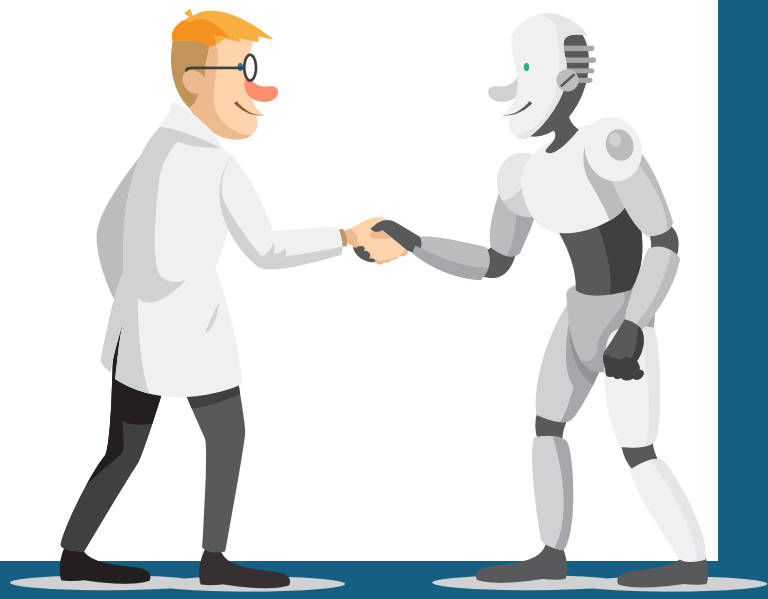
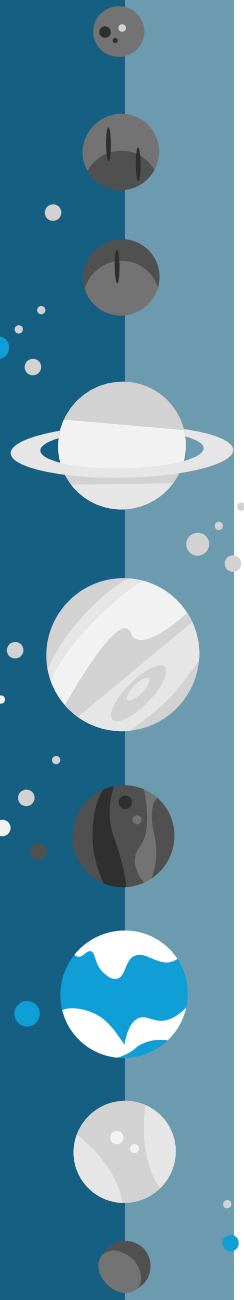
第10课:机器翻译技术及应用概论

黄婕

2024年11月20日/11月26日



本节内容：



- 1. 初识机器翻译
- 2. 机器翻译的瓶颈
- 3. 机器翻译的发展历史
- 4. 挑战与前景
- 5. 布置小组作业+汇报

1. 初识机器翻译

什么是机器翻译?

源语言
Source

机器翻译
MT

目标语言
Target

基于实例的翻译方法?

基于规则的方法?

基于神经网络的翻译方法?

基于中间语言的翻译方法?

必应

基于统计的方法?

Google

Baidu 百度

国内有哪些机器翻译团队?

5

常见机器翻译产品有哪些？

在语言服务行业的应用如何?

翻译及语言服务提供方与需求方均表示看好机器翻译前景。

被调查的企业52.9%的翻译项目使用了机器翻译。

中国翻译协会，2024中国翻译行业发展报告

应用领域广泛

人际沟通	信息分析	跨文化传播
• 个人出国翻译	• 网页翻译	• 专利摘要翻译
• 智能问答翻译	• 跨境电商翻译	• 影视字幕翻译
• 电话口译	• 跨语言舆情分析	• 图书出版翻译
• 国际会议同传	• 信息安全翻译	• 新闻传媒翻译
• 社区口译	• 行业新技术追踪	• CAT调用机器翻译
• 电子邮件翻译	• 专利信息检索	• 交互式机器翻译
• 即时通信翻译	• 自动摘要	• Office翻译插件
• …	• …	• …

2. 机器翻译的瓶颈

机器翻译的困难

自然语言中普遍存在的歧义和未知现象

- 白天鹅飞了/南京市长江大桥/她背着生病的丈夫，给毕节地区的希望小学捐款/休假式治疗/维修性拆除/节操碎了一地

机器翻译不仅仅是字符串的转换

- 青梅竹马/高山流水/江湖/印堂发黑/欲练神功，必先自宫/一饮一啄饱蘸苦辣酸甜/面子/阳春白雪/下里巴人/你妈叫你回家吃饭了

机器翻译的解不唯一，而且始终存在人为的标准

- 这也是翻译专业的同学需要面对的问题

各类修辞的文学句段

- 最是那一低头的温柔，像一朵水莲花不胜凉风的娇羞

机器翻译的困难—名词术语

夫妻肺片	
Couple’s lung slices	Sliced beef and ox tongue in chili sauce
水煮鱼	
Boiled fish	Boiled fish in chili oil
馒头	
Steamed bread	Steamed bun

机器翻译的困难—古诗词

松下问童子，言师采药去。
只在此山中，云深不知处。

谷歌翻译：
Panasonic asked the boy, herbalist herbs to go.
Only in this mountain, cloud depths do not know.

人工翻译（林语堂）：
I asked the boy beneath the pines.
He said, “The master’s gone alone,
Herb-picking somewhere on the mount.
Cloud-hidden, whereabouts unknown.”

机器翻译的困难—歧义

We do chicken right.

Google: 我们做的对。

百度: 我们是烹鸡专家。

腾讯： 我们做鸡是对的。

有道: 我们做鸡是对的。

DeepL:我们的鸡肉做得很好。

机器翻译的困难—歧义

词义消歧

原文：老子是春秋时期的教育家。
译文：Lao Zi is an educator during Spring and Autumn period.

原文：老子砍死你！
译文：I cut you dow!

原文：老子出生在鲁国。
译文：Lao Tze was born in Lu.

原文：老子出生在山东！
译文：I was born in Shandong!

机器翻译译文质量评测案例

	金融	IT	法律
百度	24.64	27.30	28.59
DeepL	29.12	26.84	26.90
GPT 3.5 turbo	25.24	26.93	24.47
GPT 4 turbo	29.88	25.69	28.89
Gemini 1.0 Pro	24.09	24.20	25.85

张井，2024中国翻译协会年会新时代翻译技术创新发展论坛，2024,3,31

机器翻译译文质量的自动评估

- **MT自动评估计算工具**
  - 输入两份译文，一键计算评测指标得分
  - 适用于MTPE教学、译文评测等场景下，计算机器译文/学生译文与参考译文的相似差异程度
- 工具链接

  https://www.shiyibao.com/tools/MTPEtest
- 工具演示

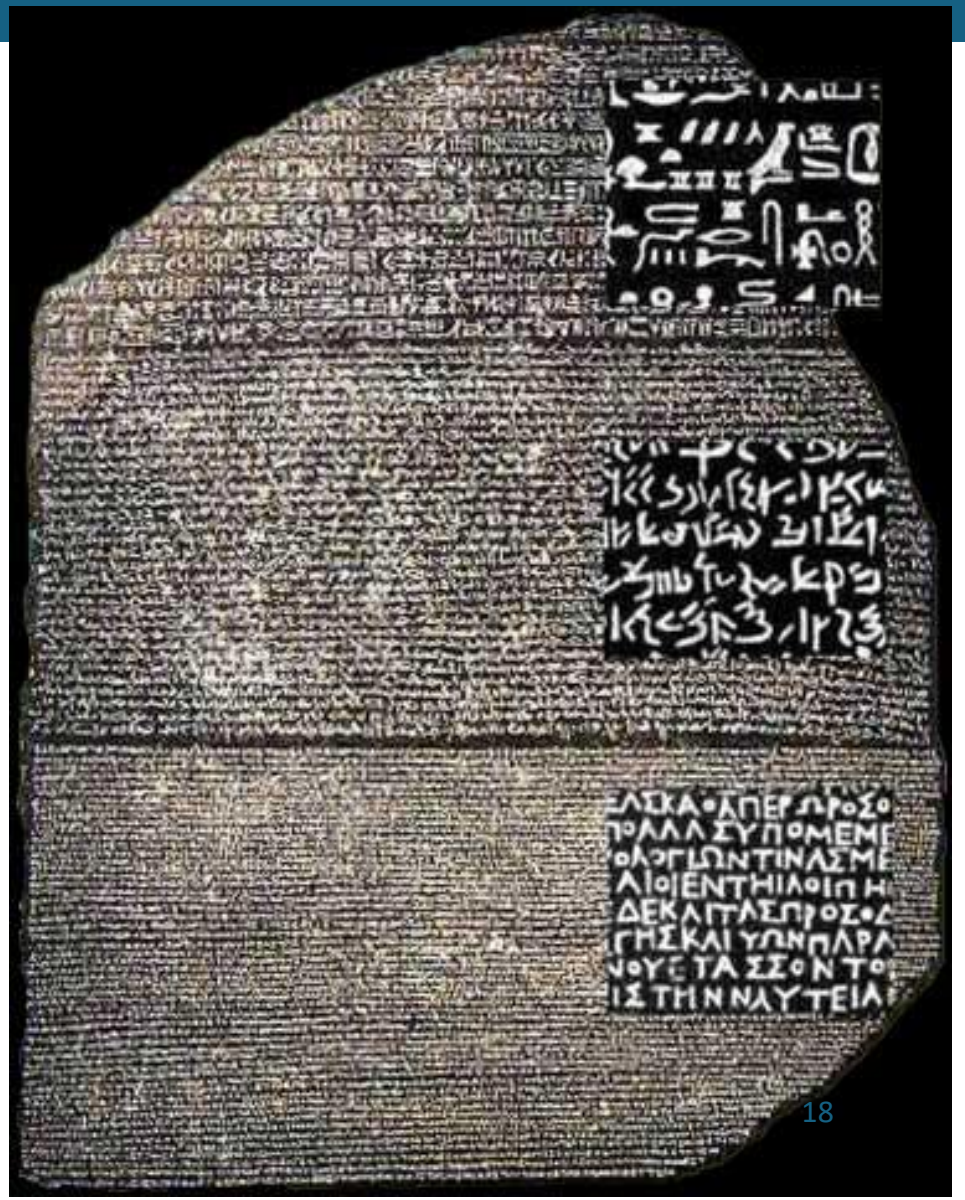
3. 机器翻译的发展历史

罗塞塔石碑？

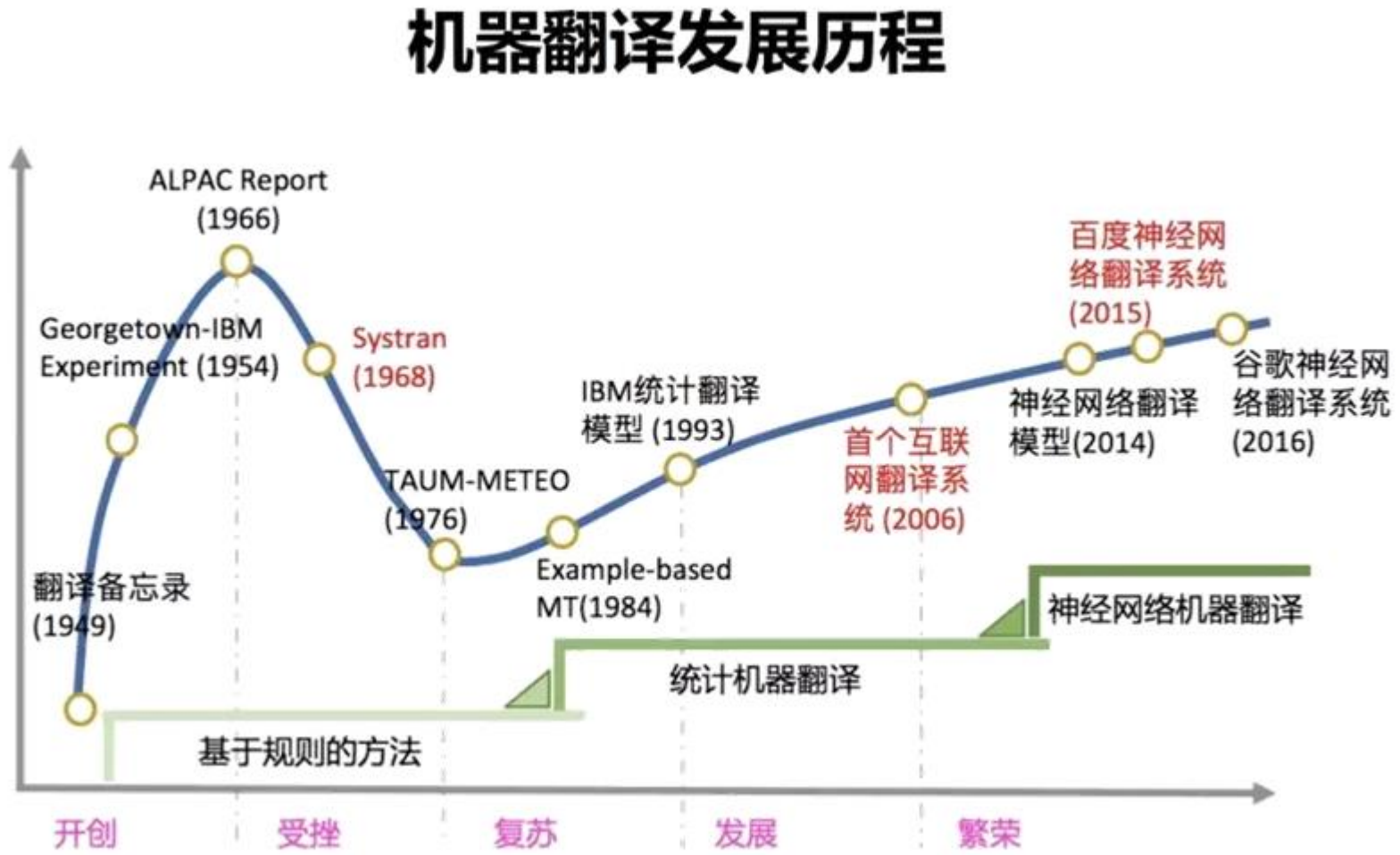
罗塞塔石碑是一块非常著名的古代文物，发现于1799年的埃及。

石碑上刻有同一段文字，分别用三种语言书写：**古埃及象形文字**、**古埃及草书**和**希腊文**。由于当时希腊文已经被解读，所以石碑提供了破译古埃及文字的线索。

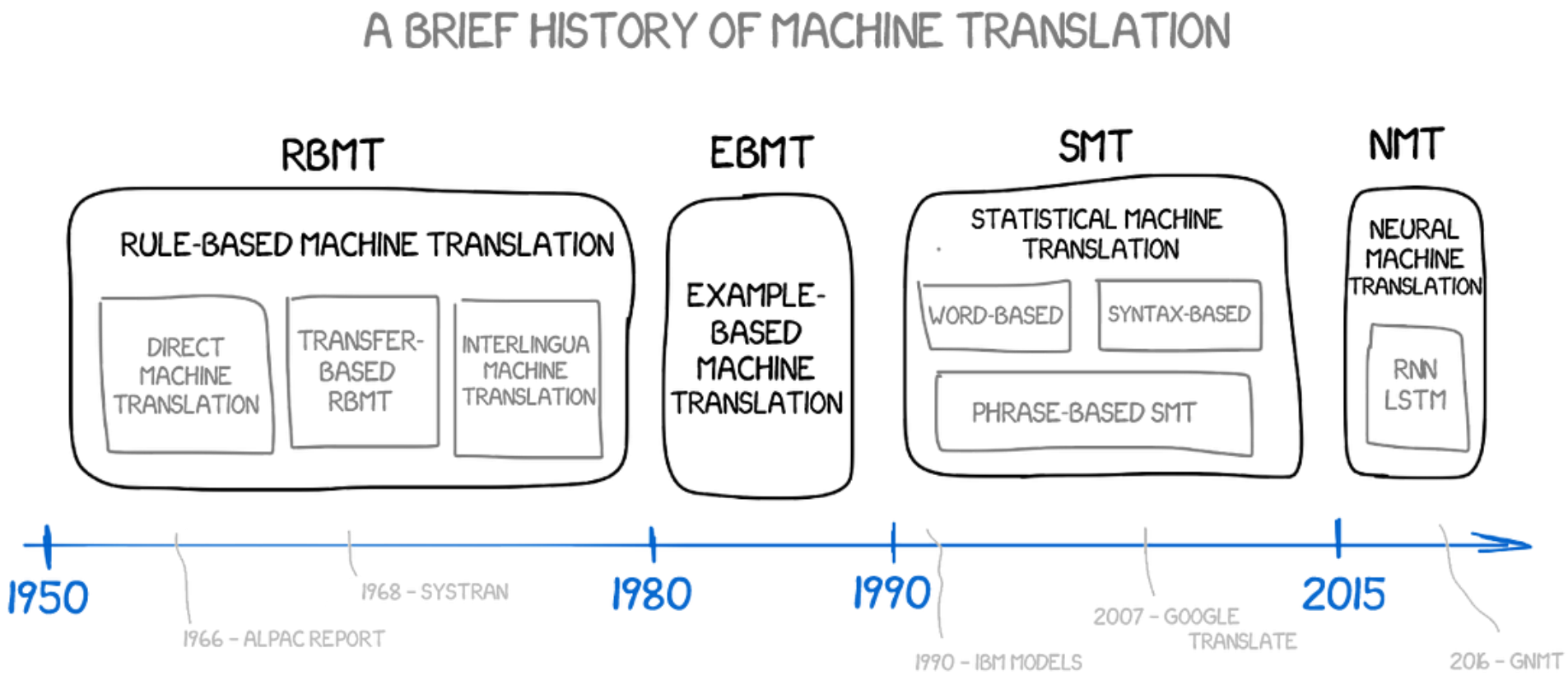
罗塞塔石碑上刻有同一段文字的三种语言版本，这启发了人们对于使用**多语言对照**的方法来解决翻译难题的思路。类似地，机器翻译也借鉴了这种思想：对比源语言和目标语言之间的多种对照版本。



机器翻译发展历程



机器翻译：几个关键阶段


A BRIEF HISTORY OF MACHINE TRANSLATION
RBMT
RULE-BASED MACHINE TRANSLATION
DIRECT MACHINE TRANSLATION
TRANSFER-BASED RBMT
INTERLINGUA MACHINE TRANSLATION
EBMT
EXAMPLE-BASED MACHINE TRANSLATION
SMT
STATISTICAL MACHINE TRANSLATION
WORD-BASED
SYNTAX-BASED
PHRASE-BASED SMT
NMT
NEURAL MACHINE TRANSLATION
RNN LSTM
1950
1966 – ALPAC REPORT
1968 – SYSTRAN
1980
1990
1990 – IBM MODELS
2007 – GOOGLE TRANSLATE
2015
2016 – GNMT

机器翻译的技术发展阶段

- 基于规则的机器翻译 (**Rule-based** machine translation)
- 基于实例的机器翻译 (**Example-based** machine translation)
- 统计机器翻译 (**Statistical** machine translation)
- 神经网络机器翻译 (**Neural** machine translation)
- 大语言模型的机器翻译 (Machine translation using **large language models**)

四种机器翻译技术的工作原理是什么？

阅读文献：第1章-机器翻译简介-肖桐-朱靖波.pdf

I. 基于规则的机器翻译 (RBMT)

资源：规则库

1: If 源 =“我”, **then** 译 =“I”

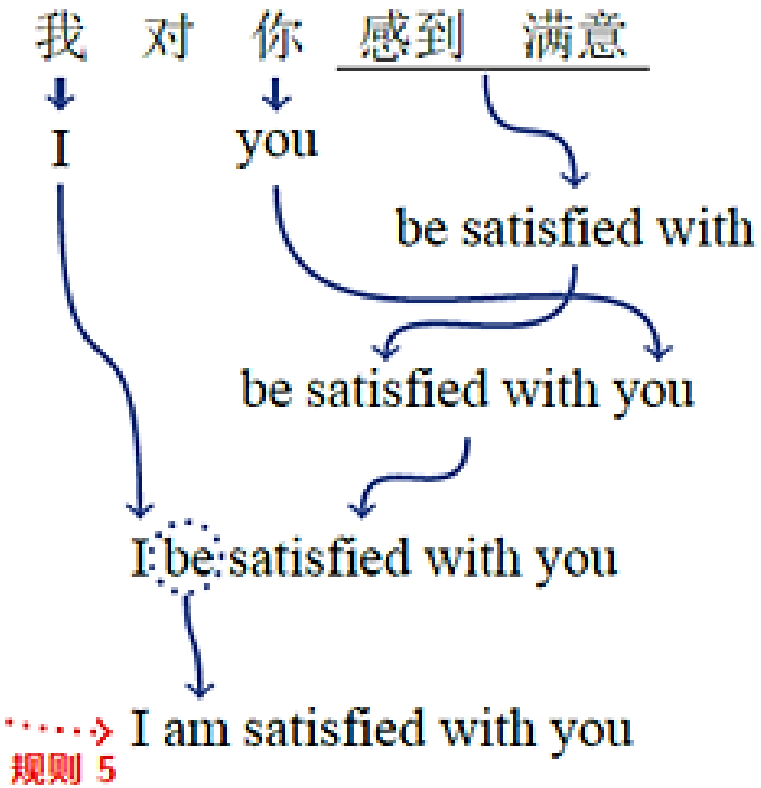
2: If 源 =“你”, **then** 译 =“you”

3: If 源 =“感到满意”,
then 译 =“be satisfied with”

4: If 源 =“对... 动词 [表态度]”
then 调序 [动词 + 对象]

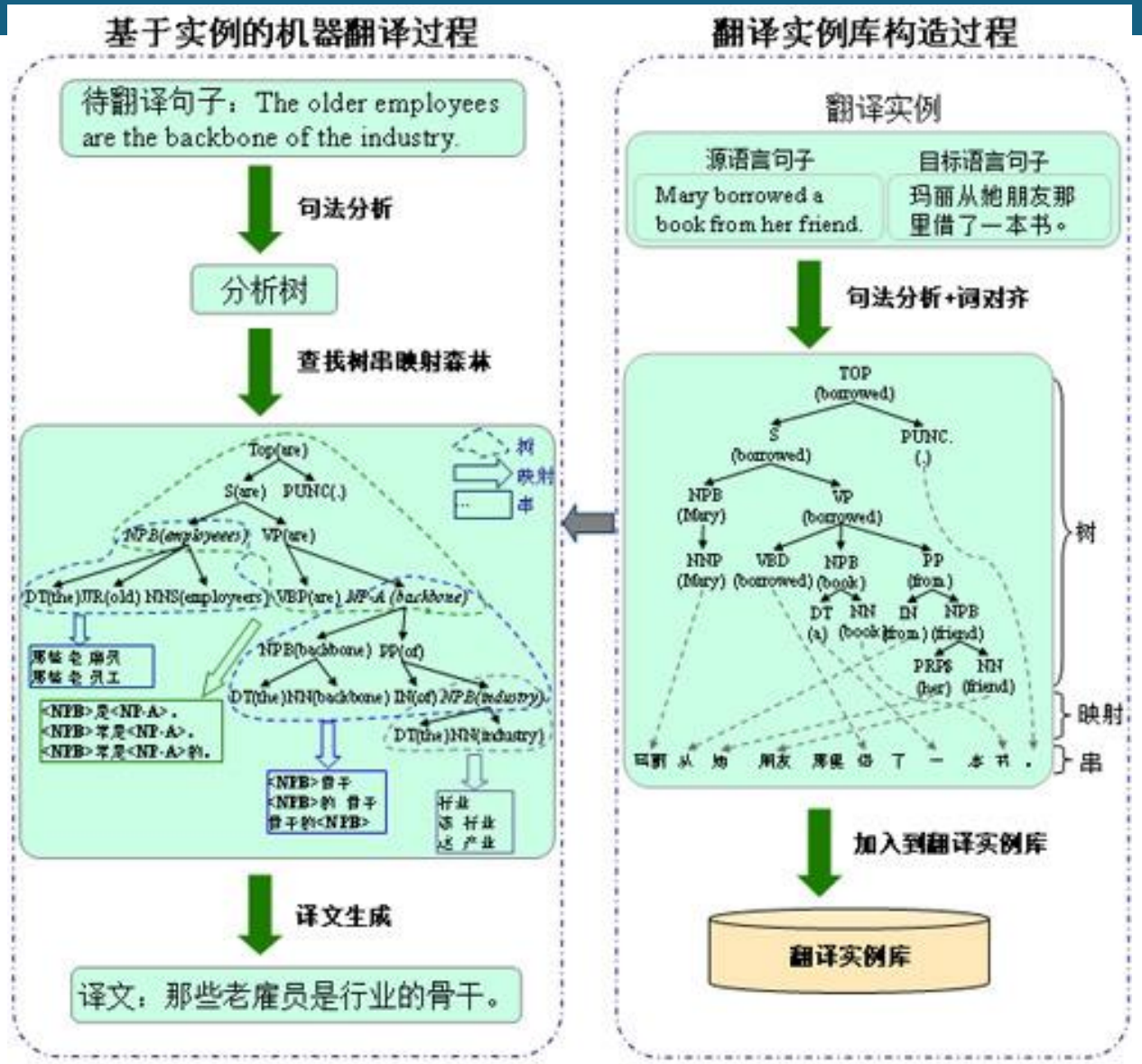
5: If 译文主语是 “I”
then be 动词为 “am/was”

6: If 源语是主谓结构
then 译文为主谓结构

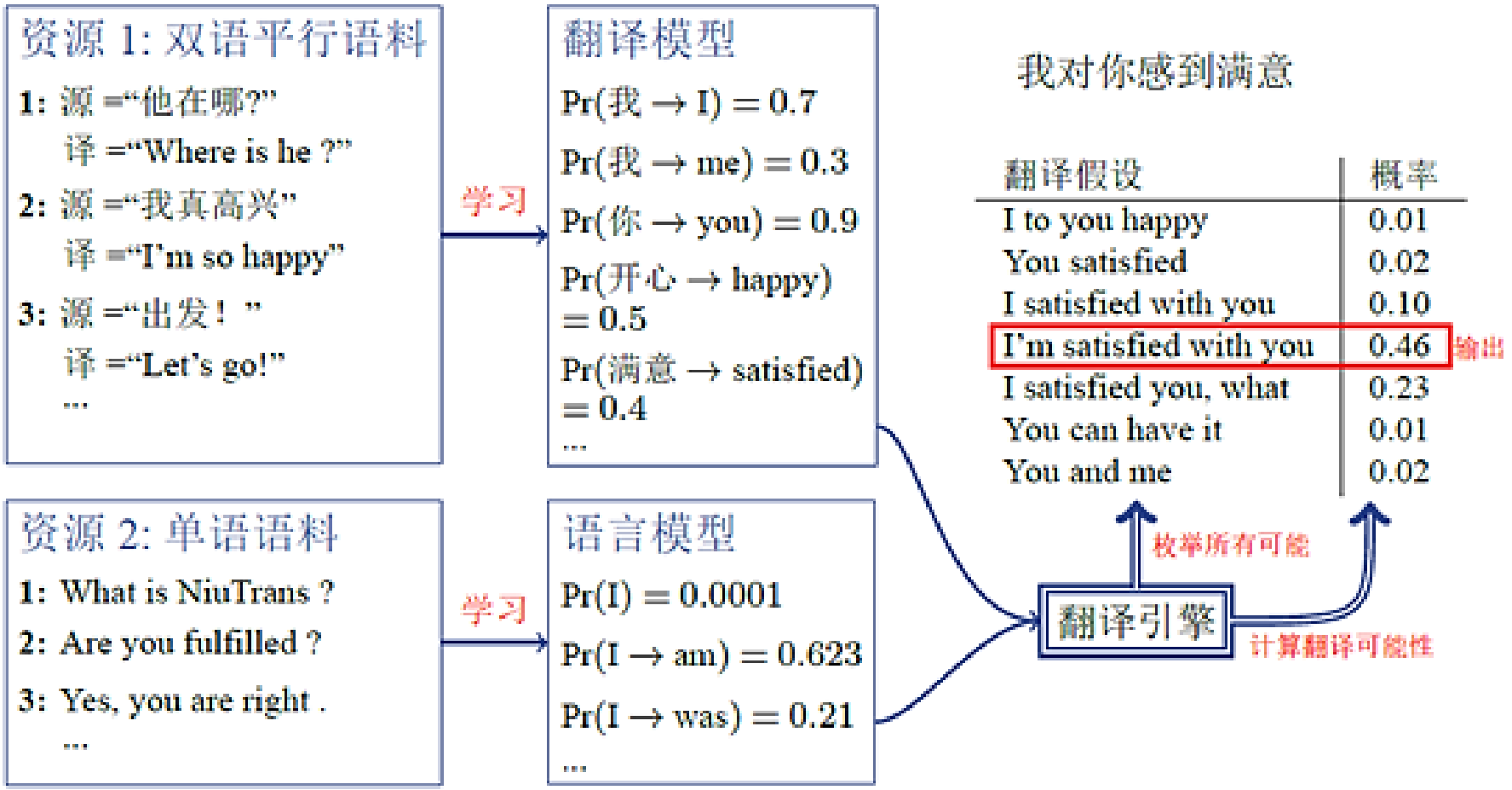


贺承浩,王泽辉,滕俊哲,等.机器翻译综述[J].电脑知识与技术,2023,19(21):31-34.

2. 基于实例的机器翻译 (EBMT)



3. 基于统计的机器翻译技术



贺承浩,王泽辉,滕俊哲,等.机器翻译综述[J].电脑知识与技术,2023,19(21):31-34.

统计机器翻译技术示例

The weather is very changeable ……
年年在这个时候天气都变化无常。

He's a very nice man.
他是个很好的人。

You're very chatty today, Alice.
艾丽斯，你今天很健谈。

今天天气晴朗。
今天天气很暖和。
起初天气很好。
天气很好,旅游是再好不过了。
旅客: 好的,今天天气很好

翻译模型　　语言模型　　调序模型……

解码器

$$\hat{T} = \arg\max_{T} p(T) \times p(S|T)$$

$$\log p(t|s) = \sum_{i=1}^{n} \lambda_i h_i(t,s)$$

对数线性模型

The weather is / nice / today / .

今天 / 天气 / 很好 / 。

4. 神经网络机器翻译

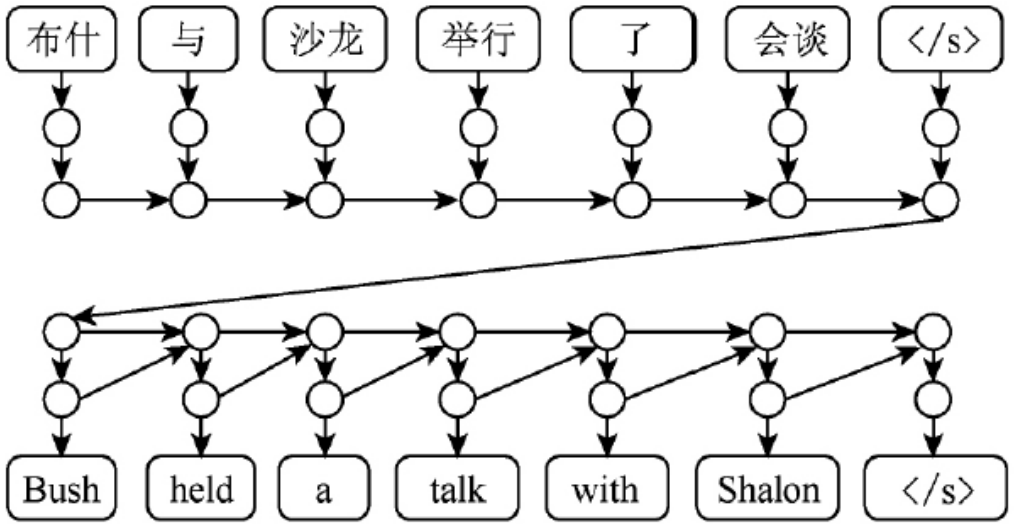


Fig. 2 The encoder-decoder framework

图 2 编码器-解码器框架

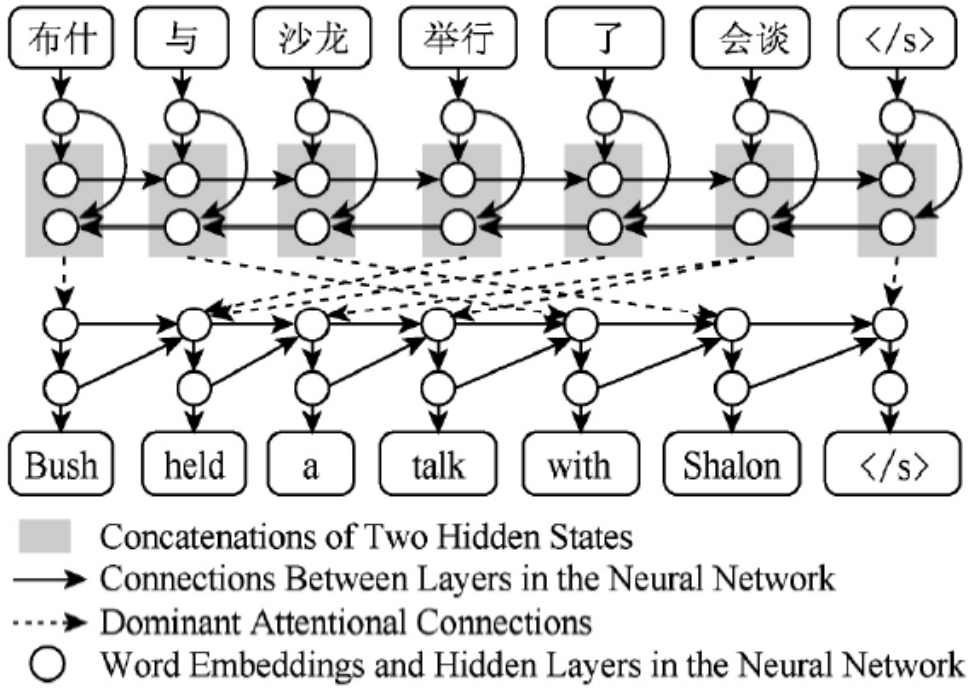


Fig. 3 Attention-based neural machine translation

图 3 基于注意力机制的神经机器翻译

刘洋.神经机器翻译前沿进展[J].计算机研究与发展,2017,54(06):1144-1149.

5. 大语言模型的机器翻译

新特性：大模型+大数据

编码器-解码器架构	循环神经网络（RNNs）或者注意力机制	词嵌入（Word Embeddings）	端到端学习	大规模数据集的训练
这是机器翻译模型的基本框架。编码器负责将输入文本编码成一个语义向量，解码器则利用这个语义向量生成目标语言的文本。	在编码器和解码器中通常会使用循环神经网络或者注意力机制来捕捉输入文本和输出文本之间的长期依赖关系和语义信息。	将单词表示为连续的向量空间中的点，这样可以更好地表示单词之间的语义关系。	大型语言模型通常采用端到端的学习方法，即直接从原始输入到目标输出进行学习，而不需要手工设计特征或者规则。	这些模型通常需要大量的双语数据来进行训练，以便学习到有效的翻译规则和语言模式。

案例：文件机器翻译

- **任务：**
  - 你是电力研究院的一名信息情报员，每天需要阅读50-100篇电力行业国外最新英文材料（Word格式），然后编译出5篇中文材料。
- **挑战：**
  - 如何快速筛选可以编译的英文材料？
  - 把Word文字复制到机器翻译原文框，获得译文后粘贴到Word中吗？

案例：文件机器翻译 – 思路1

- **思路1：**
  - 使用基于文件的机器翻译。
- **实现：**
  - 小牛快译Word和WPS插件
  - 百度翻译
  - 有道翻译
  - 云译翻译
  - 文档翻译（新译科技）
  - 文档快翻qtrans

案例：文件机器翻译 – 思路2

- **思路2：**
  - 在CAT工具中，调用机器翻译插件，进行批处理-预翻译。
- **实现：**
  - Trados Studio + DeepL API

翻译结果 – DeepL Translator provider using DeepL Trans … 片段匹配 – DeepL Translator provider using DeepL Trans … 相关搜索 备注 TQA (0) 术语识别 术语库搜索

EN_Facts-1.doc.sdlxliff [翻译]

	EN_Facts-1.doc		EN_Facts-1.doc	
1	***Sharing our passion for learning and science to create opportunities for youth***	NMT	***分享我们对学习和科学的热情，为青年创造机会***	P
2	**The Golden Sun** is a global non-profit education program that serves students aged 10-18.	NMT	**金太阳** ″ 是一个全球性的非营利教育项目，服务对象是 10 至 18 岁的学生。	P
3	THE GOLDEN SUN has grown out of the spirit of goodwill and close ties between Company people and the communities where they live and work.	NMT	金太阳公司的发展源于公司员工与其生活和工作所在社区之间的友好精神和紧密联系。	
4	THE GOLDEN SUN began in 1998 as a way for Company employees, spouses and retirees to share their time, experience and passion for learning and science through a variety of volunteer activities with younger generations of learners.	NMT	金太阳 ″始于 1998 年，旨在让公司员工、配偶和退休人员通过各种志愿活动，与年轻一代的学习者分享他们的时间、经验以及对学习和科学的热情。	P
5	THE GOLDEN SUN provides access to technological and knowledge resources for underserved students and teachers in communities where Company people live and work.	NMT	金太阳 ″计划为 ″公司人 ″生活和工作的社区中得不到充分服务的学生和教师提供获取技术和知识资源的机会。	P
6	These include a range of project-based activities provided through an extensive multilingual website, hands-on science education workshops, and collaborative international projects.	NMT	这些活动包括通过广泛的多语种网站提供的一系列以项目为基础的活动、实践科学教育讲习班和国际合作项目。	
7	In these ways, THE GOLDEN SUN is building a learning community that creates connections among youth around the world and expands their understanding of science.	NMT	通过这些方式，″金太阳 ″正在建立一个学习社区，在世界各地的青少年之间建立联系，并扩大他们对科学的了解。	
8	In addition, the THE GOLDEN SUN Action Fund provides financing to young people for local initiatives addressing sustainability issues in their communities, for example in relation to water and energy.	NMT	此外，金太阳行动基金（THE GOLDEN SUN Action Fund）还为年轻人提供资助，帮助他们在当地开展活动，解决其所在社区的可持续发展问题，例如与水和能源有关的问题。	
9	THE GOLDEN SUN plans and carries out activities through three integrated programs:	NMT	金太阳 ″通过三个综合计划来规划和开展活动：	P
10	**The Golden Sun Program** invites qualified underserved schools to apply for funding that supports three goals:	NMT	**金太阳计划** 邀请符合条件的服务欠缺学校申请资助，以实现三个目标：	P
11	*Infrastructure*:	NMT	*基础设施*	P
12	Providing technical and financial support to connect underserved educational organizations to the Internet;	NMT	提供技术和资金支持，将服务不足的教育组织接入互联网；	
13	*Collaboration*:	NMT	*合作*：	P
14	Facilitating opportunities to participate in projects, events and partnerships with other educational organizations;	NMT	促进参与项目、活动和与其他教育组织合作的机会；	

案例：文件机器翻译 – 思路3

- **思路2：**
  - 使用能够分析文件的大模型工具进行翻译+分析。
- **实现：**
  - Kimi.ai
  - Sider
  - 更多?

4. 挑战和前景

机器翻译存在的问题：2024

Pang, J., Ye, F., Wang, L., Yu, D., Wong, D. F., Shi, S., & Tu, Z. (2024). Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models. arXiv preprint arXiv:2401.08350.

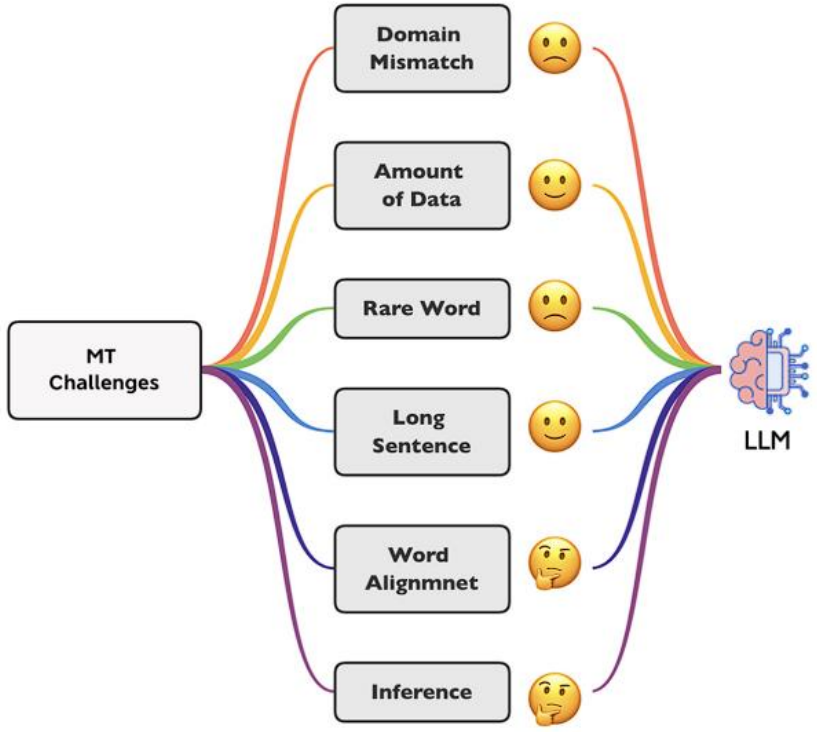


Figure 1: The six MT challenges are revisited in the context of LLM. Emojis symbolize the key findings: a 'smiley face' for issues largely addressed, a 'thinking face' indicating partial alleviation with ongoing concerns, and a 'sad face' for persisting unresolved challenges. This can be treated as the architecture of our paper.

机器翻译存在的问题：2013

张家俊&宗成庆. 2013.
机器翻译研究进展与趋势

长句和复杂句式的处理问题

弱规范、非规范化文本的翻译问题

双语资源缺乏问题

缺乏基于理解的翻译模型

篇章级翻译问题

增长式学习问题

反馈学习问题

机器翻译评测指标问题

应用创新问题

资源共享问题

机器翻译的技术应用前景

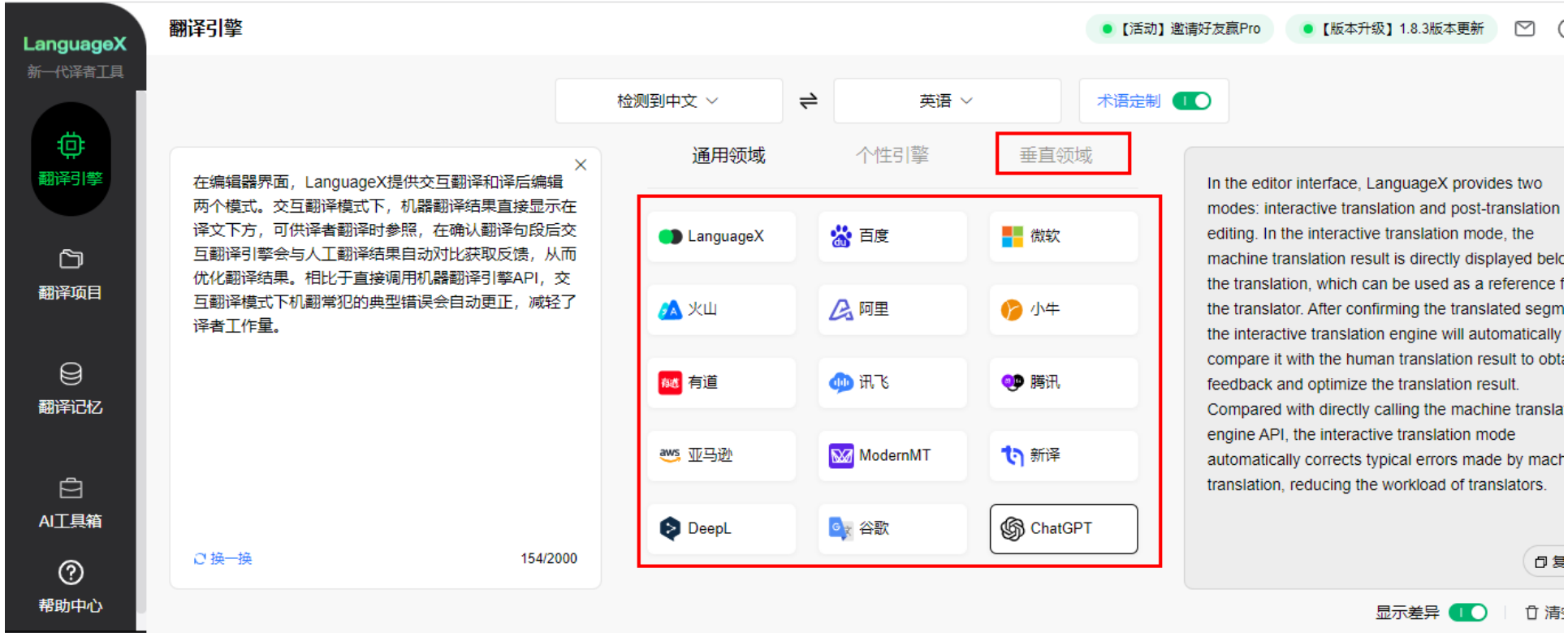
前景1:垂直行业的机器翻译

- 使用垂直行业的语料训练机器翻译，提高机器翻译译文质量，在垂直行业翻译中比通用翻译质量更好。

前景2: 交互机器翻译

Language的网址：https://languagex.com/

个性化翻译引擎：官方多款个性化机翻引擎可选，一键训练自己的个性化引擎；
通用翻译引擎：全球顶尖十余款机器翻译引擎，多个专业领域，并可定制术语
AI辅助翻译：基于术语、记忆库和译者输入，实时推荐译文的交互式翻译，而非被动译后编辑



前景2: 交互机器翻译

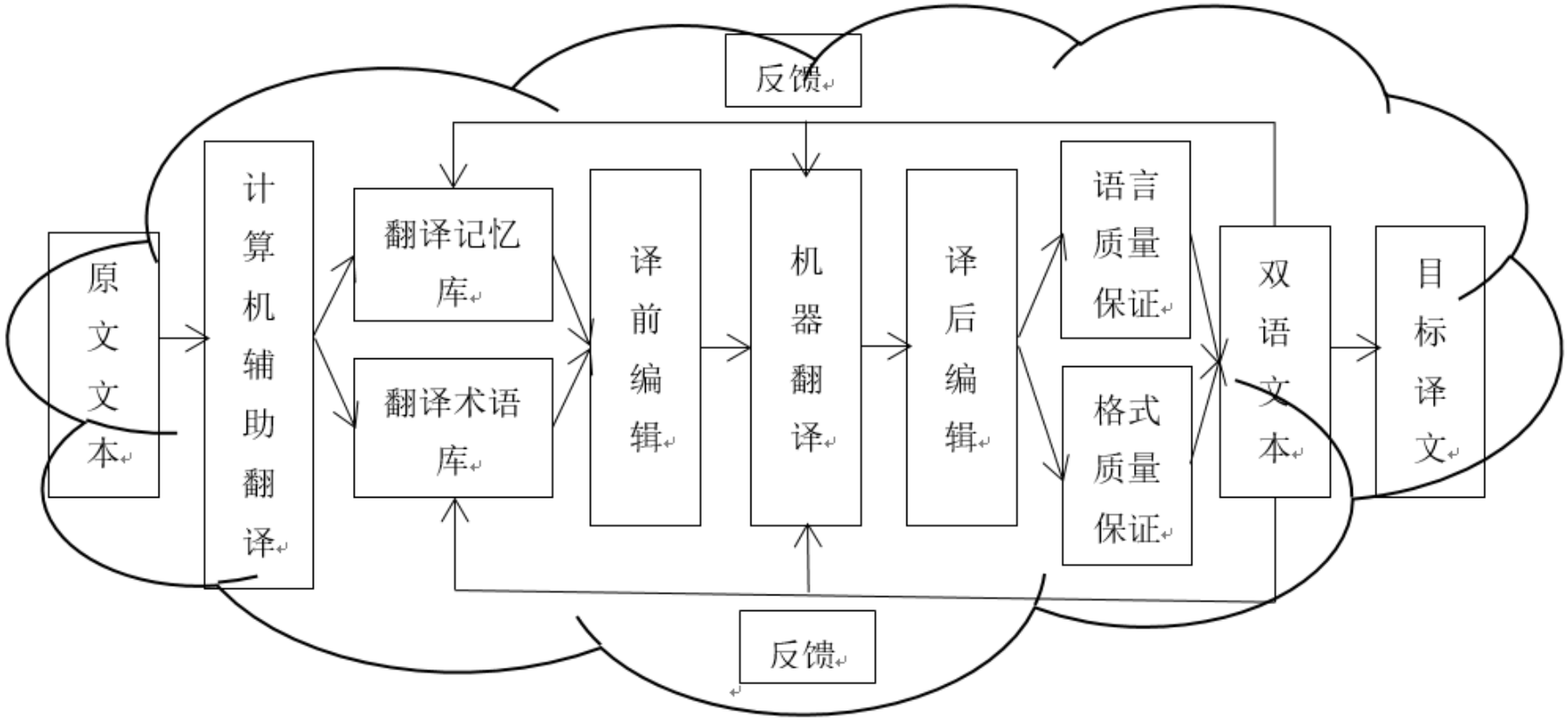
腾讯辅助翻译Transmart
https://transmart.qq.com/index

前景3: 大语言模型LLM机器翻译： ChatGPT

Microsoft Edge浏览器的“扩展”插件

OpenAI

前景4: CAT+MTPE融合应用



王萌，崔启亮，基于人机结合翻译模型的团队翻译技术策略研究，2020

趋势与展望

更深层次的语言理解：

- 语法、语义、逻辑和语境的理解：更复杂的模型架构和更高级的自然语言处理技术。

跨语言学习和迁移学习：

- 利用不同语言之间的相似性和共同特征来提高翻译效果。

多模态翻译：

- 同时处理文本、图像、视频等多种类型的数据，并有机结合。

增强式学习和交互式学习：

- 从与用户的交互中不断改进和优化翻译效果。

个性化翻译：

- 根据用户的偏好和习惯生成个性化的翻译结果。

机器翻译对翻译专业的影响与启示

积极跟踪、学习和应用机器翻译技术与工具

理解机器翻译的技术原理

创建和对齐语料库，为机器翻译输送高质量数据

对机器翻译的译文质量进行专业评估（人工评估和及自动评估）

根据译文评估结果，选择适当的机器翻译工具（通用机器翻译、垂直机器翻译）

应用基于文档的机器翻译工具（Office的MT插件，支持文档翻译的MT）

应用CAT+MT+GPT+PE模式从事翻译实践

5. 小组作业+汇报

END