Suda&Alibaba at CGED-7: Ensembles of Error Detection and Correction Models for Chinese Grammatical Error Diagnosis (E2DC)

李嘉诚[1]，沈嘉钰[1]，包祖贻[2]，章波[2]，章岳[1]，李辰[2]，李正华[1]

1. 苏州大学 人类语言技术研究所

2. 阿里巴巴 达摩院

评测介绍

- 错误分类：CGED-7评测将中文语法错误分为四类。

错误类型	原始句子	正确句子
缺失(M)	每个城市的超市能看到这些食品。	每个城市的超市都能看到这些食品。
冗余(R)	我和妈妈是不像别的母女。	我和妈妈不像别的母女。
替换(S)	最重要的是做孩子想学的环境。	最重要的是创造孩子想学的环境。
乱序(W)	“静音环境”是对人体应该有危害的。	“静音环境”应该是对人体有危害的。

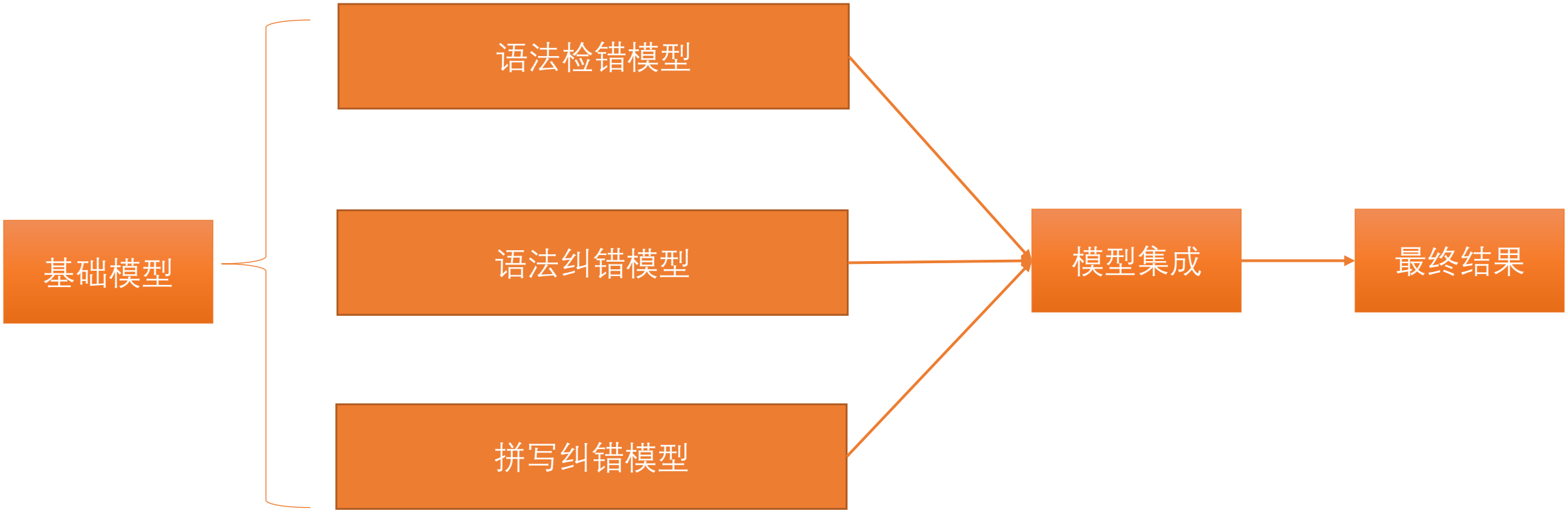
- 评测要求：输入一个可能包含语法错误的中文句子，参赛系统应判断该句子是否包含语法错误，如果包含了语法错误，要求给出该语法错误在句子中的位置以及错误类型，对误用(S)和缺失(M)需要给出对应的纠正结果。
- 举例：

输入句子：1000 我根本不能了解这妇女辞职回家的现象。在这个时代，为什么放弃自己的工作，就回家当家庭主妇。

输出结果：1000，6，7，S，理解

1000，8，8，R

系统总架构



语法检错模型

- 模型：我们把CGED任务看作一个序列标注任务，采用基于BIOES的BERT-CRF模型。
- 举例：

输入	星	期	五	我	上	汉	语	果	。	晚	上	我	有	约	会	跟	朋	友	。
输出	O	O	O	O	O	O	O	S-S	O	O	O	O	B-W	I-W	I-W	I-W	I-W	E-W	O

合并错误类型一致的标签

输出结果：

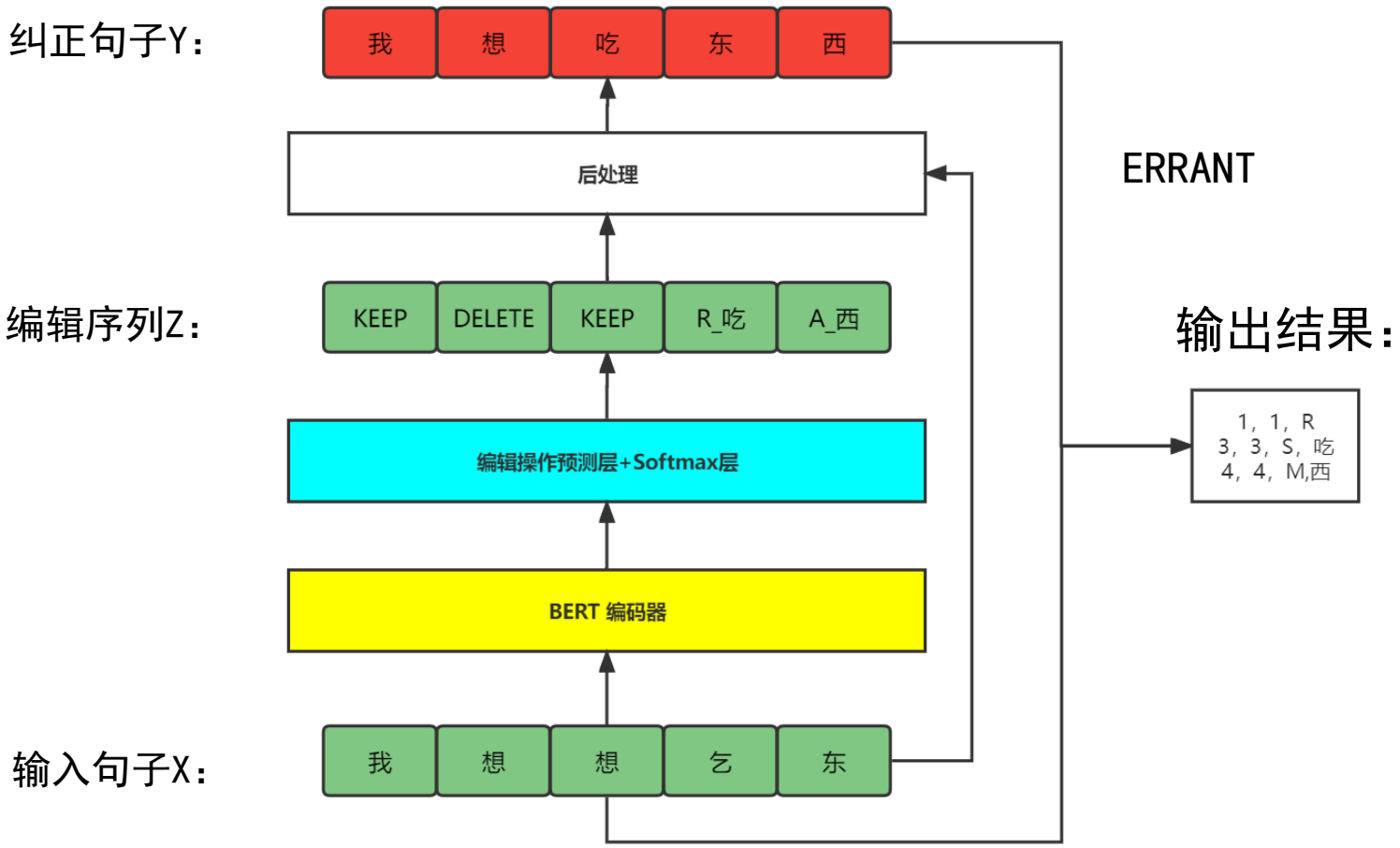
1001 7,7, S
1001 12,17, W

语法检错模型-数据转换

- 现状：
  - CGED任务的数据：
    - 标注了错误点的开始，结束位置，错误类型，可转化为语法检错模型训练所需要的BIOES格式的数据。
    - 数据量较少，2014-2021年7届CGED比赛，总共开源5万句左右数据。
  - CGEC任务的数据：
    - 错误-正确平行句对。
    - 数据量相对较多，Lang8（120万）；HSK（15万）。
- 思路：
  - 我们使用ERRANT（错误编辑抽取工具）将相对数据较多的CGEC任务的数据转换为CGED任务检错模型所需要的格式，使用该数据对检错模型进行预训练。

语法纠错模型

- 模型：基于序列到编辑的GECToR模型。
- 举例：

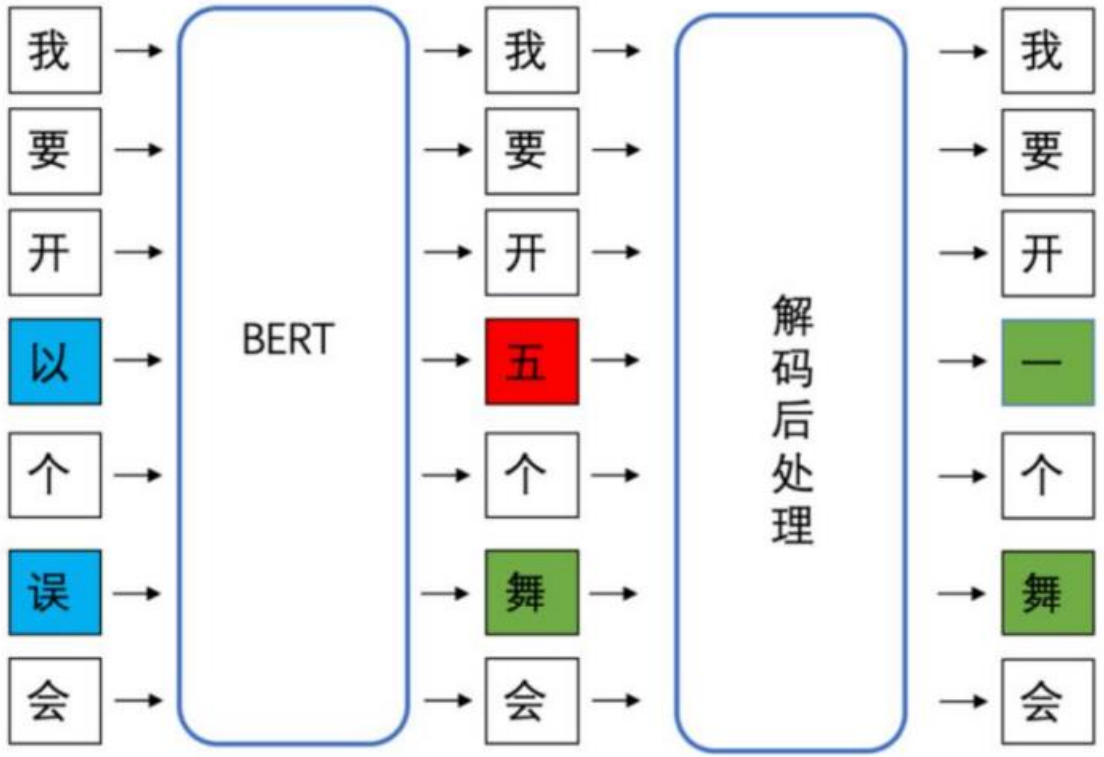


语法纠错模型-数据增强

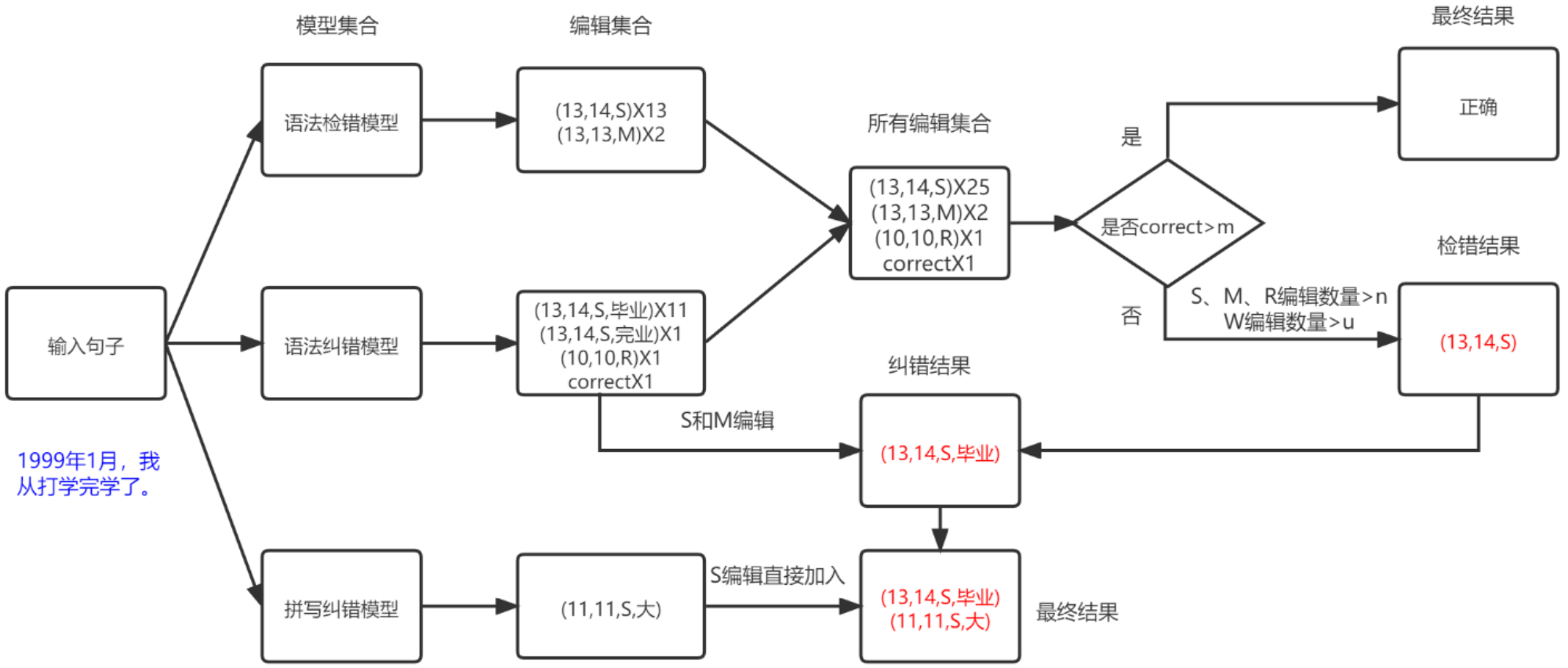
- 方法：将基于混淆集的规则数据增强和基于反向翻译的神经网络数据增强结合起来生成人造数据。
  - 基于混淆集的规则数据增强：赋予模型解决已有训练集中没出现过的语法错误的能力。
  - 基于反向翻译的神经网络（GECTOR模型）数据增强：模拟现有的二语学习者的数据分布，赋予模型更好地解决二语学习者语法错误的能力。

拼写纠错模型

- 模型：基于BERT的序列标注模型。
- 举例：



模型集成-基于编辑级别投票的模型集成



实验-检错模型（开发集结果）

- 开发集：选择CGED2020年的测试集作为开发集

	检测层			识别层			定位层		
检错模型	P	R	F	P	R	F	P	R	F
GED	**91.88**	88.61	90.22	**73.18**	57.30	64.27	44.53	32.49	37.57
GED+PT	88.60	**92.61**	**90.56**	72.53	**63.65**	**67.80**	**44.76**	**37.83**	**41.00**

- 模型：
  - GED：检错模型在CGED的数据上训练。
  - GED+PT：检错模型先在CGEC数据集转换后的数据上预训练，然后在CGED的数据上微调。
- 结论：
  - 采用CGEC转换后的数据预训练的检错模型比基准模型在三个指标中都有较大的提升。

实验-纠错模型（开发集结果）

	检测层			识别层			定位层			修正层		
纠错模型(概率平均集成)	P	R	F	P	R	F	P	R	F	P	R	F
CSC	**97.22**	27.39	42.74	**91.36**	13.72	23.86	**68.42**	7.11	12.88	**51.98**	7.35	12.88
GEC	92.02	84.26	87.97	70.01	54.10	61.03	46.61	30.66	36.99	30.08	18.98	23.28
GEC(3Ens.)	91.16	80.70	85.61	72.63	52.29	60.80	51.78	31.46	39.14	32.63	20.33	25.05
GEC+PT	90.00	**86.87**	**88.41**	70.28	**57.67**	**63.36**	46.85	**34.00**	39.40	29.54	**22.34**	25.44
GEC+PT(4Ens.)	92.20	81.22	86.36	73.82	52.16	61.13	53.15	31.43	**39.50**	35.55	21.15	**26.52**

- 模型：
  - CSC：拼写纠错模型。
  - GEC：语法纠错模型在CGEC的数据上训练。
  - GEC+PT：语法纠错模型先在人造数据上预训练，然后在CGEC的数据上微调。
  - GEC(nEns.)：对n个采用不同预训练模型初始化的语法纠错模型，采用基于概率平均的模型集成方法进行模型集成。
- 结论：
  - 拼写纠错模型在四个指标上的准确率都很高，
  - 使用人造数据预训练的语法纠错模型，比基准模型在四个指标上都有一定的提升。
  - 不管是否使用数据增强，基于概率平均的模型集成的语法纠错模型，在定位层和修正层上均比基准模型高，而在检测层和识别层上，均比基准模型低。

实验-编辑投票模型集成（开发集结果）

总结果：

	检测层			识别层			定位层			修正层		
编辑投票集成	P	R	F	P	R	F	P	R	F	P	R	F
Ensemble	93.01	**91.39**	**92.19**	73.20	70.79	71.98	50.24	49.23	49.73	31.03	28.98	29.97
Ensemble-W	93.01	91.39	92.19	72.95	**71.40**	**72.16**	50.19	**49.59**	**49.89**	31.03	**28.98**	**29.97**

不同错误类型结果：

	替换（S）			乱序（W）			缺失（M）			冗余(R)		
编辑投票集成	P	R	F	P	R	F	P	R	F	P	R	F
Ensemble	54.01	54.28	54.14	**64.62**	38.30	48.09	42.66	43.06	42.86	45.68	48.44	47.02
Ensemble-W	54.01	54.28	54.14	62.05	**42.25**	**50.27**	42.66	43.06	42.86	45.68	48.44	47.02

- 模型：
  - Ensemble：对所有的错误类型采用同一个阈值。
  - Ensemble-W：对W单独设置一个阈值，对另外三种错误类型设置同一个阈值。
- 结论：
  - 使用Ensemble-W方法虽然在整体性能上提高不大，但却显著提高了对W错误的检测能力（F值提高了两个点左右）。

实验-评测结果

- 成绩： 我们提交的系统在三个指标上(识别层；定位层；修正层)得到了第一名，在一个指标(检测层)得到了第三名。
- 具体指标：

TOP3

检测层

	A	B	C	D	E
1	Team	Runs	Pre	Rec	F1
2	NJUNLP	run1	0.8442	0.9141	0.8778
3	LDU	run2	0.819	0.9406	0.8756
4	NJUNLP	run3	0.813	0.9368	0.8706
5	S&A	run1	0.8706	0.8673	0.869
6	S&A	run2	0.8677	0.8699	0.8688
7	S&A	run3	0.8677	0.8699	0.8688
8	LDU	run3	0.9489	0.7985	0.8672
9	NJUNLP	run2	0.879	0.8534	0.866
10	hitmitlab	run1	0.8182	0.8983	0.8564

TOP1

识别层

	Team	Runs	Pre	Rec	F1
2	S&A	run1	0.4961	0.5368	0.5157
3	S&A	run3	0.4932	0.5353	0.5134
4	S&A	run2	0.4931	0.5353	0.5133
5	YYDS	run2	0.5406	0.4757	0.5061
6	YYDS	run3	0.5669	0.4485	0.5008
7	YYDS	run1	0.4945	0.5053	0.4998
8	NJUNLP	run2	0.4823	0.4435	0.4621
9	KS-NLP	run1	0.4465	0.4785	0.462
10	KS-NLP	run3	0.4507	0.4713	0.4608

	A	B	C	D	E
1	Team	Runs	Pre	Rec	F1
2	S&A	run1	0.698	0.7229	0.7102
3	S&A	run2	0.6956	0.7225	0.7088
4	S&A	run3	0.6956	0.7225	0.7088
5	NJUNLP	run1	0.6461	0.7419	0.6907
6	NJUNLP	run2	0.7154	0.6614	0.6874
7	YYDS	run1	0.6641	0.7081	0.6854
8	YYDS	run2	0.7031	0.6623	0.6821
9	NJUNLP	run3	0.6001	0.7873	0.6811
10	KS-NLP	run1	0.6537	0.7	0.6761

定位层

TOP1

	Team	Runs	Pre	Rec	F1
1	Team	Runs	Pre	Rec	F1
2	S&A	run3	0.4153	0.2585	0.3186
3	S&A	run2	0.4148	0.2585	0.3185
4	S&A	run1	0.4139	0.2587	0.3184
5	YYDS	run1	0.4806	0.2265	0.3079
6	YYDS	run2	0.5087	0.2167	0.3039
7	YYDS	run3	0.5251	0.2083	0.2982
8	corrector_dl	run2	0.3737	0.2015	0.2619
9	corrector_dl	run3	0.2984	0.228	0.2585
10	X!!	run2	0.3439	0.1979	0.2513

纠正层

TOP1

总结

- 本次评测，我们在三个指标的F值上创造了历史新高。
  - 对于定位层，从2020年的67%提高到71%。
  - 对于识别层，从2020年的40%提高到52%
  - 对于纠正层，从2020年的19%提高到32%

- 我们尝试了利用CGEC的数据来帮助CGED的任务，使得语法检错模型的性能得到较大的提升。

- 我们尝试了将基于规则的数据增强和基于反向翻译的神经网络（GECTOR模型）数据增强结合起来生成人造数据，使得纠错模型的性能得到较大的提升。

- 在模型集成时，我们尝试了一种新的模型集成策略，提高了对乱序错误的检错能力。

Thanks