第一届全国自然语言生成与智能写作技术评测报告

2021 年 6 月到 2022 年 4 月，中国中文信息学会（CIPSC）自然语言生成与智能写作专业委员会（筹）组织了第一届全国自然语言生成与智能写作技术评测。本届评测于 2022 年 4 月 22-23 日首届自然语言生成与智能写作大会（NLGIW 2022）上举行技术评测论坛。专委会主席哈尔滨工业大学赵铁军教授宣布启动第二届评测筹备工作。

一、评测任务

本届评测论坛供发布四项面向技术前沿和领域落地的任务，包括面向事实一致性的生成评测、基于大纲的条件故事生成、图像描述生成评价方法评测与中文句法错误检测技术评测四项内容。详情如表 1 所示。

表 1 评测任务一览（按报名时间先后排序）

标题	单位	任务主席
文本生成一致性评测	清华、哈工大（深圳）、百度	肖欣延（百度 NLP）
基于大纲的条件故事生成评测	清华	黄民烈（清华）
图像描述生成评价方法评测	青海师大、中央民大	李琳（青海师大）
中文句法错误检测技术评测	北语	饶高琦（北语）

1. 文本生成一致性评测

任务简介：我们计划使用三个任务数据集测试参赛系统的生成能力，包括文案生成、摘要生成和问题生成：（1）文案生成根据结构化的商品信息生成合适的广告文案；（2）摘要生成是为输入文档生成简洁且包含关键信息的简洁文本；（3）问题生成则是根据给定段落以及答案生成适合的问题。显然，这三个任务对生成结果的事实一致性均有较高要求。

评测指标：事实一致性指标：由于任务 1 的输入有明确的事实数据，所以采用专门的评估指标，对任务 2、3 采用通用的评估指标，具体如下：

a) 对任务 1，采用 PARENT 指标，同时将生成的句子和参考文本、输入表格信息比较，在兼顾参考答案的同时，评价生成内容是否忠于输入表格的信息

b) 对任务 2、3，利用提前构建的文本蕴含模型，通过衡量参考答案与文本生成结果的

蕴含关系，作为衡量事实一致性的通用评估指标

文本流畅性指标：使用 BLEU-4，基于参考答案和预测结果，计算 n-gram 的匹配度。

2. 故事生成技术评测

任务简介：故事生成是指给机器一些故事相关的信息，让机器生成一个故事。故事生成是自然语言处理和人工智能领域的重要前沿课题，对于提升机器对语言的理解能力、生成能力等具有重要价值，近年来受到学术界和工业界的广泛关注。

然而一般的故事生成任务，给定的输入信息很少，导致输出可以十分多样，造成了自动评价的困难性。因此本任务给定了一个无序的情节序列作为输入，要求机器能够合理地利用这些情节，产生一个自然、流畅、有趣味性的长故事。

评测指标：本任务增大了输入的信息量，从而很好地缩减了输出的空间，不仅给自动评价提供了便利，而且也能更好地检验模型生成的可控性，同时还对模型进行情节规划的能力提出了更高的要求。数据来源于从网上爬取的中文故事，情节序列利用 RAKE 算法从故事中抽取得到，每个故事至多抽取 8 个情节，每个情节不超过 8 个词。

3. 图像描述生成评价方法评测

任务简介：本共享任务邀请参与者提交一个或多个面向图像描述生成任务的自动评测算法，算法的目标是使自动评测方法给出的分数与人工评测的分数尽可能一致。我们将为参与者提供研究所需数据集，并采用客观的评价指标结果作为提交算法的最终成绩。本评测要求选手提出面向图像描述生成任务的评测方法，利用该方法对自动生成的图像描述进行打分，并使自动评测结果尽量接近于人工评测结果。数据集（数据截取自公开数据集）：数据集包括以下三部分内容：（1）自动生成的图像描述文本集合；（2）人工撰写的图像描述文本集合；（3）文本集（1）的人工评分集。请参与者提交以下材料：（1）提出的自动评测算法的相关技术文档；（2）提出的自动评测算法在我们提供的测试集上的评分结果。

评价指标： 我们将通过衡量参与者提出的自动评测方法与人工评测方法之间的相关性来评价自动评测算法的好坏，评价指标将采用 Kendall 协调系数（Kendall Correlation）。

4. 中文句法错误检测技术评测

作文错误自动批改是对外汉语教学走向智能化过程中的重要环节。汉语中介语作文错误自动批改技术评测（Chinese grammatical error diagnosis，CGED）是目前对汉语作为第二语言自动批改领域持续时间最长的技术评测。

任务简介：我们模仿 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。

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。系统可以决定提交最多 3 个可能答案。

二、评测结果

第一届全国自然语言生成与智能写作技术评测的宣发和评测阶段开始于 2021 年 6 月，任务一、二、四陆续于 2022 年前结束竞赛任务，受疫情影响任务三于 2022 年 4 月结束竞赛。本届技术评测共获得海内外近七百支队伍的关注和参加。现就各任务评测结果进行简要介绍。

1. 文本生成一致性评测

文本生成一致性评测共吸引 577 名高校、企业参赛者，其中：57 支参赛队提交有效结果；30 支参赛队自动评测指标超过基线系统；排名 Top10 队伍中，收到 9 份参赛系统总结报告。参赛队伍来自腾讯、搜狗、小米 AI Lab、联想、OPPO、思必驰、QTrade 等知名企业，以及清华、中科院、哈工大、北交大、北理工、华南理工、湖南大学等顶尖高校和科研院所。任务要求各参赛团队提交详细技术报告，才算作有效成绩，因

此未提交报告的团队不予记录成绩。其中成绩有效的 TOP9 参赛团队信息如表 2 所示。来自腾讯的“对对队”、来自中科院自动化所和湖南大学的“后浪”联队和来自作业帮教育科技有限公司的 zipyourbelt 三支队伍分获冠亚季军。

表 2. 任务一前九名团队

排名	队伍	单位
1	对对队	腾讯-荔枝 LicheeGen
2	后浪	中科院自动化所&湖南大学
3	zipyourbelt	作业帮教育科技有限公司
4	策马奔腾	哈尔滨工业大学
5	IchbinDerek	QTrade
6	Anonymous 的团队	Sogou Inc
7	AIspeech	思必驰科技股份有限公司
8	破影 518	北京交通大学
9	LambdaX 的团队	小米 AI Lab

2. 基于大纲的故事生成技术评测

基于大纲的故事生成技术评测共分为主赛道和开放资源赛道两部分。其中开放资源赛道允许队伍使用包括悟道大规模预训练模型在内的其他资源。本任务共吸引 32 名高校、企业参赛者。参赛队伍来自腾讯等知名企业，以及中科院、哈工大、北交大、苏州大学、湖南大学等顶尖高校和科研院所。其中两个赛道的成绩优异队伍信息如表 3 和 4 所示。最终任务要求各参赛团队提交详细技术报告，才算作有效成绩，因此未提交报告的团队不予记录成绩。最终来自广西大学的“以上队伍成绩无效”、来自哈工大的“中国著名画家王羽熙”和来自南京大学的“不知道取啥名字”队分获主赛道冠亚季军。来自哈工大的“中国著名画家王羽熙”和来自中科院自动化所&湖南大学联队的“灵境”并列开放资源赛道冠军。来自南京大学的“不知道取啥名字”和来自广西大学“也不知道取啥名字”分获开放资源赛道亚军和季军。

表 3. 任务二主赛道前八名团队

排名	队名	参数规模（M）	bleu-1	bleu-2	distinct-3	distinct-4	cov	order	overall
1	以上队伍成绩无效	223	29.46	18.45	19.93	35.25	100	68.14	34.92

2	中国著名画家王羽熙	1000	28.76	17.48	20.62	36.41	95.83	67.54	34.14
3	不知道取啥名字	1000	27.87	17.77	19.68	34.25	100	63.59	33.67
4	IIE-NLP-Eyas	60	30.3	18.16	18.58	31.37	91.61	64.36	33.41
5	灵境	1000	27.24	15.79	20.06	35.77	89.54	65.67	32.29
6	你说的队	60	29.02	17.45	13.59	25.08	87.24	64.64	31.45
7	兔兔说的队	60	27.43	16.24	13.51	25.3	88.69	64.34	30.71
8	NLPer	60	8.11	3.7	12.3	23.96	70.22	54.97	18.51

表 4. 任务二开放资源赛道前五名团队

排名	队名	额外数据	额外模型	参数规模（M）	bleu-1	bleu-2	distinct-3	distinct-4	cov	order	overall
1	中国著名画家王羽熙	CMRC，LOT	CPT-large	450	73.19	68.41	24.24	42.35	98.96	88.48	67.75
2	灵境	悟道，LOT	CPT	400	71.77	67.64	27.36	46.71	98.37	87.79	67.69
3	不知道取啥名字	CRMC,wikidata		1000	35.11	22.69	21.98	38.03	100	63.95	37.89
4	也不知道取啥名字		CPM-1-Distill	109	21.27	12.73	33.08	52.3	99.86	63.47	33.06
5	国双科技	THUCTC,nlp_chinese_corpus	T5-pegasus	292.25	22.94	12.67	22.05	36.85	87.63	65.18	30.14

3. 图像描述生成评价方法评测

受到疫情影响，本评测赛程大幅度延期，原本征 18 家团队报名。直到 NLGIW2022 大会召开时，仅有一份评测结果返回，为来自华南理工大学的谢嘉嘉团队。该团队提交结果的 Kendall 协调系数（Kendall Correlation）为：训练集 0.2100，验证集 0.1931。

4. 中文句法错误检测技术评测

第七届中文句法错误检测技术评测吸引到 30 支队伍报名（企业包括华为、阿里、网易等，高校和院所包括南京大学、苏州大学、北京大学、新华社技术局等），18 支队伍提交了 45 个系统结果（本任务允许一支队伍提交至多三个系统）。竞赛分为五个赛道：假阳性（FPR）、错误包含（Detection）、错误类型（Identification）、错误定位（Position）和错误修正（Correction）。本次竞赛各赛道前三名和最优系统性能如表 5 所示。最终来自苏大&阿里联队（S&A）、网易有道（YYDS）、南京大学（NJU）、华为教育（corrector_dl）、新华社技术局（intospace）、北信科和鲁东大学（LDU）的团队在各赛道斩获佳绩。

总体而言本届评测各队伍在错误定位和错误修正两方面均取得了较大的提升。错误定位赛道首次突破 F1 0.5 大关，较之 CGED2020 有 10 个点的惊人提升。错误修正赛道也首次突破 F1 0.3。且这些突破出现在多支团队职工，这体现了全行业技术进步带来的普遍性能增长。近十年来的 CGED 技术评测见证了本任务系统从完全学术探索逐步走向实用雏型的历程。

赛道	第一名/F1	第二名/F1	第三名/F1
假阳性*	LDU**/0.0141	北信科/0.1224	intospace/0.135
错误包含	NJU/0.8778	LDU/0.8756	S&A/0.869
错误类型	S&A/0.7102	NJU/0.6907	YYDS/0.6854
错误定位	S&A/0.5157	YYDS/0.5061	NJU/0.4621
错误修正	S&A/0.3186	YYDS/0.3079	corrector_dl/0.2619

*假阳性率为假阳性数量/测试集样例数，其余赛道均为 F1 值

**每支队伍允许提交多个系统，此处列举的是 F1 值最优系统的成绩

三、评测论坛

本届评测于 2022 年 4 月 22-23 日首届自然语言生成与智能写作大会（NLGIW 2022）上举行技术评测论坛。论坛由共同评测主席饶高琦博士主持。本届评测论坛共发布了四项任务及其成绩。每项内容包括综述、成绩公布、优秀队伍进行技术分享等板块。

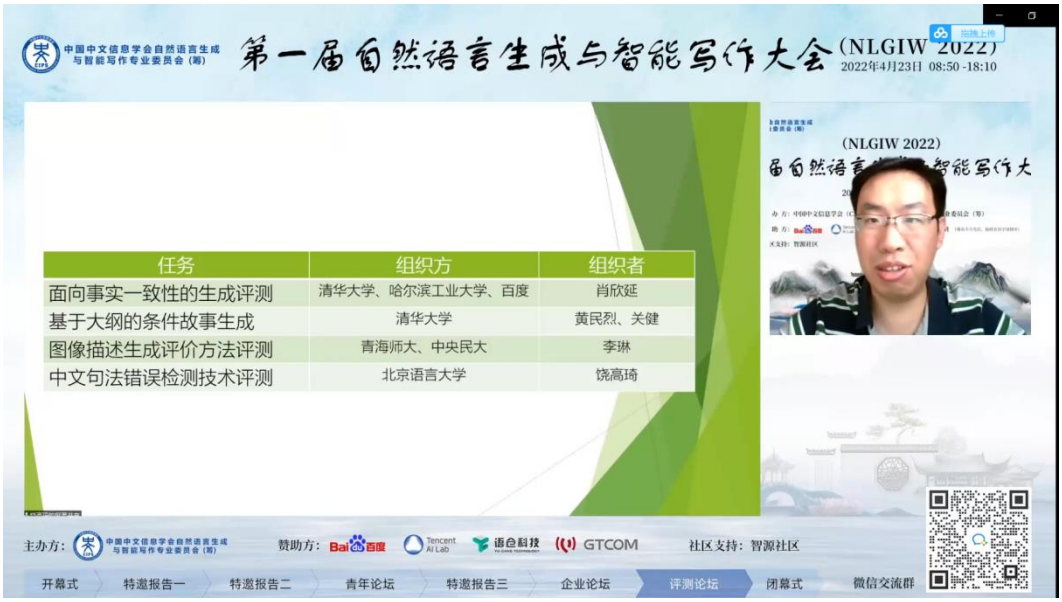

图 1. 共同评测主席饶高琦博士主持评测论坛

任务一文本生成一致性技术评测由清华大学、哈尔滨工业大学（深圳）和百度共

同组织，任务主席为百度主任架构师肖欣延。百度刘家辰对任务一进行了综述报告。

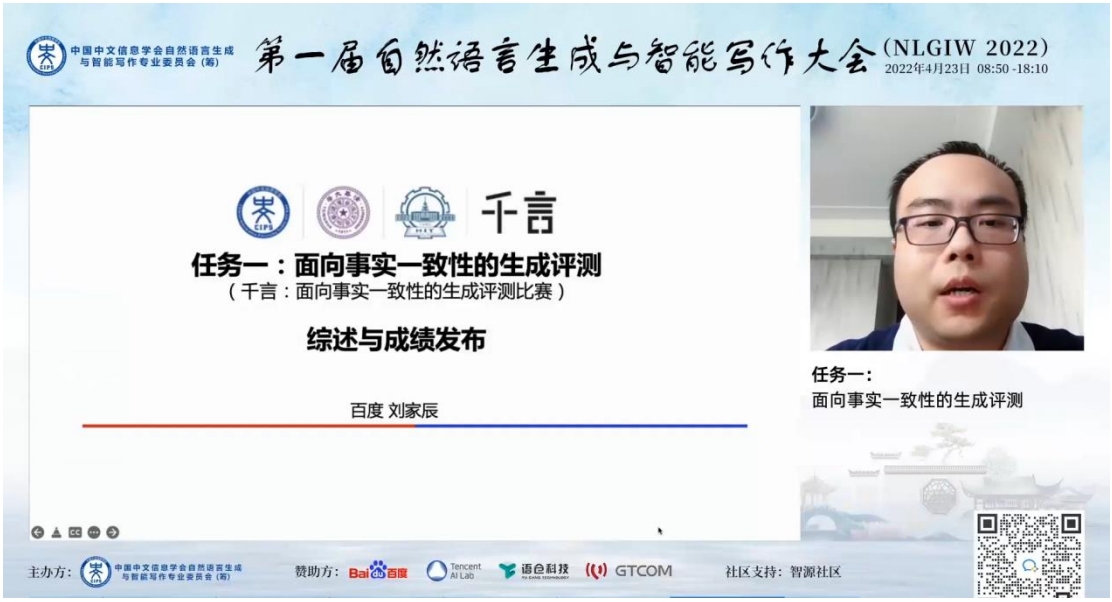

图 2. 百度刘家辰老师综述任务一内容并发布成绩

评测任务冠军为来自腾讯的“对对对”团队。刘辉代表团队介绍了基于领域预训练的事实一致性生成解决方案。该团队采用预训练+领域预训练+微调的三阶段训练模式大幅度提高了文本生成的效果并且极大限度保持生成文本的事实一致性。同时，也使用对抗训练、模型融合等方法进一步提高模型效果，最终取到了该评测的第一名。

来自哈尔滨工业大学的“策马奔腾”队获得评测第四名，龚恒博士针对团队解决方案作报告《基于预训练语言模型和数据增强的文本生成系统》。该团队选用同时兼顾了自然语言理解和自然语言生成两类任务的预训练语言模型进行微调，在大幅提升生成文本流畅性的同时保持较好的事实一致性。另外，团队也通过数据增强的方法，进一步提升文本生成的总体效果。

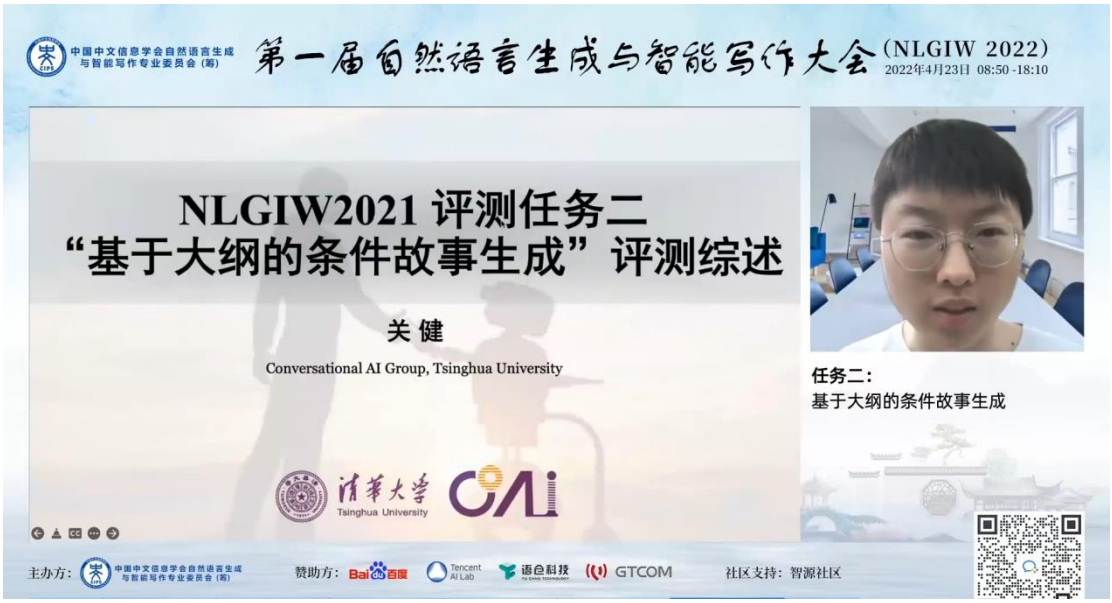

图 3. 清华大学关健博士做任务二综述并发布成绩

本届评测任务二为基于大纲的条件故事生成。该任务由清华大学组织，任务主席为黄民烈副教授。关健博士发布了评测成绩并做综述。任务冠军团队来自广西大学。廖泽明博士作报告《基于模型融合和数据增强的条件大纲故事生成》，介绍了其团队在主赛道上的冠军解决方案。在具体的实践中，团队结合了数据增强、数据预处理和模型融合的方法，在现有模型基础上取得了一定的提升效果。

哈尔滨工业大学的钟蔚弘博士作报告《基于数据增强和任务解构的条件故事生成》介绍了开放资源赛道冠军解决方案和主赛道亚军技术方案。该团队通过数据增强和任务解构的方法，促进了模型对于任务形式的学习理解，同时将一步生成转化为多步生成，进一步降低了生成难度，提高了生成过程的可控性。

来自湖南大学的李宾博士作报告《Transfer and Denoising Learning For Story Generation》，介绍了其在开放资源赛道并列冠军技术方案。该团队设计了一种 rake 训练方法，以弥补训练前任务和下游任务之间的差距。团队引入子调整方法来学习与任务相关的知识，并使用去噪学习来获得一个完整且逻辑流畅的故事。

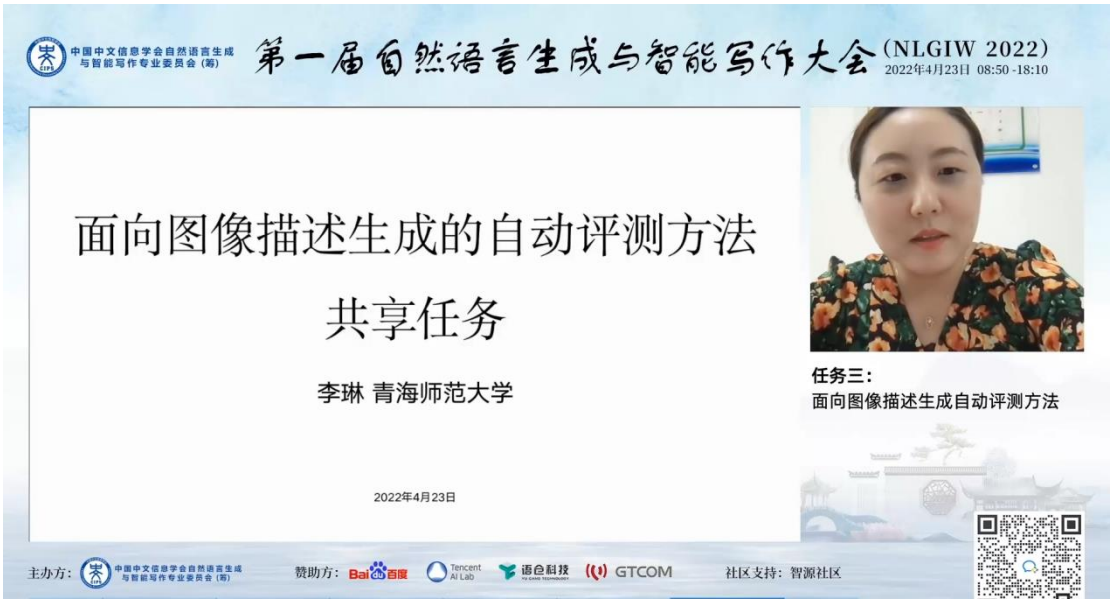

图 4. 青海师范大学李琳副教授综述评测任务三

青海师范大学和中央民族大学联合组织了本届大会的任务三图像描述生成评价方法。青海师范大学李琳副教授综述了图像描述自动生成技术、数据与测试。该评测要求团队提出面向图像描述生成任务的评测方法，利用该方法对自动生成的图像描述进行打分，并使自动评测结果尽量接近于人工评测结果。

图 5. 北京语言大学饶高琦博士综述评测任务四并发布成绩

本届大会任务四为第七届中文句法错误检测（Chinese grammatical error diagnosis，CGED-7）。该评测由北京语言大学饶高琦博士组织。该评测是目前对汉语作为第二语

言自动批改领域持续时间最长的技术评测。相较以往，今年的评测成绩有了较大提高。在错误定位和错误修正两个较受关注的赛道上，S&A 和 YYDS 两支队伍均获得冠军和亚军。

苏州大学李嘉诚博士代表团队作报告 Suda&Alibaba at CGED-7: Ensembles of Error Detection and Correction Models for Chinese Grammatical Error Diagnosis(E2DC)，介绍了苏州大学和阿里巴巴联队在综合评价指标上刷新纪录的技术方案。该团队采用了两类模型，即基于序列标注的语法检错模型和基于序列到编辑的语法纠错模型。与此同时他们也尝试了不同的数据增强策略来缓解训练语料不足的问题。对于检错子任务，该团队利用 ERRANT 将额外的语法纠错数据（错误-正确平行句对）转换为语法检错模型的训练数据；对于纠错子任务，团队采用基于规则和基于反向翻译的数据增强策略生成大量的伪训练数据。

来自网易有道的方美媛博士则以《高精确率导向的中文句法错误诊断系统》为题介绍了亚军团队的独特工作，深入讲解了深耕精确率并创造新高的方法。

以上部分技术报告和评测综述详见自然语言生成与智能写作专委会网站。

四、第二届评测活动筹备

目前第二届全国自然语言生成与智能写作技术评测已启动筹备工作。专委会主席赵铁军教授在 NLGIW 大会上发布了 2023 年度评测计划，欢迎广大同行积极参与新一轮评测活动。第二届评测任务拟于 2022 年 6 月开始宣发，持续到 2022 年底。专委会诚挚地向相关领域的学者、研究机构及企业征集评测任务方案。评测任务包括而不限于以下主题：自然语言生成基础任务（人机对话、自动问答、自动文摘、图片/视频描述）、智能写作相关任务（文案生成、新闻写作、自动作文、作文批改、作文评分、文本校对）。方案中应详细描述任务内容、评价标准、评测数据的准备情况及大致赛程，请有意设置任务的主体通过邮件发送评测方案至评测联系人饶高琦博士（raogaoqi At blcu.edu.cn）并抄送专委会（nlgiw2021 At 163.com）。