第二十一届全国机器翻译与多语言信息处理大会
机器翻译评测大纲

发布时间：2025.04.08

（CCMT 2025 MT Evaluation: http://mteval.cipsc.org.cn:81/CCMT2025/index.html)

1. 引言

第二十一届全国机器翻译与多语言信息处理大会（CCMT 2025）将于 2025 年 9 月在甘肃兰州举行。根据惯例，本次会议将继续组织统一的机器翻译评测。

CCMT 2025 继续组织机器翻译评测，其主要内容如下：

- 延续经典翻译评测任务，包括由 CCMT 与 WMT 合作组织的中英、英中新闻领域的翻译评测；维汉、蒙汉、藏汉的双语翻译评测；自动译后编辑评测。
- 延续多领域机器翻译任务。

与往届评测相同，本次评测不设置统一发放训练数据的时间，各参评单位报名之后即可获取数据并进行系统训练，测试数据将在统一时间发放。本次评测的评测论文（含系统评测报告）将与 CCMT2025 学术论文一同接受匿名审稿，并择优推荐与 CCMT2025 学术论文共同发表。

希望本次评测能够促进国内外科研单位、产业界相关单位之间的学术交流和联系，共同推动机器翻译研究和技术的发展。

本次评测的组织信息如下：

评测主办机构：
中国中文信息学会

评测组织单位：
中国中文信息学会机器翻译与多语言信息处理专业委员会

评测资源提供单位（按单位名称拼音首字母顺序）：
阿里巴巴达摩院
北京语智云帆科技有限公司
点通数据有限公司
东北大学
华为技术有限公司
南京大学
内蒙古大学

青海师范大学
网智天元科技集团股份有限公司
西北民族大学
西藏大学
厦门大学
中国科学院计算技术研究所
中国科学院新疆理化技术研究所
中国科学院自动化研究所
中央民族大学

评测委员会联合主席：
毛存礼（昆明理工大学）
王　瑞（上海交通大学）

评测委员会委员（按姓氏拼音首字母排序）：
陈科海（哈尔滨工业大学（深圳））
陈毅东（厦门大学）
飞　龙（内蒙古大学计算机学院）
冯　洋（中国科学院计算技术研究所）
冯　冲（北京理工大学）
何中军（百度公司）
侯宏旭（内蒙古大学）
黄书剑（南京大学）
毛存礼（昆明理工大学）
那顺乌日图（内蒙古大学）
仁青东主 (西藏大学)
施艳蕊（中译语通科技股份有限公司）
王龙跃（腾讯 AI Lab）
王　瑞（上海交通大学）
魏勇鹏（北京语智云帆科技有限公司）
肖　桐（东北大学）
杨雅婷（中国科学院新疆理化技术研究所）
杨沐昀（哈尔滨工业大学）
杨　浩（华为技术有限公司 2012 实验室）
杨宝嵩（阿里巴巴达摩院）
尹　曦（鹏城实验室）
于满泉（网智天元科技集团股份有限公司）
张家俊（中国科学院自动化研究所）
张　珮（阿里巴巴达摩院）

有关评测最新信息参见：http://mteval.cipsc.org.cn:81/CCMT2025/index.html

2. CCMT2025 评测任务一览

本次机器翻译技术评测由双语翻译、自动译后编辑、多领域翻译任务组成，我们将为每项评测项目的参评单位提供相应的训练语料、开发语料，测试语料。今年的评测不再设置在线评测环节，具体的评测方法详见大纲后续部分的说明。

注：本届 CCMT 将对评测打分工具和评测指标进行更新，具体评测指标以后期发布为准。

本次翻译技术评测任务的具体评测项目如下所示：

表 1 CCMT 2025 任务 1：双语翻译评测项目表

序号	项目代号	评测项目名称	翻译方向	领域
1	CE	中英新闻领域机器翻译	汉语->英语	新闻领域
2	EC	英中新闻领域机器翻译	英语->汉语	新闻领域
3	MC	蒙汉综合领域机器翻译	蒙语->汉语	综合领域
4	TC	藏汉综合领域机器翻译	藏语->汉语	综合领域
5	UC	维汉新闻领域机器翻译	维语->汉语	新闻领域

表 2 CCMT 2025 任务 2：自动译后编辑评测项目表

序号	项目代号	评测项目名称	翻译方向	领域
1	CE-PE	中英机器翻译自动译后编辑	汉语->英语	ICT
2	EC-PE	英中机器翻译自动译后编辑	英语->汉语	ICT

表 3 CCMT 2025 任务 3：多领域机器翻译评测项目表

序号	项目代号	评测项目名称	翻译方向	领域
1	CE-CA	多领域机器翻译	汉语->英语	汽车领域
2	CE-EL	多领域机器翻译	汉语->英语	电子领域
3	CE-EM	多领域机器翻译	汉语->英语	能源领域
4	CE-FI	多领域机器翻译	汉语->英语	经济领域
5	CE-IT	多领域机器翻译	汉语->英语	IT 领域
6	CE-LI	多领域机器翻译	汉语->英语	文学领域
7	CE-MA	多领域机器翻译	汉语->英语	机械领域
8	CE-ME	多领域机器翻译	汉语->英语	医药领域

3. 双语翻译评测项目概况

3. 1 任务介绍

与往年一致，翻译任务主要评测参评单位在双语翻译任务上的性能。评测翻译语言对包括英中、中英、蒙汉、藏汉、维汉等。

3.2 评测指标

本次评测中继续沿用自动评测方式，即利用自动评价工具对参评单位提交的最终翻译结果文件进行评价。依据惯例，计划采用多种自动评价标准。**本届CCMT 计划采用 COMET 和 BLEU 两种评价指标**，最终具体评测指标以后期发布为准。评测组织方进行自动评价时将采用如下设置：

- 所有自动评测将采用大小写敏感（case-sensitive）的方式，评测结果中也包含部分大小写不敏感的评价作为参考；
- 英中、藏汉、维汉和蒙汉四个方向将采用基于字符（character-based）的评价方式；
- 英中、藏汉、维汉和蒙汉四个方向中，评测组织方将对 GB2312 编码的 A3 区字符进行全角到半角的转换；
- 中英项目则采用基于词（word-based）的评价方式。

3.3 评测数据

本次评测由主办方提供全部训练、开发、测试集数据，感谢提供并授权使用数据的各个研发单位。本次评测的数据设置方式基本和 CCMT2024 保持一致，将使用 CCMT2025 开发的测试集。

(1)训练数据

本次评测训练数据的情况请见下表（提供单位排名不分先后，以汉语拼音为序）。

表 4 CCMT 2025 双语翻译任务训练数据情况

评测项目名称	训练规模（句）	提供单位	说明
中英-英中新闻领域机器翻译	9023475	点通数据有限公司、东北大学、中国科学院计算技术研究所、中国科学院自动化研究所	平行语料
	5281	中国科学院计算技术研究所	汉语和四个英语参考译文
	1000	中国科学院计算技术研究所	英语和四个汉语参考译文
	1000	南京大学	CWMT2017

			开发集、测试集
	约 1100 万词	厦门大学	汉语单语语料
蒙汉综合领域机器翻译	1259777	中央民族大学、中国科学院自动化研究所、内蒙古大学、中国科学院计算技术研究所	平行语料
	1000	内蒙古大学	CCMT2020 开发集
藏汉综合领域机器翻译	1657580	北京网智天元科技股份有限公司西北民族大学、中国科学院自动化研究所、青海师范大学、西藏大学、厦门大学、中国科学院计算技术研究所	平行语料
	1000	青海师范大学	CCMT2020 开发集
维汉新闻领域机器翻译	171061	中国科学院计算技术研究所、中国科学院新疆理化技术研究所	平行语料
	1000	中国科学院新疆理化技术研究所	CCMT2020 开发集

其中，中英与英中评测项目的训练数据继续与 WMT2025 共享，WMT2025 在 News 任务下提供的中英双语训练数据（含单语数据）也可以在本评测对应的中英和英中项目使用，见 http://www.statmt.org/wmt25/translation-task.html。

(2)开发数据

原则上，本次评测开发数据使用的是 CCMT2023 线下评测的测试集，即 CCMT2024 在线评测的测试集。下表简要给出各项目本次测试数据的具体规模，并向各数据提供单位表示感谢。

备注：在本评测中，WMT 2025 News 任务指定的开发数据可以作为训练集使用(http://www.statmt.org/wmt25/translation-task.html)，而不应作为开发集使用。

表 5 CCMT 2025 双语翻译任务开发数据

评测项目名称	**规模（单位：句）**	**提供单位**	**说明**
中英-英中新闻领域机器翻译	500-中英 500-英中	东北大学	CCMT2023 线下测试集 即 CCMT2024 在线测试集
蒙汉日常用语	1000	内蒙古大学	CCMT2023 线下

评测项目名称	规模 （单位：句）	提供单位	说明
机器翻译			测试集 即 CCMT2024 在线测试集
藏汉政府文献 机器翻译	1000	青海师范大学	CCMT2023 线下 测试集 即 CCMT2024 在线测试集
维汉新闻领域 机器翻译	1000	中国科学院新疆理 化技术研究所	CCMT2023 线下 测试集 即 CCMT2024 在线测试集

(3)测试数据

本次评测测试数据的设计见下表，具体以测试数据发放时的情况为准。

表 6 CCMT 2025 双语翻译任务测试数据情况

评测项目名称	规模 （单位：句）	提供方	说明
中英新闻领域 机器翻译	待定	东北大学	单参考译文
英中新闻领域 机器翻译	待定	东北大学	单参考译文
蒙汉综合领域 机器翻译	待定	内蒙古大学	单参考译文
藏汉综合领域 机器翻译	待定	西藏大学	单参考译文
维汉新闻领域 机器翻译	待定	中国科学院新疆理 化技术研究所	单参考译文

3.4 系统训练要求

对于每个评测项目，参评单位可以自由选择所采用的机器翻译技术（如：基于规则的机器翻译技术、基于实例的机器翻译技术、统计机器翻译技术、神经网络机器翻译技术及**基于大模型的机器翻译技术**等）。参评单位也可以使用系统融合技术，但在系统描述中必须做出明确说明，并在技术报告中说明进行系统融合的各个单系统的性能。此处，系统融合技术指使用两个及两个以上单系统的翻译结果进行字、词、短语、句子级别的重构或选择的技术。没有明确产生两个或两个以上单系统翻译结果的技术，如统计机器翻译中的协同解码、

神经网络机器翻译的输出层 ensemble、单个系统结果的重排序等，本次评测不认定为系统融合技术。评测组织方在发布评测结果时，将对使用系统融合技术的系统进行标注说明。

对于使用以上机器翻译技术的参评系统，可以以“受限”和“非受限”两种方式参与评测。下面对两种方式进行详细说明：

- **受限训练：**受限训练是只使用评测组织方指定范围的数据进行训练。具体说明如下：
  - 参评单位提交的“主系统”必须采用受限训练。
  - 对于以基于规则的机器翻译技术为主的参评系统，允许采用通过人工方式构造的翻译知识（如规则、模板、词典等），但要在系统描述和技术报告中对所使用的翻译知识的规模、构造和使用方式等给出清晰的说明。
  - 单语分析工具可以使用外部数据，如词法分析、句法分析及命名实体识别工具等可以使用外部数据，在评测报告中应对此进行积极性说明；涉及双语翻译的工具不能使用外部数据，包括命名实体翻译、音字转换工具等（数词和时间词翻译不受此约束）。
  - **对于基于大模型的机器翻译技术为主的参评系统**，允许使用模型总参数数量小于 20B 的开源模型框架。这些模型需遵循允许非商业用途的无限制使用的开源许可协议，例如 Apache、MIT 等。以下是符合条件的模型列表（包括但不限于）：Llama-7B、Llama-13B、Qwen2.5-7B、Ministral-8B、Mistral-7B、Aya-Expanse-8B、Aya-101 -13B、NLLB，**所有模型需满足参数规模限制与开源协议双重规范。**
  - 对于综合使用多种不同机器翻译技术或进行系统融合的参评系统，**其所有单系统均须满足受限训练和受限系统要求**。
  - 每个评测项目只允许使用评测组织方发布的该项目相关的训练数据，不可以使用其他评测项目的数据。即对于参加多个评测项目的单位，不同项目提供的数据不可以联合使用。
  - 与 WMT 联合组织的中英、英中领域评测项目的受限训练语料包括 CCMT 方提供的数据；也包括由 WMT 组织提供的数据。为便于比较，请参评单位提交中英、英中领域系统的评测报告时说明是使用 CCMT 数据还是 WMT 数据还是两者皆有，评测组织方将在发布评测报告时对相应的系统结果予以标识。
- **非受限训练：**非受限训练是指可以使用评测组织方指定范围数据之外的数据进行训练。具体说明如下：
  - 参评单位提交的 “对比系统” 可以采用非受限训练。
  - **参评单位可以使用预训练语言模型及闭源大模型（如 GPT-4、PaLM、DeepSeek 满血版等，参数量不受限制）构建翻译模型。**

- 上述系统将被认定为非受限系统。请在系统描述和技术报告中对所使用的预训练语言模型和大模型情况进行详细说明。评测组织方将在发布评测报告时对相应的系统结果予以标识。
- 采用非受限训练方式的系统，需要在系统描述和技术报告中对使用的数据进行说明（如数据规模和领域类型、是否为可公开获取的数据等。若为可公开获取的数据，则需说明数据出处）。

3.5 结果提交

参评单位收到测试数据后，应在规定时间内提交最终翻译结果文件。对于每个评测子项，参评单位应提交 1 个主系统翻译结果（Primary Result）、以及最多 2 个对比系统的翻译结果（Contrast Result）。提交的每个结果文件都应包含详细的系统描述。

4．自动译后编辑任务评测方法

4.1 任务介绍

自动译后编辑（Automatic Post-Editing, APE）任务通过学习人工对机器翻译译文的修改，从而实现自动更正机器翻译译文、提升翻译质量。APE 的应用场景包括不限于：（1）解决一些因编解码器黑盒特性或者训练开销太大而导致的难以干预或优化的翻译错误；（2）将面向通用领域输出的翻译结果改写成更适合特定垂域的翻译结果；（3）降低人工编辑的代价，提升翻译效率。

本任务包括一个语向：中文到英语，领域为 ICT。

本任务由华为公司联合设立。

4.2 评测指标

本任务采用自动评价指标 HTER 和 BLEU。最终系统排名基于 HTER，BLEU 仅用作相关性参考。

HTER 采用 TERCOM 工具（http://www.cs.umd.edu/~snover/tercom/），设置为：分词（tokenized），大小写敏感（case sensitive）。

4.3 评测数据

该任务提供了一个包含 5000 句训练集、1000 句开发集、1000 句测试集的训练数据。具体说明如下：

- 数据来源：华为消费者官网（https://consumer.huawei.com/）可公开获取的手册类内容，由 CCMT 进行数据挖掘、过滤，调用任务中提供的基线机器翻译接口生成机器翻译译文，并由华为翻译中心专业译员生成人工译后编辑译文。

- 数据组成：所有集合中，均为华为终端的手册类内容。
- 数据格式：训练集和开发集每条数据包括一个完整的三元组：中文原文（source）、机器翻译译文（target）、人工译后编辑译文（post-edit）。测试集每条数据包括：中文原文（source），机器翻译译文（target）。
- 数据授权：由 CCMT 统一进行授权。

4.4 系统训练要求

参评单位可选择基于评测组织方提供的训练集，采用**传统 APE 方法或基于大模型的 APE 方法进行模型训练**，也可使用其他可公开获取的数据集。所有数据集均需在系统说明和评测报告中予以说明。参评单位可以以“受限”和“非受限”两种方式参与评测。下面对两种方式进行详细说明：

- **受限训练**：受限训练是只使用评测组织方指定范围的数据进行训练。具体说明如下：
  - 参评单位提交的“主系统”必须采用受限训练。
  - **采用基于大模型的 APE 方法的参评系统，**允许使用模型总参数数量小于 20B 的开源模型框架。这些模型需遵循允许非商业用途的无限制使用的开源许可协议，例如 Apache、MIT 等。以下是符合条件的大模型列表（包括但不限于）：Llama-7B、Llama-13B、Qwen-2.5 -7B、Ministral-8B、Mistral-7B、Aya-Expanse-8B、Aya-101-13B、NLLB，**所有模型需满足参数规模限制与开源协议双重规范。**
- **非受限训练**：非受限训练是指可以使用评测组织方指定范围数据之外的数据进行训练。具体说明如下：
  - 参评单位提交的 “对比系统” 可以采用非受限训练。
  - **参评单位可以使用预训练语言模型及闭源大模型（如 GPT-4、PaLM、DeepSeek 满血版等，参数量不受限制）构建 APE 模型。**
  - 上述系统将被认定为非受限系统。请在系统描述和技术报告中对所使用的预训练语言模型和大模型情况进行详细说明。评测组织方将在发布评测报告时对相应的系统结果予以标识。
  - 采用非受限训练方式的系统，需要在系统描述和技术报告中对使用的数据进行说明（如数据规模和领域类型、是否为可公开获取的数据等。若为可公开获取的数据，则需说明数据出处）。

4.5 提交结果

每个参评单位允许提交两个不同的系统结果，分别为受限训练系统和非受限训练系统，需在结果文件名中明确标识系统类型为受限还是非受限。结果文件中，APE 译文的格式如下：

[测试集行号]tab[APE 译文]

5. 多领域机器翻译任务评测方法

5.1 任务介绍

多领域机器翻译任务旨在评估和提升机器翻译系统在不同专业领域下的翻译能力，目标是在多样化领域中都能实现高质量的翻译效果。现实世界的文本来源于多种多样的领域，如法律、医疗、科技、金融、新闻等。每个领域都有其独特的术语、语法结构和表达方式，而机器翻译系统通常倾向于在大规模的新闻等通用性语料上进行训练，这可能会导致在特定领域上翻译品质不佳。

本任务的翻译语向为中文到英文，主要领域涉及 IT、金融、医学、能源矿物、工业汽车、机器设备、电子技术、科技文献。

5.2 评测指标

本任务评测采用自动评测方式，即用自动评价工具对参评单位提交的最终翻译结果文件进行评价，计划采用 COMET 和 BLEU 两种评价指标，最终具体评测指标以后期发布为准。自动评价指标采用的设置为：分词（tokenized），大小写敏感（case sensitive）。

5.3 评测数据

(1)训练数据

本任务不提供训练数据集，参评单位可根据开发集自行收集和构建需要的平行语料用以训练翻译系统和增强翻译系统的领域适应能力。

(2)开发数据

本任务提供 8 个中英领域的开发集，包括 CCMT2024 离线评测的开发集及测试集，领域为 IT、金融、医学、能源矿物、工业汽车、机器设备、电子技术、科技文献。

(3)测试数据

本任务提供 8 个中英领域的测试集，由阿里巴巴集团提供，具体以测试数据发放时的情况为准。

表 7 CCMT 2025 多领域翻译任务测试数据情况

评测领域	规模 （单位：句）	说明
IT	待定	单参考译文
金融	待定	单参考译文
医学	待定	单参考译文
能源矿物	待定	单参考译文
工业汽车	待定	单参考译文
机器设备	待定	单参考译文
电子技术	待定	单参考译文
科技文献	待定	单参考译文

5.4 系统要求

参评单位可以自由选择所采用的机器翻译技术，如：基于规则的机器翻译技术、基于实例的机器翻译技术、统计机器翻译技术、神经网络机器翻译技术以及**基于大模型的机器翻译技术**等。

参评单位可使用各种预训练语言模型和大模型构建翻译模型，包括但不限于以下模型：mBART、BERT、RoBERTa、XLM-RoBERTa、sBERT、LaBSE、Llama-7B、Llama-13B、Qwen-2.5-7B、Ministral-8B、Mistral-7B、Aya-Expanse-8B、Aya-101-13B、NLLB、GPT-4、PaLM、DeepSeek 满血版等。

参评单位需要在系统描述和技术报告中对使用的模型和训练数据进行说明（如训练数据来源、训练数据规模和领域类型、是否为可公开获取的数据等。若为可公开获取的数据，则需说明数据出处）。

5.5 提交结果

参评单位收到测试数据后，应在规定时间内提交所有领域的最终翻译结果文件。结果文件中，翻译译文的格式如下：

[测试集行号]tab[翻译译文]

6. 提交技术报告及参加评测会议

评测结束后，每个评测项目的参评单位应向 CCMT 2025 会议提交一份详细的技术报告，说明系统的架构、原理、使用的主要技术以及数据使用的情况。评测报告将与 CCMT2025 投稿论文采用相同的匿名同行评审机制，并择优推荐与 CCMT2025 论文共同发表。

评测报告发表前，应针对学术论文投稿要求进行修改。欢迎将评测中的技术成果提炼形成学术论文形式，直接提交 CCMT 投稿论文。

参加评测单位应派至少一人参加 CCMT 2025 会议进行相应技术交流。

10. 评测日程

日期	评 测 环 节
2025.04.08	发布评测大纲，评测报名开始。评测组织方向报名单位提供训练集、开发集数据（通过邮件发送下载方式）
2025.05.20	报名截止
2025.06.03-06.09	发布测试集、提交参测结果。
2025.06.09-06.27	参评单位提交评测技术报告（通过会议投稿系统）
2025.07.02-07.07	评测组织方向参评单位通知初步评测结果
2025.07.17	评测报告评审结果通知
2025.08.02	评测技术报告终稿提交
2025.09	CCMT2025 会议，发布评测总结