第七届中文句法错误检测技术评测（CGED-7）

1. 任务背景

在众多教育技术中，面向第二语言的作文自动批改是最有价值也是最能检验自然语言处理应用能力的任务。可使用的汉语二语数据匮乏是制约这一领域技术发展的重要原因。**因此 NLP-TEA 研讨会首创中文句法错误自动检测技术评测活动。**ICCE-2014 面向教育应用的自然语言处理技术研讨会（NLP-TEA）上组织了第一届评测句法错误自动检测评测。第三届评测由北京语言大学、台湾师范大学共同举办。评测活动与当年 COLING2016 会议共同举行。第四届及以后均由北京语言大学举办，与当年 IJCNLP2017 共同举行。

句法错误自动检测评测的目标在于为学界提供一套供检验语言学特征、统计方法和与计算模型的可比较的通用数据平台。技术评测活动也将促进业界学者的经验交流和技术、方法共享，是目前自然语言处理领域推动技术进步、系统性能提升的重要动力。

第五届评测与 ACL2018（2018 年 7 月墨尔本）共同举办，评测增加对错字（词）和缺字（词）类型错误的修改评测。参赛系统需要在这两类错误中提交推荐答案，至多三个，用以评价 Top3 答案推荐的准确率。5043 个错误点和 1562 个无错误测试单元。哈工大-讯飞联合实验室取得最好成绩（36.12%）。

2020 年刚刚结束的第六届评测有 31 支队伍报名，17 支队伍完成评测，11 支队伍提交了论文。各项技术指标较之以往取得长足进步。科大讯飞不再垄断各指标第一。外研社、南京大学、网易有道、阿里巴巴均取得非常优异的成绩。大量新技术的应用使得错误判别赛道首次突破 90%大关，错误定位赛道首次突破 40%大关。

今年我们将在第一届自然语言生成大会（CCNLG）上继续举行第七届 CGED 技术评测。

2. 任务描述

本任务模仿 CoNLL 评测的通行做法，将汉语水平考试（HSK）原始数据中精细的错误分类归并为四类：字符串冗余（R）、字符串缺失（M）、字符串错误（S）和语序错误（W）。评测任务要求参加评测的系统输入中介语句子（群），其中包含有一个或多个错误的错误。参赛系统应判断该输入是否包含错误，并识别错误类型，标记出其在句子中的位置和范围。

3. 评测指标

评测在假阳性评价的基础上，从四个方面以精确率、召回率和 F1 值对系统性能进行评价：

1．侦测层（Detective-level）：对段落单元是否包含错误做二分判断。

2．识别层（Identification-level）：本层子任务为多分类问题，即给出错误点的错误类型。

3．定位层（Position-level）：对错误点的位置和覆盖范围进行判断。

4．修正层（Correction-level）：参赛系统被要求提交针对错误字符串（S）和字符串缺失（M）两种错误类型的修正答案。

4. 训练集与测试集

CGED-7 不再提供新训练集，参赛团队可以采用前六届提供的训练集、测试集（报名后提供）和所有外源性数据资源。测试集为 id 与 utf8 文本（作文语句），输出应包含：本 id 是否正确、错误类型、错误位置等信息，详情将在报名后提供样例。

5. 重要日期

报名时间：2021 年 10 月 20 日-11 月 11 日

测试集发布：2021 年 11 月 12 日 14:00

答案提交：发布时间+3 小时以内

评测结果发布（网站）：2021 年 11 月 15 日

NLGIW2021 评测研讨会：2021 年 11 月 20 日

6. 发起单位

发起单位：北京语言大学

评测委员会：

饶高琦，北京语言大学

荀恩东，北京语言大学

张宝林，北京语言大学

报名邮箱：raogaoqi@blcu.edu.cn

请告知团队单位、团队名称、团队成员、联系人信息

7. 反作弊声明

(1) 参与者禁止注册多账户报名，经发现将取消成绩并严肃处理。

(2) 参与者禁止在指定考核技术能力的范围外利用规则漏洞或技术漏洞等不良途径提高成绩排名，经发现将取消成绩并严肃处理。

(3) 可以接触到赛题相关数据的人员，其提交结果将不计入排行榜及评奖。

8. 交流平台

主办方建立技术讨论群（微信群），供选手讨论、沟通，主办方也将安排工作人员定期在群内答疑，且后续的相关活动信息均会在群内发布。