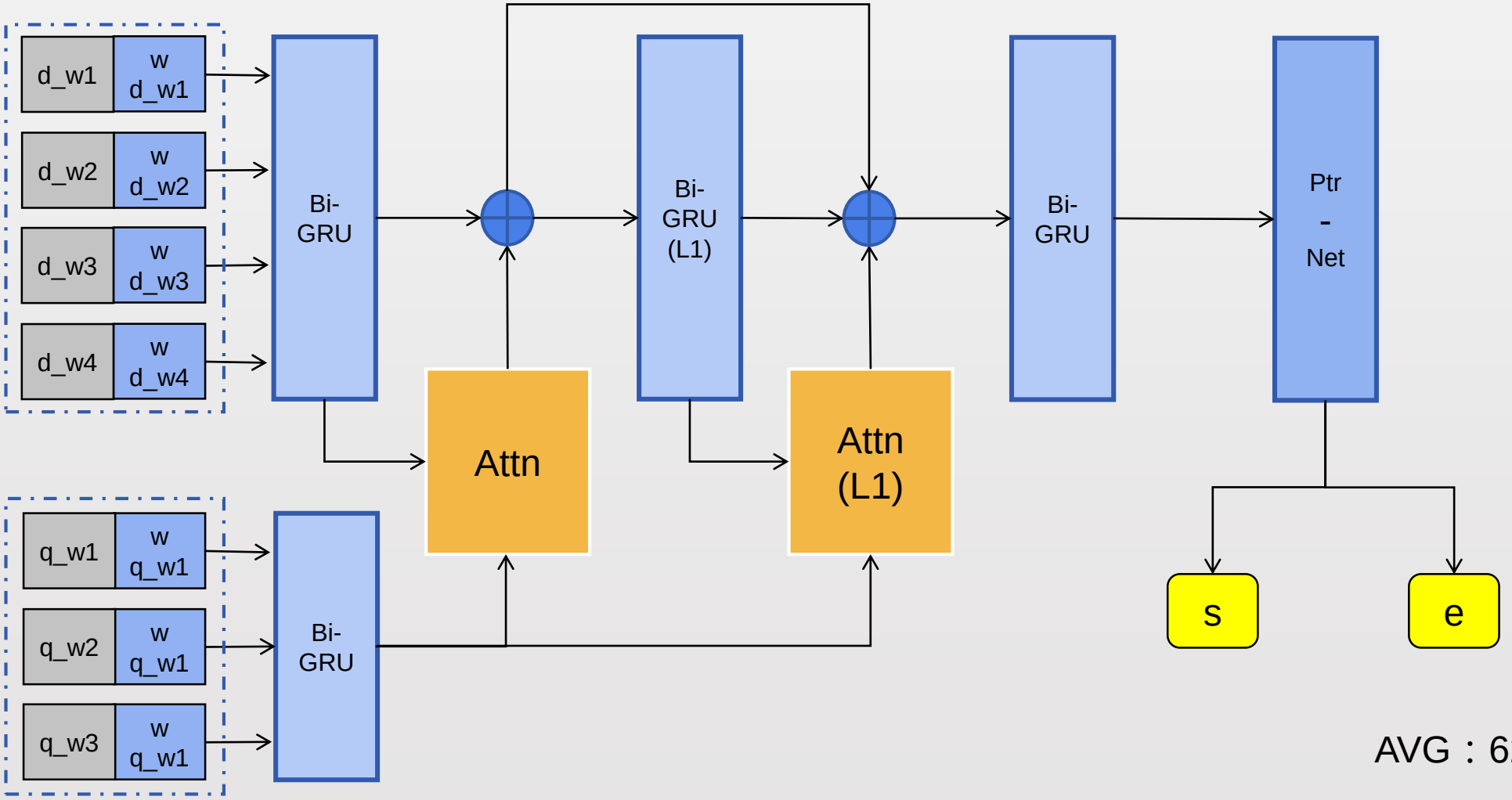
+ Dense connect



AVG：59.935

02

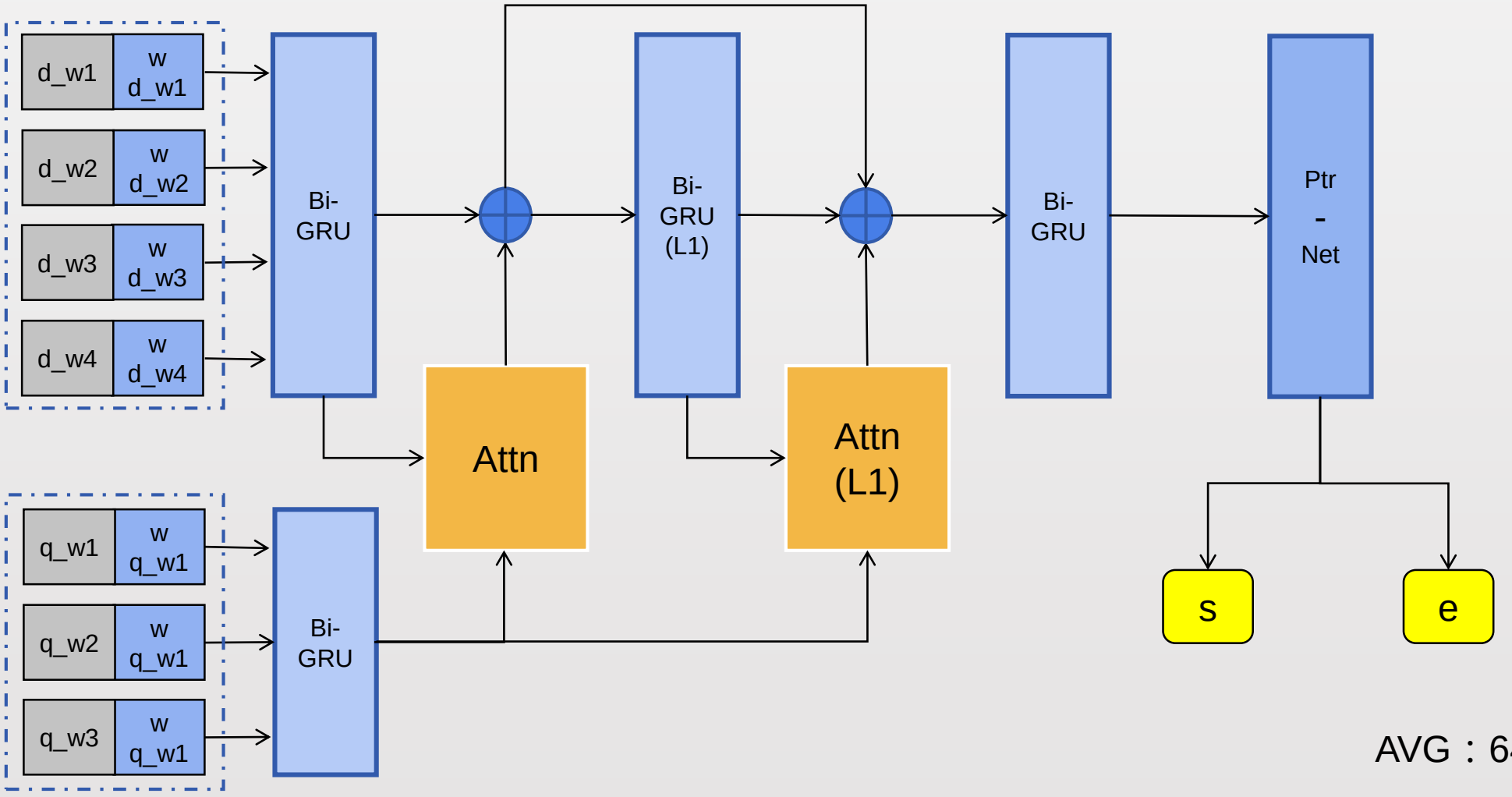
+ Word level feature



AVG：62.420

02

+ Data amplification(CMRC2017)
+ UNK word replace

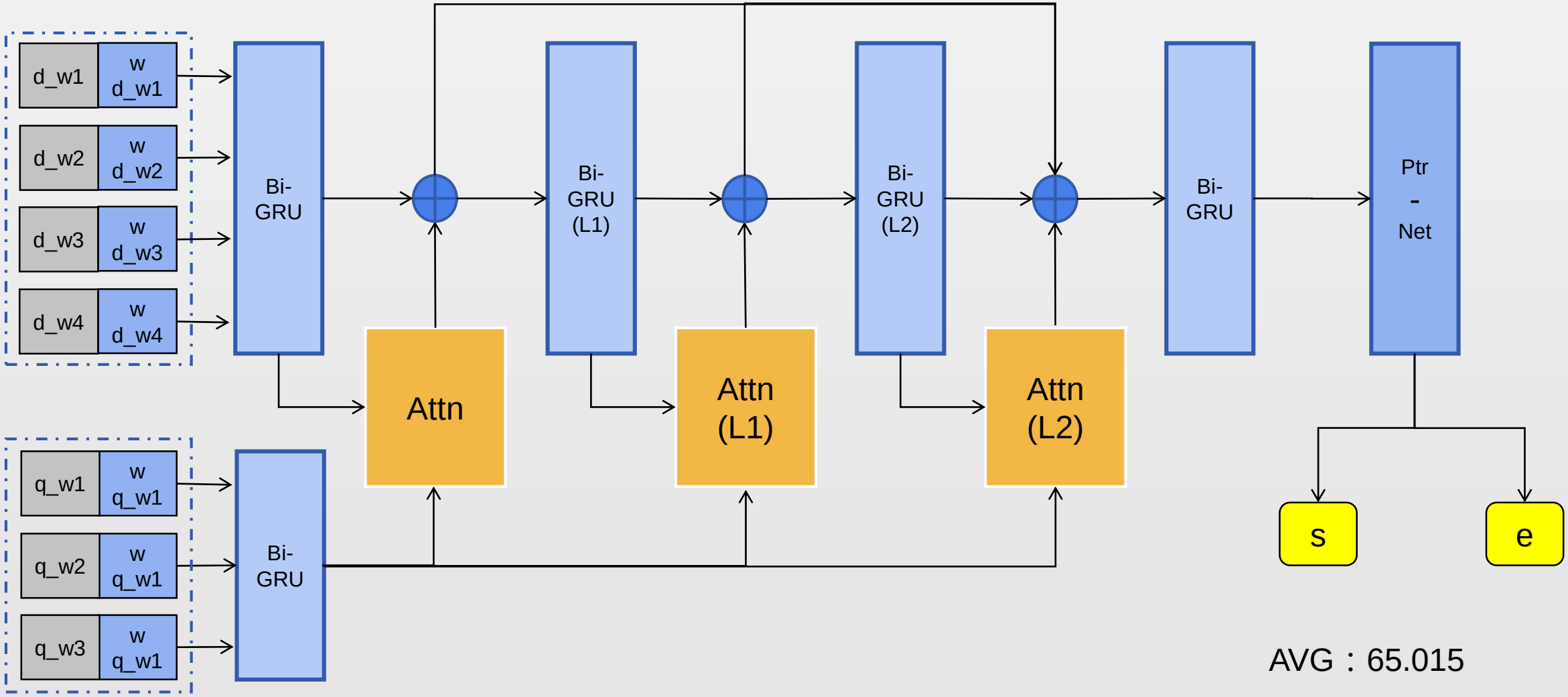


AVG：64.265

https://github.com/ymcui/cmrc2017

02

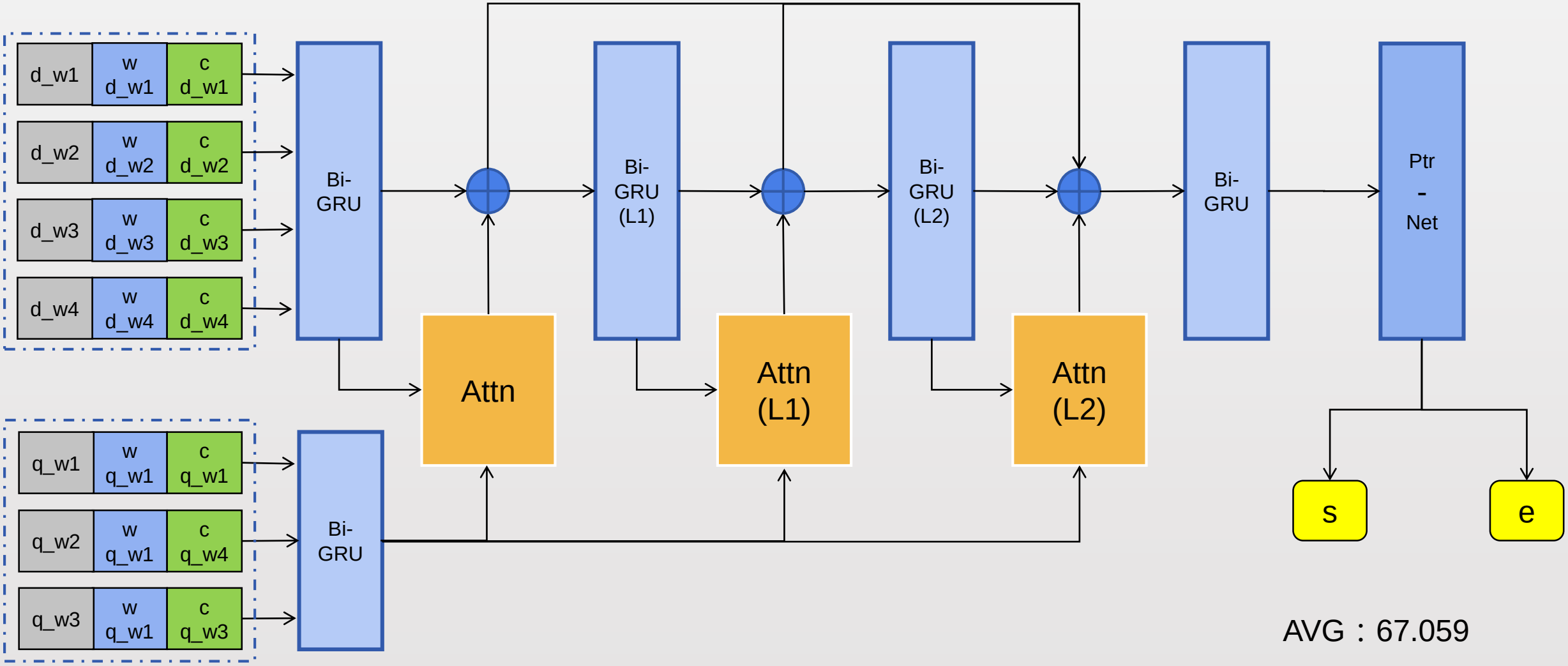
+ Layer2



AVG : 65.015

02

+ Char level feature



AVG：67.059

02

+ Post processing

例：

森林北部有动植物公园（Jardin d'acclimatation）、民间艺术与传统博物馆（Musée des Arts et Traditions populaires），东北濒临马约门广场 (place de la Porte de Maillot)，西北部有训练场（Champ d'entraînement）、运动场（Terrain de Sports）和巴加特勒公园 (Parc Bagatelle)，西南部有隆尚跑马场 (Hippodrome de Longchamp)。

问题：

进行中文分词后，会忽略掉原文中英文部分的空格。在给出最终输出结果时，若缺少空格，则会判定为预测不完全正确，EM值为0。

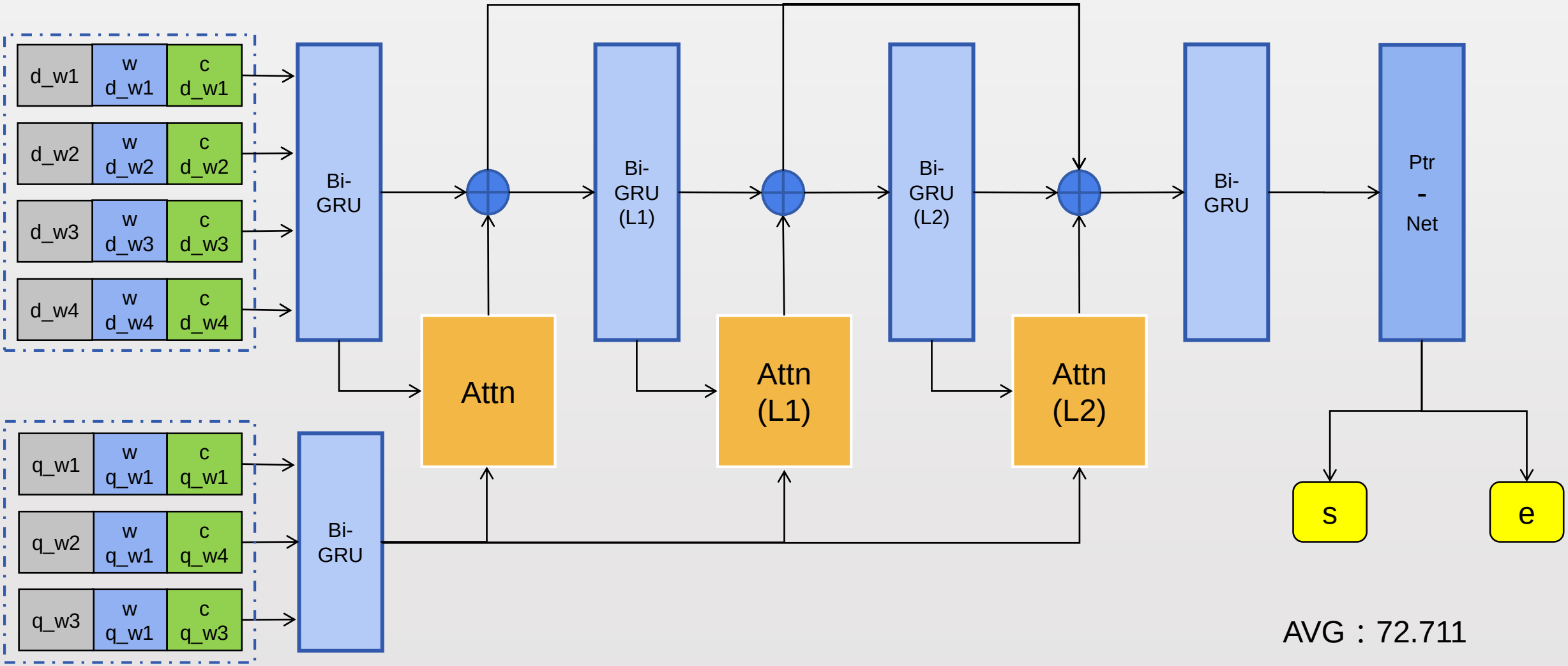
解决办法：

预测起始位置后，将范围内的词使用 ”\s*” 进行拼接，将拼接后的结果带回原文中，查找是否有匹配的句子，将匹配的结果进行输出。

AVG：68.397

02

+ Data amplification (DRCD)
+ word segmentation with NER

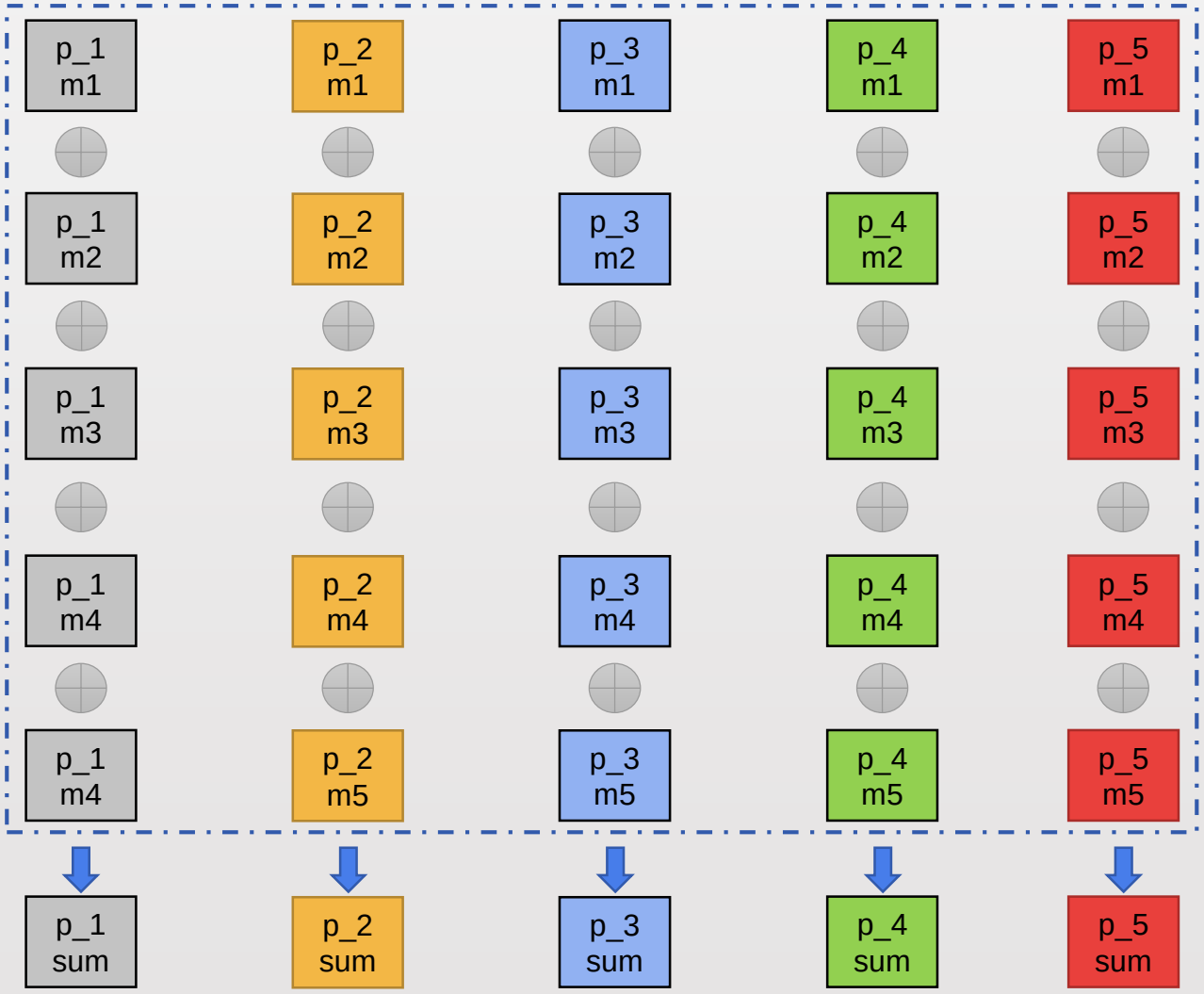


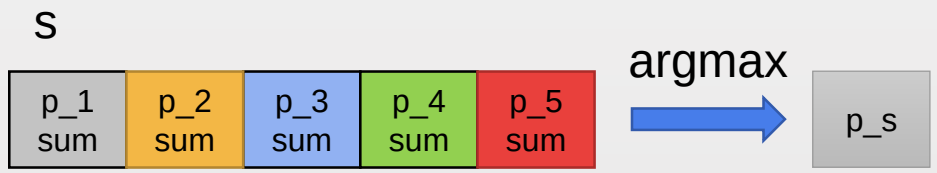
AVG：72.711

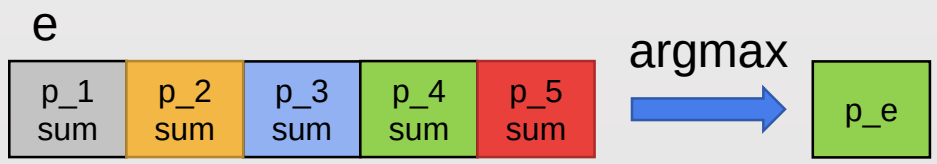
https://github.com/DRCSolutionService/DRCD

02

+ Ensemble







AVG : 74.697

02

最终排名

排名	时间	系统名称	Pre-Test	Test		
			Average	EM	F1	Average↓
1	2018/9/17	**Z-Reader (single)** ZhuiYi	**81.608**	**74.178**	**88.145**	**81.161**
2	2018/9/17	**MCA-Reader (ensemble)** 北京信息科技大学智能信息处理实验室	79.147	71.175	88.090	79.632
3	2018/9/17	**RCEN (ensemble)** 6ESTATES PTE LTD	77.978	68.662	85.743	77.203
4	2018/9/17	**MCA-Reader (single)** 北京信息科技大学智能信息处理实验室	76.696	68.335	85.707	77.021
5	2018/9/17	**OmegaOne (ensemble)** 复旦大学	74.232	66.272	82.788	74.530
6	2018/9/17	**RCEN (single)** 6ESTATES PTE LTD	75.228	64.576	83.136	73.856
7	2018/9/17	**GM-Reader (ensemble)** CIST - 北京邮电大学	73.720	64.045	83.046	73.546
8	2018/9/17	**OmegaOne (single)** 复旦大学	72.475	64.188	81.539	72.864
9	2018/9/17	**GM-Reader (single)** CIST - 北京邮电大学	70.615	60.470	80.035	70.252
10	2018/9/17	**R-NET (single)** SXU - 山西大学	62.976	50.112	73.353	61.733

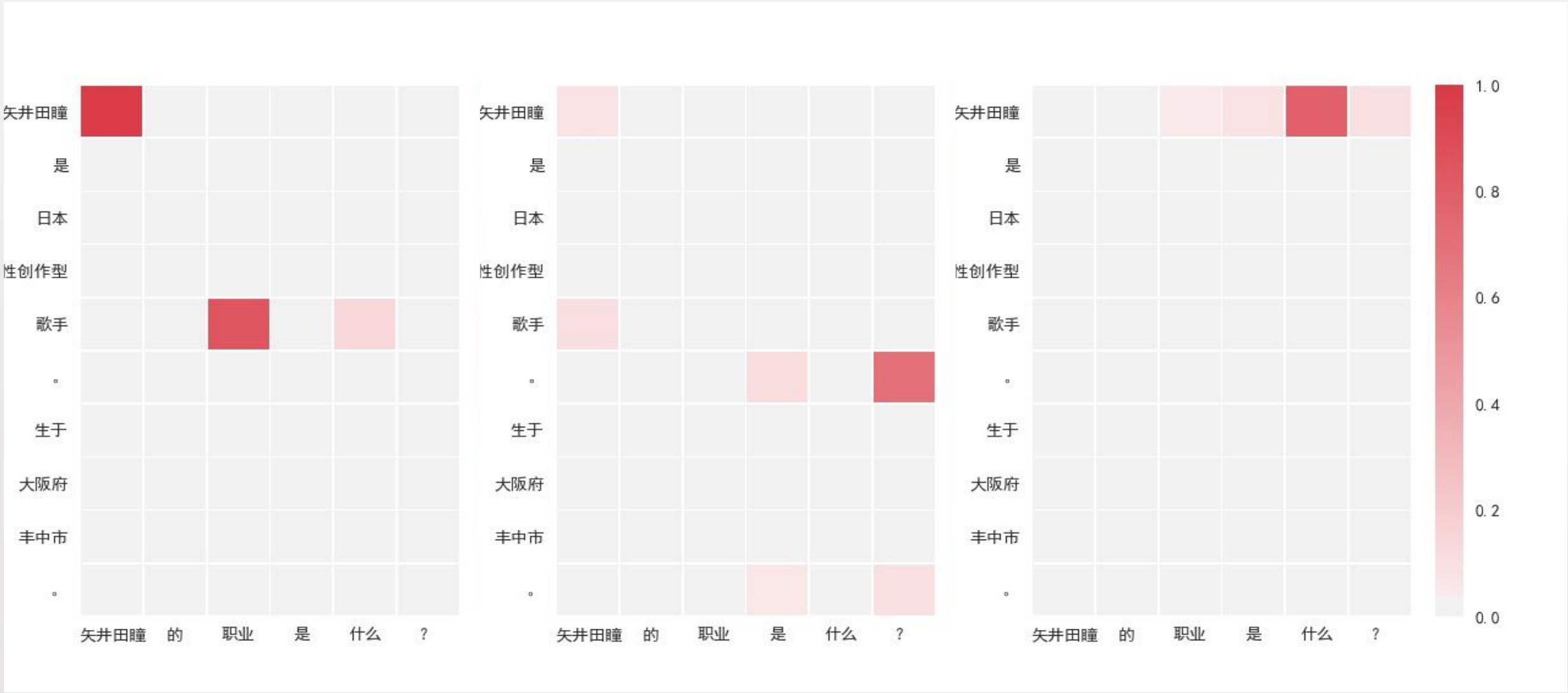
模型分析

03

模型分析

03

模型分析



03

模型分析


哪个
哪些
谁
为什么
哪
哪部
多少
什么
哪里
哪年
哪位
如何
哪几个
几种
多少岁
几届
事情
一类
几年
时间
职业
哪条
哪一部
地方
怎么样
哪国
多长
哪一
几个
怎么
哪家
几枚
多少倍
大
什么样
时候
怎样
哪些地方
多久
几位
一种
两种
几层
三种
写作
为何
朝代
第几
几张
几辆
几条
几岁
几座
几架
由谁来
类型
学科
动物
国人
一档
一时期
一项
一场
一段
季节
第几名
几班
几辑
几级
几月
几批
几项
几部
几号
几场
几分
几只
几名
几球
几岁
几站
几轮
几类
几列
颜色
哪句
哪篇
哪几种
哪一位
哪些因素？
哪些方面
哪场
哪次
哪种
哪项
多长时间
多少度
多重
东西
官职
书
那个
单位
有何
行业
意思
啥
远
一首歌
篇章
篇文章
时代
时期
朝
物质
第几代
疾病
两位
两个
两大
角色
由来
干什么
何时
何
大学
大约
人
女朋
怎么办
一话
什么游戏
还是
种
是

欢迎大家提问