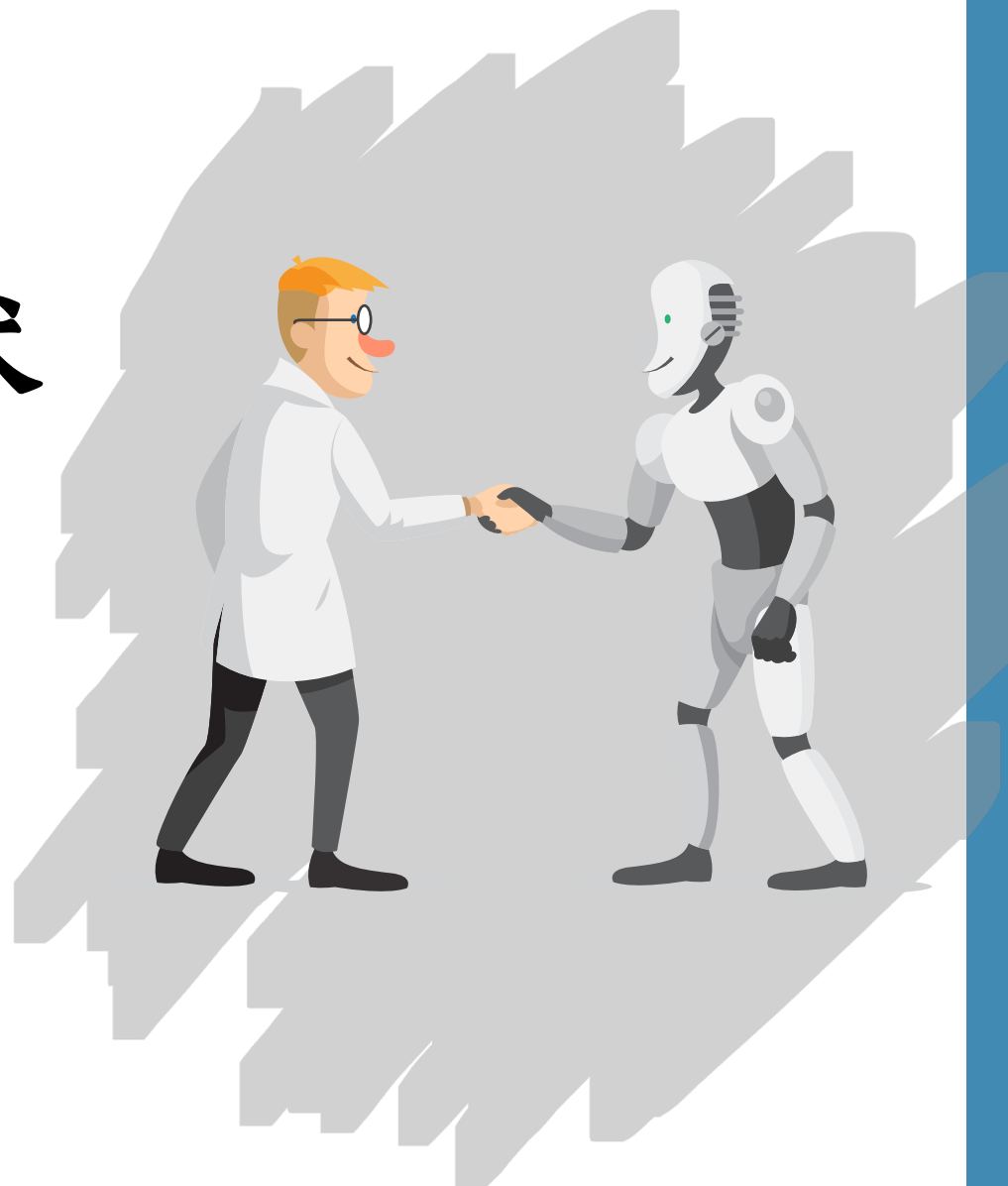


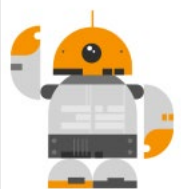
第10讲 机器翻译技术 及应用概论

黄婕

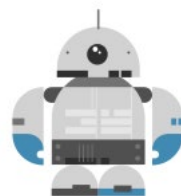
2025-5-15



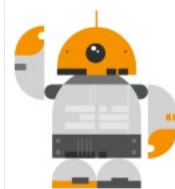
本节内容



1. 初识机器翻译

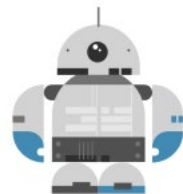


2. 机器翻译的瓶颈

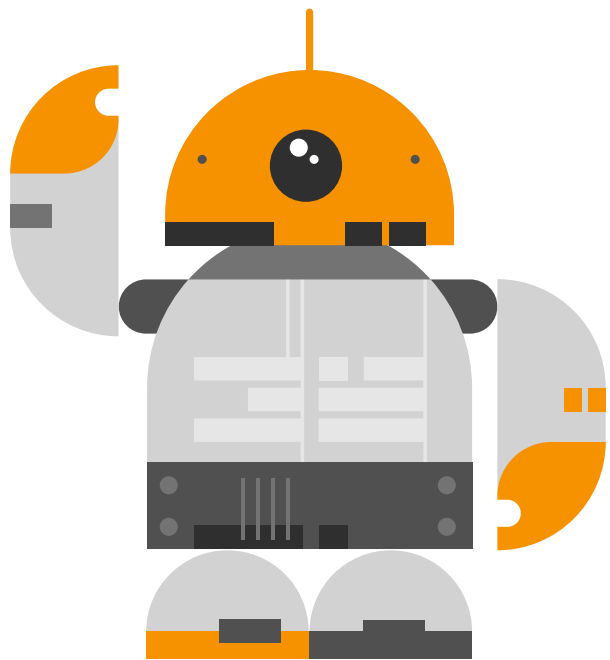


3. 机器翻译的发展历史

4. 挑战和前景



5. 小组作业+汇报



1. 初识机器翻译

什么是机器翻译?

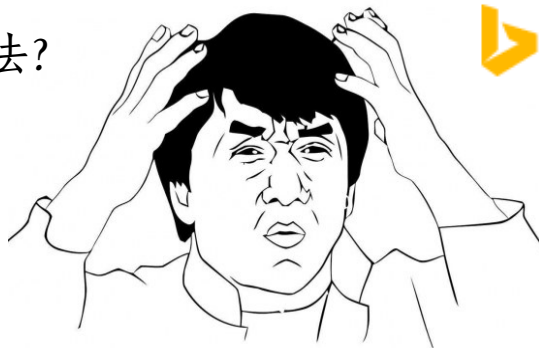


基于实例的翻译方法?

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?



基于统计的方法?



国内有哪些机器翻译团队？



中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES



哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY



東北大學
Northeastern University



清華大學
Tsinghua University



廈門大學
XIAMEN UNIVERSITY



南京大學
NANJING UNIVERSITY

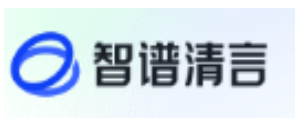
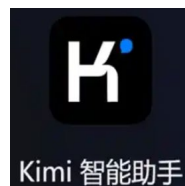
常见机器翻译产品有哪些？



垂直领域机器翻译



常见的大语言模型有哪些？



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

- 个人出国翻译
- 智能问答翻译
- 电话口译
- 国际会议同传
- 社区口译
- 电子邮件翻译
- 即时通信翻译
- ...

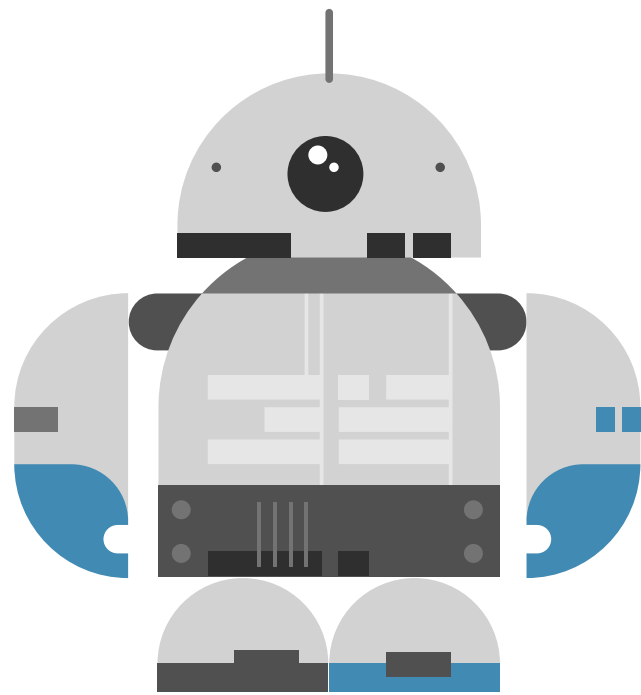
信息分析

- 网页翻译
- 跨境电商翻译
- 跨语言舆情分析
- 信息安全翻译
- 行业新技术追踪
- 专利信息检索
- 自动摘要
- ...

跨文化传播

- 专利摘要翻译
- 影视字幕翻译
- 图书出版翻译
- 新闻传媒翻译
- CAT调用机器翻译
- 交互式机器翻译
- 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

 一者科技

5款主流机器翻译质量评测

评测金融、IT、法律三个领域中英各50句，5款主流机器翻译中到英BLEU对比结果

	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

练习：机器翻译的质量评估

常见机器翻译自动质量评估的指标

BLEU

2002年 – 基于n-gram匹配的统计方法

METEOR

2007年 – 融合语言学资源的混合方法

TER

2006年 – 基于编辑距离的算法

COMET

2020年 – 基于深度学习的神经网络方法

MT 自动质量评估的核心工作原理

BLEU

- 计算机器翻译与参考翻译之间的n-gram重叠度，依赖于精确词汇匹配

TER (Translation Edit Rate)

- 计算将机器翻译转换为参考翻译所需的最少编辑操作数（插入、删除、替换、移动），基于表面形式的比较

METEOR








- 结合精确率、召回率和词序信息，同时考虑同义词、词干和词形变化，语言资源依赖（需要同义词词典等）

新指标：COMET

- 使用预训练多语言模型（如XLM-RoBERTa）学习源文本、机器翻译和参考翻译之间的关系

Summary



	BLEU	TER	COMET
Evaluation Focus	 Accuracy	 Word error rate	 Transiation quality
Approuch	Matching h-grams	Edit distance	Neural model 
Relationship to Human Judgment	 Indirect	 Indirect	 Direct
Complexity	★★	★★★★☆	★★★★★
Correlation with Human Judgment	★★★★☆	★★★★☆	★★★★★☆☆
Resource Requirements	★★★★☆	★★★☆☆	★★★★★☆☆

练习: BLEU, TER, METEOR

▶ 常见指标:

▶ BLEU

▶ TER

▶ METEOR

▶ 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

计算结果:

指标名称	计算值
BLEU ?	0.3862
TER ?	0.4828
METEOR ?	0.5691
综合得分 ?	49

计算评测指标

待评测译文

个目标的资金：
基础设施：提供技术和财政支助，将服务不足的教育组织连接到互联网；
协作：促进参与项目、活动和与其他教育组织建立伙伴关系的机会；
可持续性：帮助学校确保项目在最初拨款后的连续性。

教育项目为金太阳网络学校的学生和教师提供专门的第
一手和在线学习体验，使用基于项目的方法。这些经验

参考译文

关系提供机会。
可持续性：帮助学校确保计划在得到最初许可后能够持续下去。

教育计划通过基于项目的方法，为金太阳网络学校的学生和教师提供第一手的专业知识和在线学习体验。频繁举行的实际操作研讨会和协作项目创建了一个由来自不同国家的人组成的全球学习社区，从而解决了与科学和技术相关的各个主题方面的要求。不论是自己单独学

练习：COMET

COMET 译文质量评估

- COMET (Crosslingual Optimised Metric for Evaluation of Translation) 是由 Unbabel 团队开发的开源神经框架，专注于机器翻译质量评估 (MT Quality Estimation) 。
- 通过神经网络模型预测人类对翻译质量的主观评价 (如 MQM 分数) ， 是当前与人工评测相关性最高的自动化评估工具之一。

练习：尝试使用COMET进行翻译质量评估

方法1：直接使用Python 调用

方法2：使用大语言模型计算？

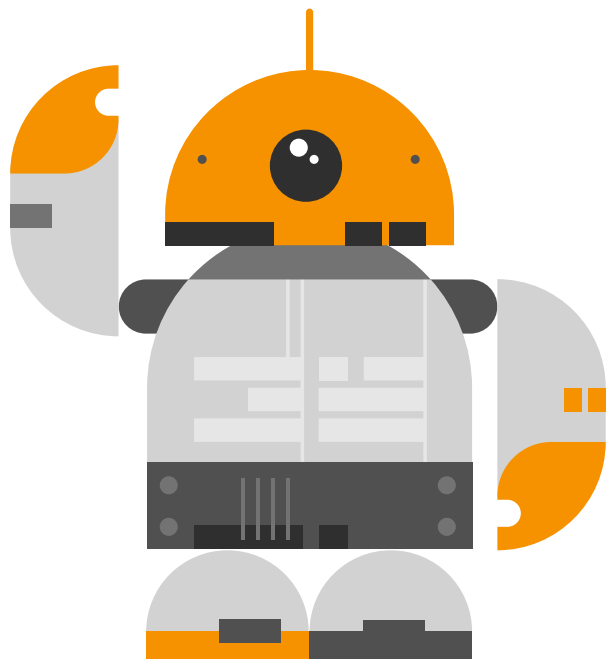
方法3：找到现成的工具？



试一试！

步骤：阿里云天池 Notebook

1. 打开阿里云天池 Notebook，登录我的实验室：
<https://tianchi.aliyun.com/notebook-ai/>
2. 上传 ipython script:
[2_Cases/2_ipython_script/MT_quality_evaluation.ipynb](#)
3. 编辑 Notebook, 进入编程控制台，左上角上传ST+TT文档
([ST.txt](#), [TT_deepl.txt](#), [TT_deepseek.txt](#), [TT_human.txt](#))
4. 开始逐行运行代码块 ([Shift +Enter](#)) 。



3. 机器翻译的发展历史

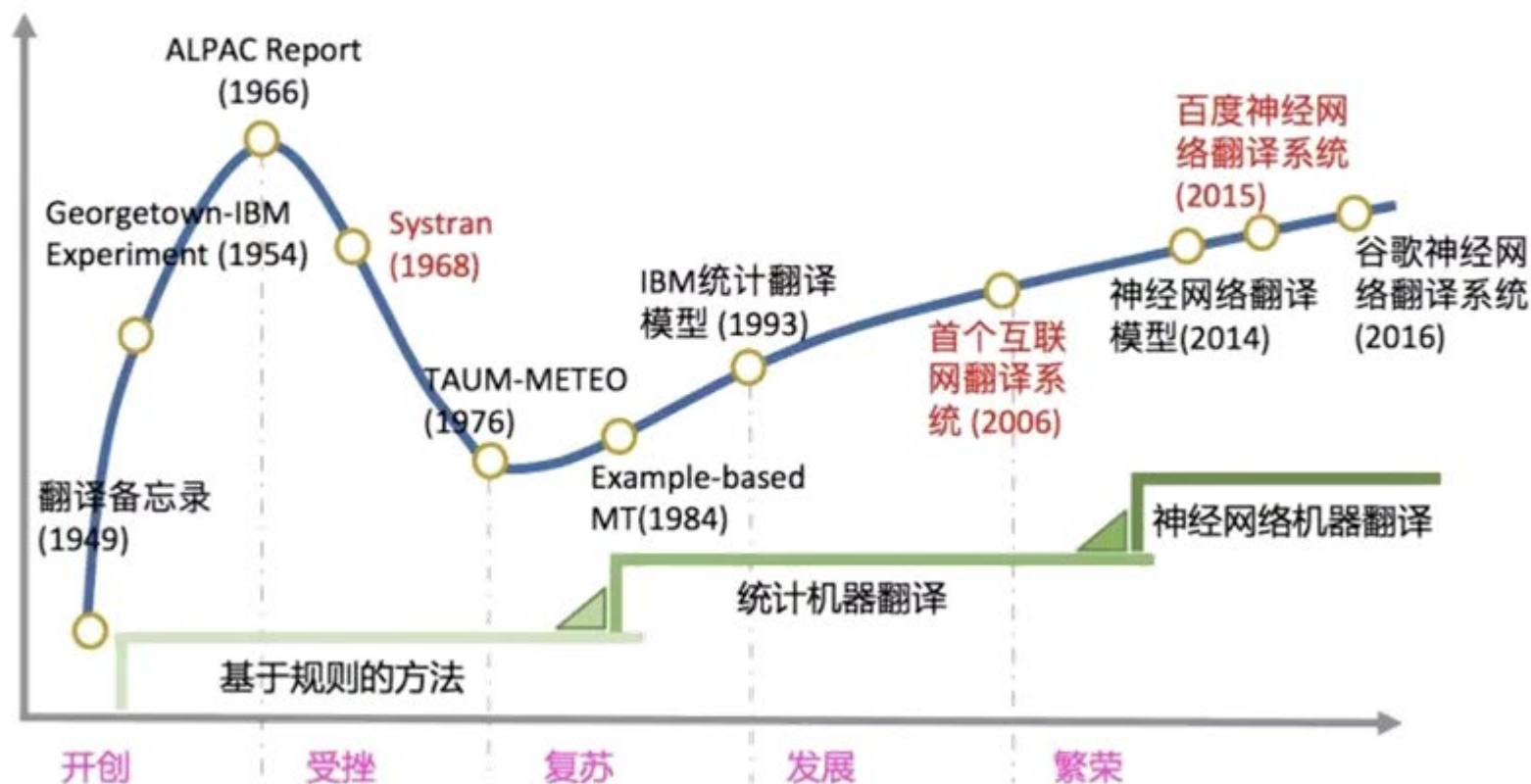
罗塞塔石碑？

- 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。



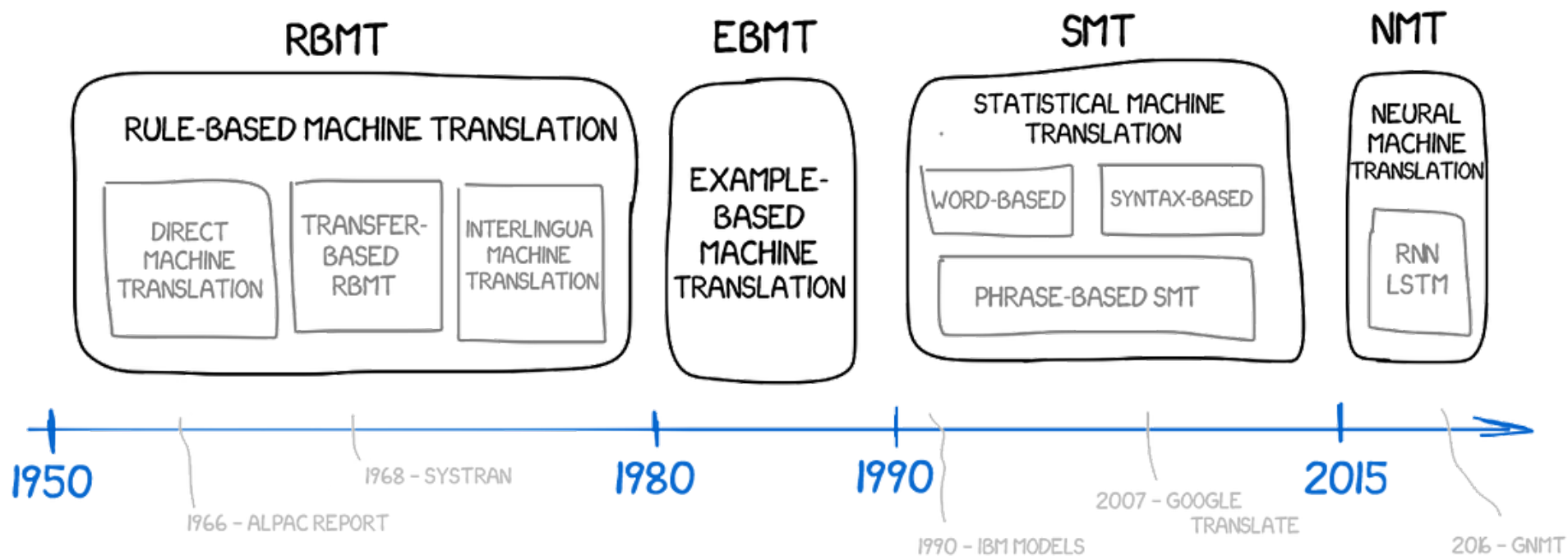
机器翻译发展历程

机器翻译发展历程



机器翻译：几个关键阶段

A BRIEF HISTORY OF MACHINE TRANSLATION



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**)

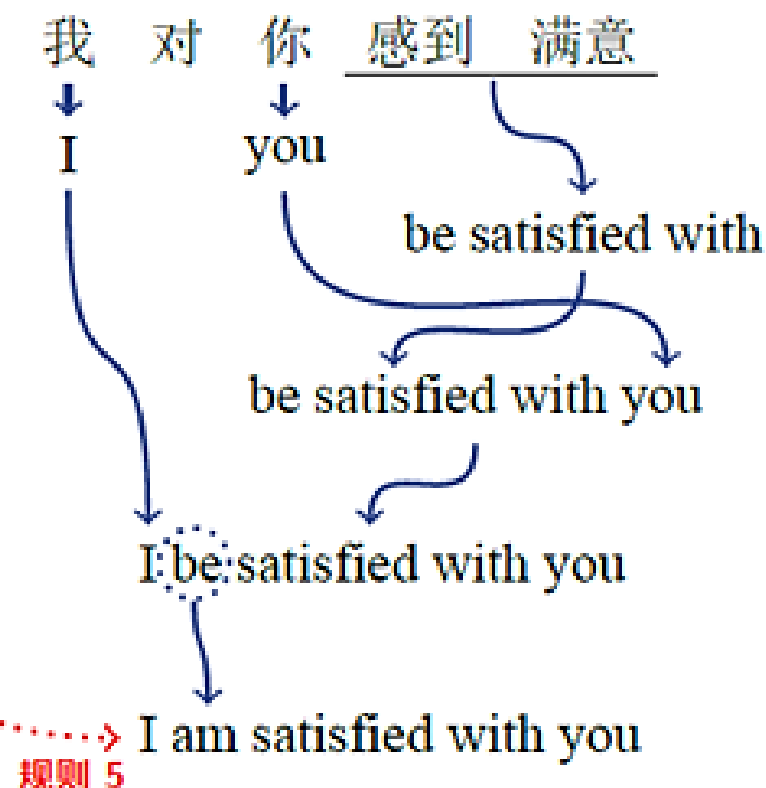
四种机器翻译技术的工作原理是什么？

阅读文献：第1章-机器翻译简介-肖桐-朱靖波.pdf

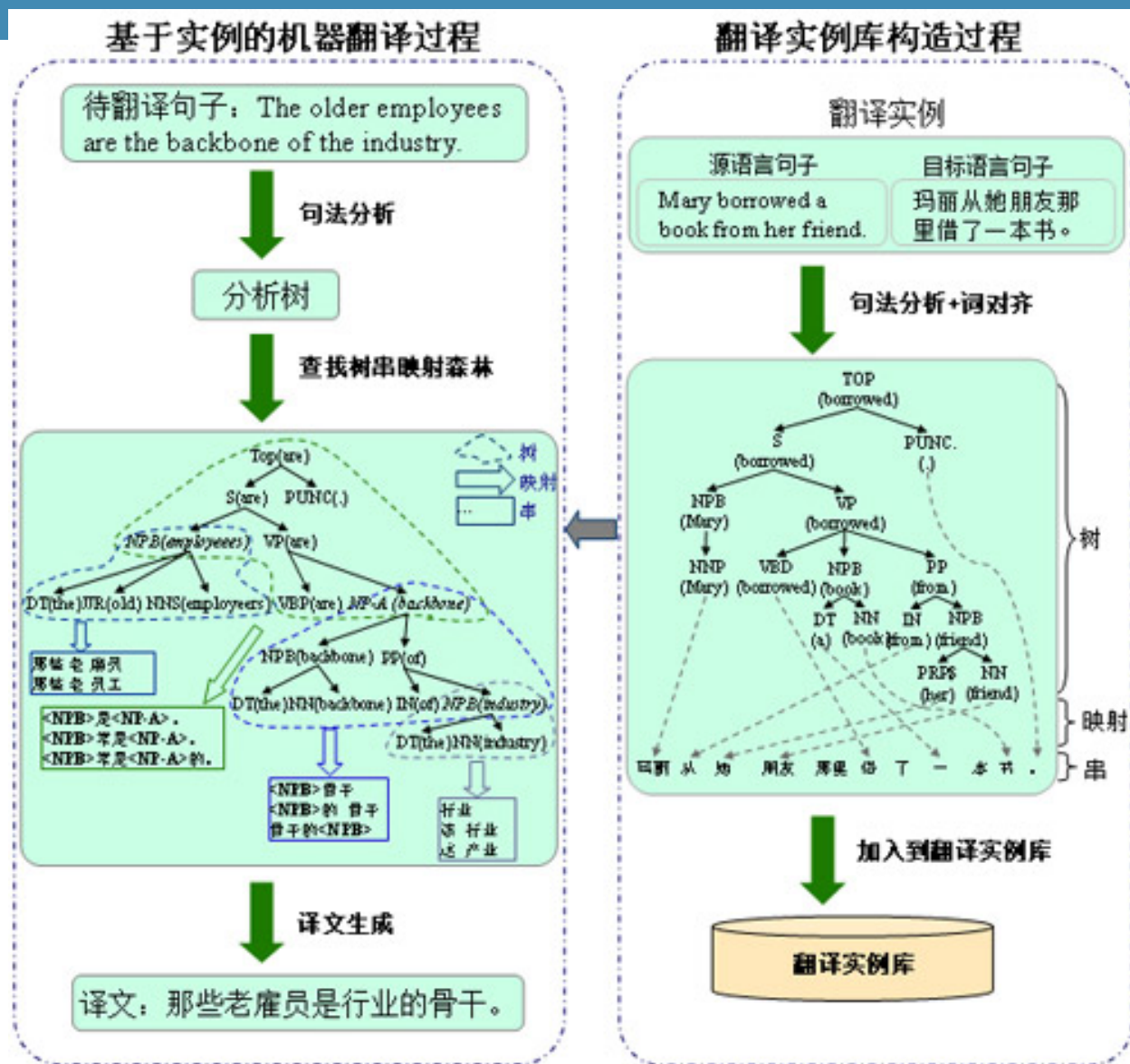
1. 基于规则的机器翻译 (RBMT)

资源：规则库

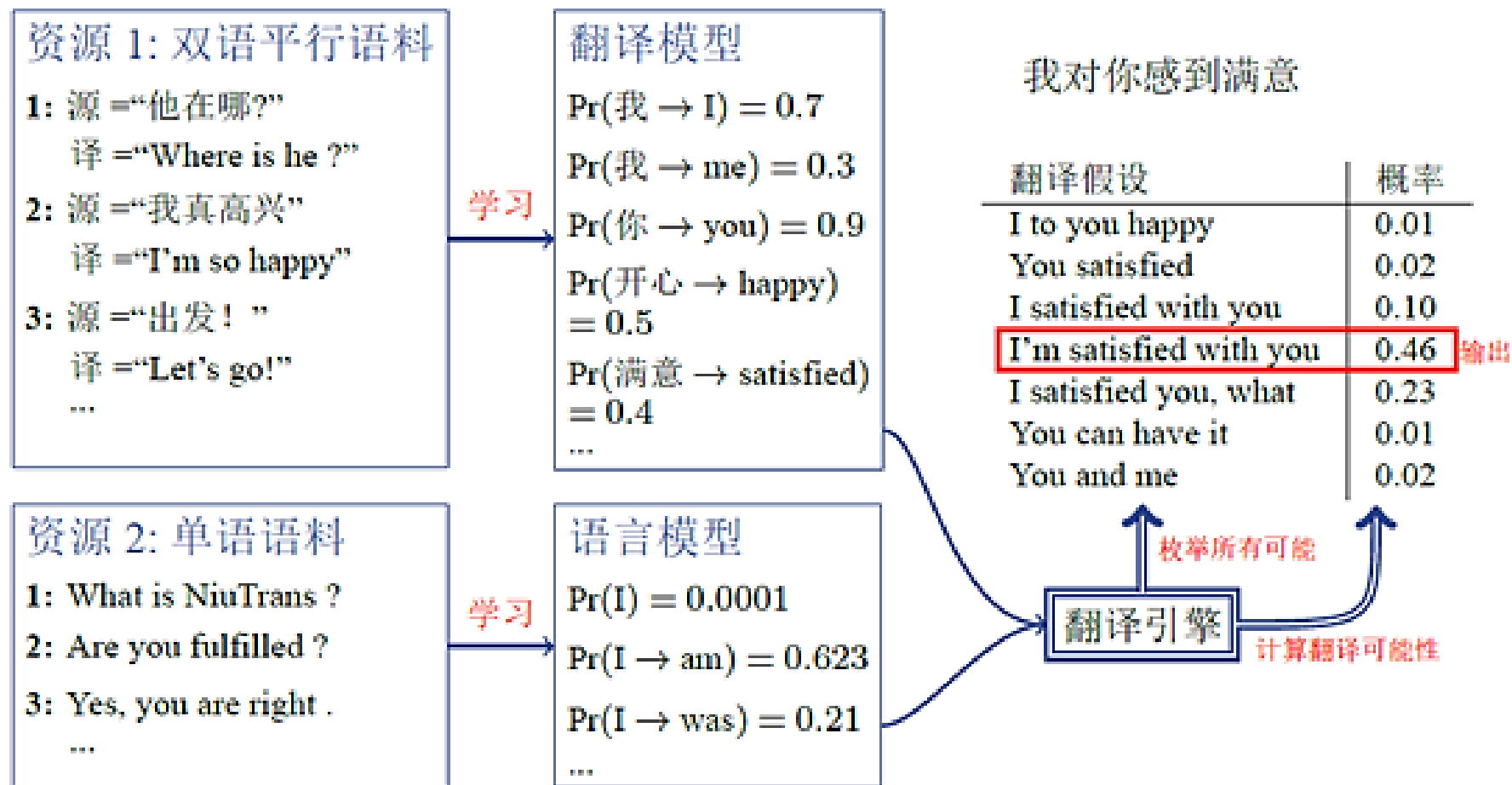
- 1: If 源 = “我”, then 译 = “I”
- 2: If 源 = “你”, then 译 = “you”
- 3: If 源 = “感到满意”,
then 译 = “be satisfied with”
- 4: If 源 = “对... 动词 [表态度]”
then 调序 [动词 + 对象]
- 5: If 译文主语是 “I”
then be 动词为 “am/was”
- 6: If 源语是主谓结构
then 译文为主谓结构



2. 基于实例的机器翻译 (EBMT)



3. 基于统计的机器翻译技术



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

$$\hat{T} = \underset{T}{\operatorname{argmax}} p(T) \times p(S|T)$$



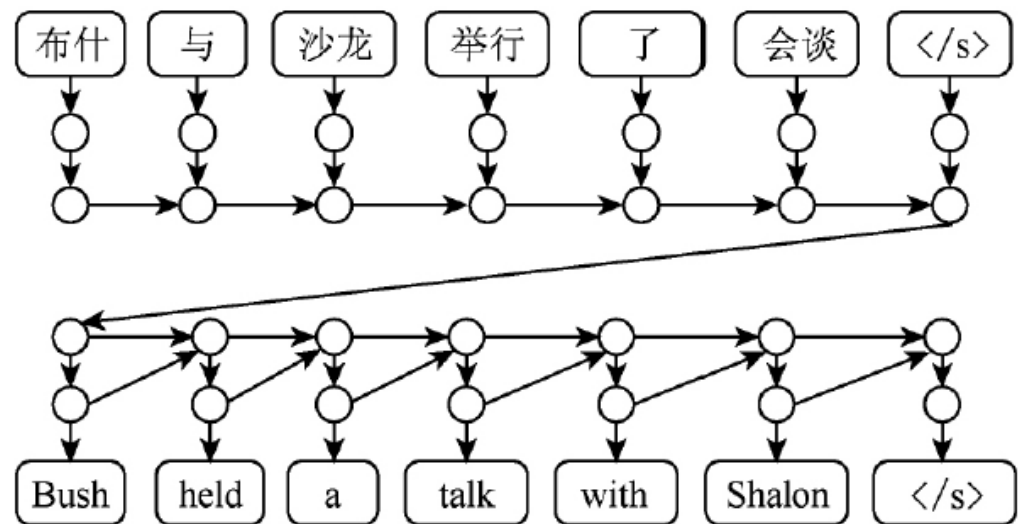
$$\log p(t|s) = \sum_{i=1}^n \lambda_i h_i(t, s)$$

对数线性模型

The weather is / nice / today / .

今天 / 天气 / 很好 / 。

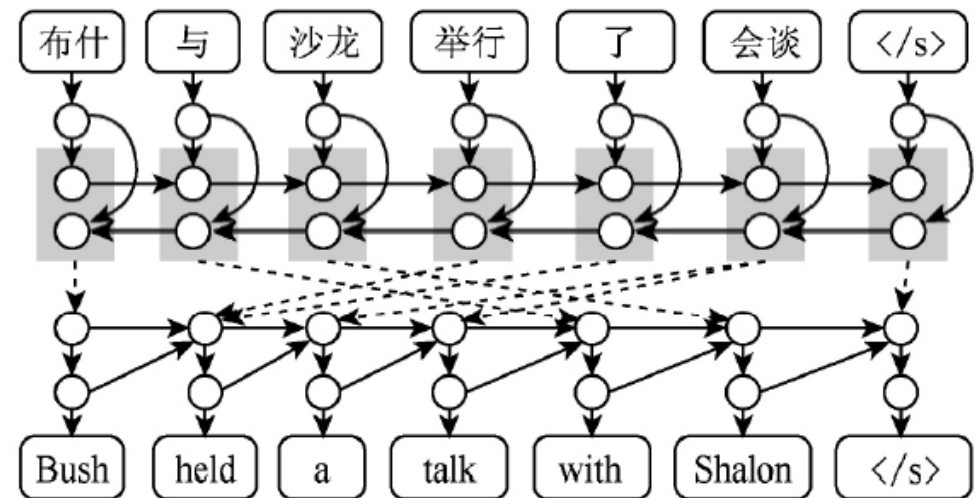
4. 神经网络机器翻译



○ Word Embeddings and Hidden Layers in the Neural Network

Fig. 2 The encoder-decoder framework

图 2 编码器-解码器框架



■ Concatenations of Two Hidden States
— Connections Between Layers in the Neural Network
--- Dominant Attentional Connections
○ Word Embeddings and Hidden Layers in the Neural Network

Fig. 3 Attention-based neural machine translation

图 3 基于注意力机制的神经机器翻译

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

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。

循环神经网络
(RNNs) 或者注
意力机制

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。

词嵌入 (Word
Embeddings)

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。

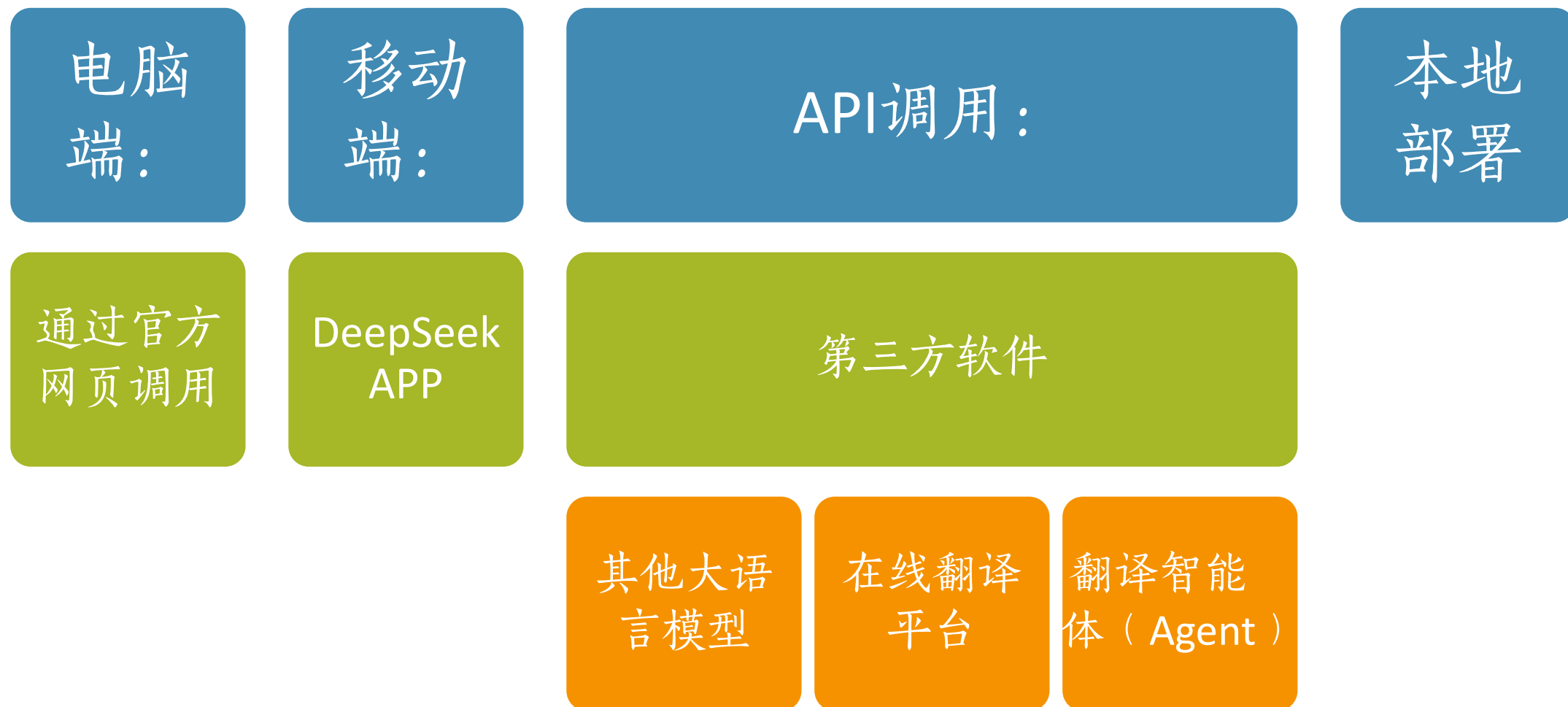
端到端学习

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。

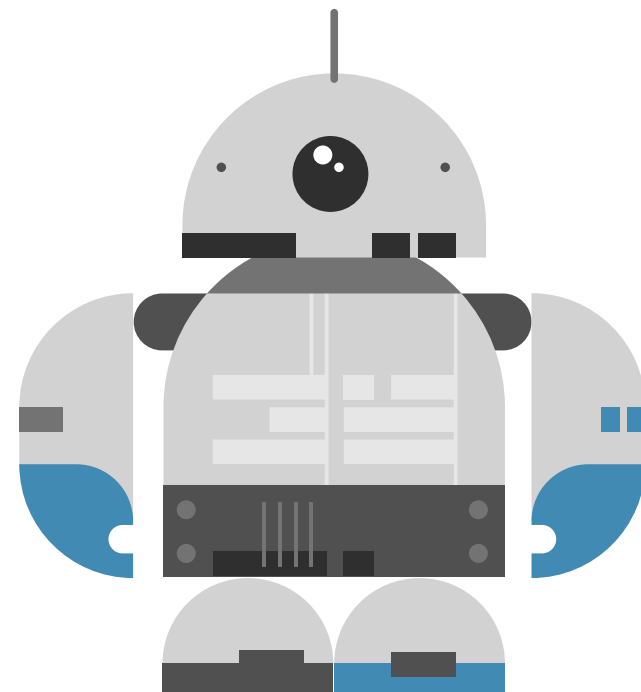
大规模数据集的
训练

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：大语言模型翻译应用—DeepSeek为例



4.挑战 and 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.

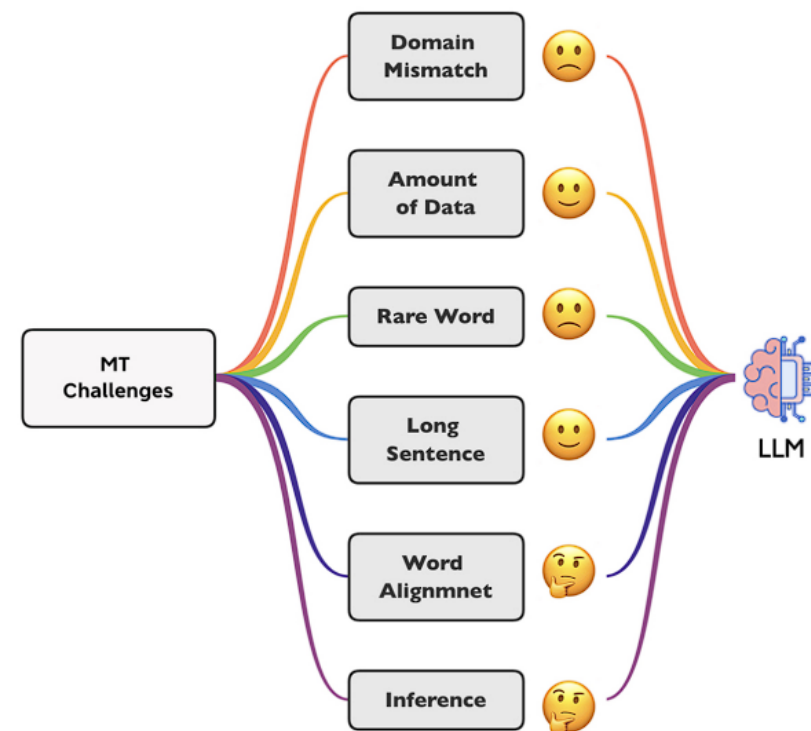


Figure 1: The six MT challenges are revisited in the context of LLM. Emojis symbolize the key findings: a 'smiley face' for issues largely addressed, a 'thinking face' indicating partial alleviation with ongoing concerns, and a 'sad face' for persisting unresolved challenges. This can be treated as the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

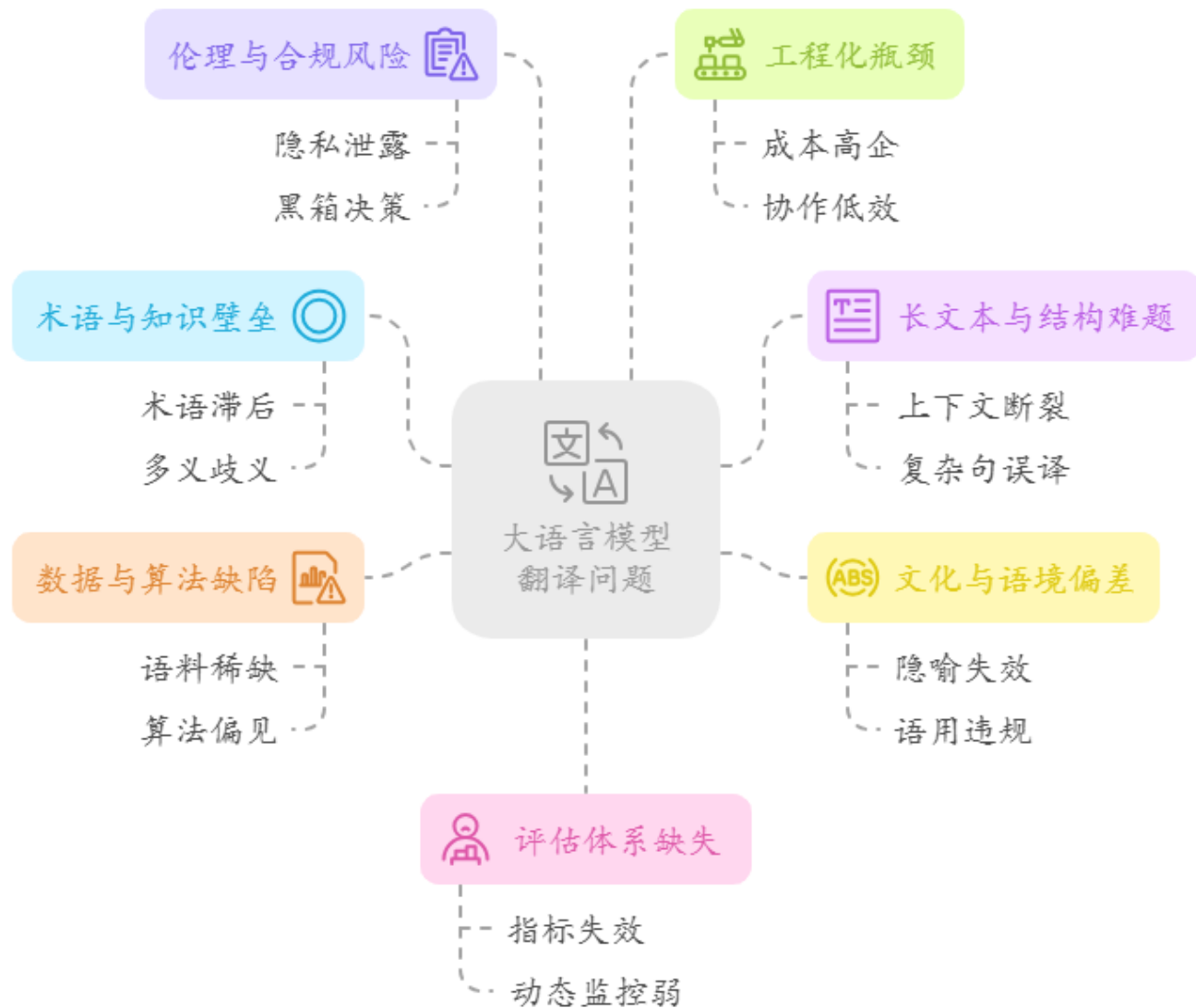
机器翻译存在的问题：2013

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



机器翻译与大语言模型翻译存在的问题



机器翻译的技术应用前景

垂直行业机器翻译

交互式机器翻译

大语言模型机器翻译

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>



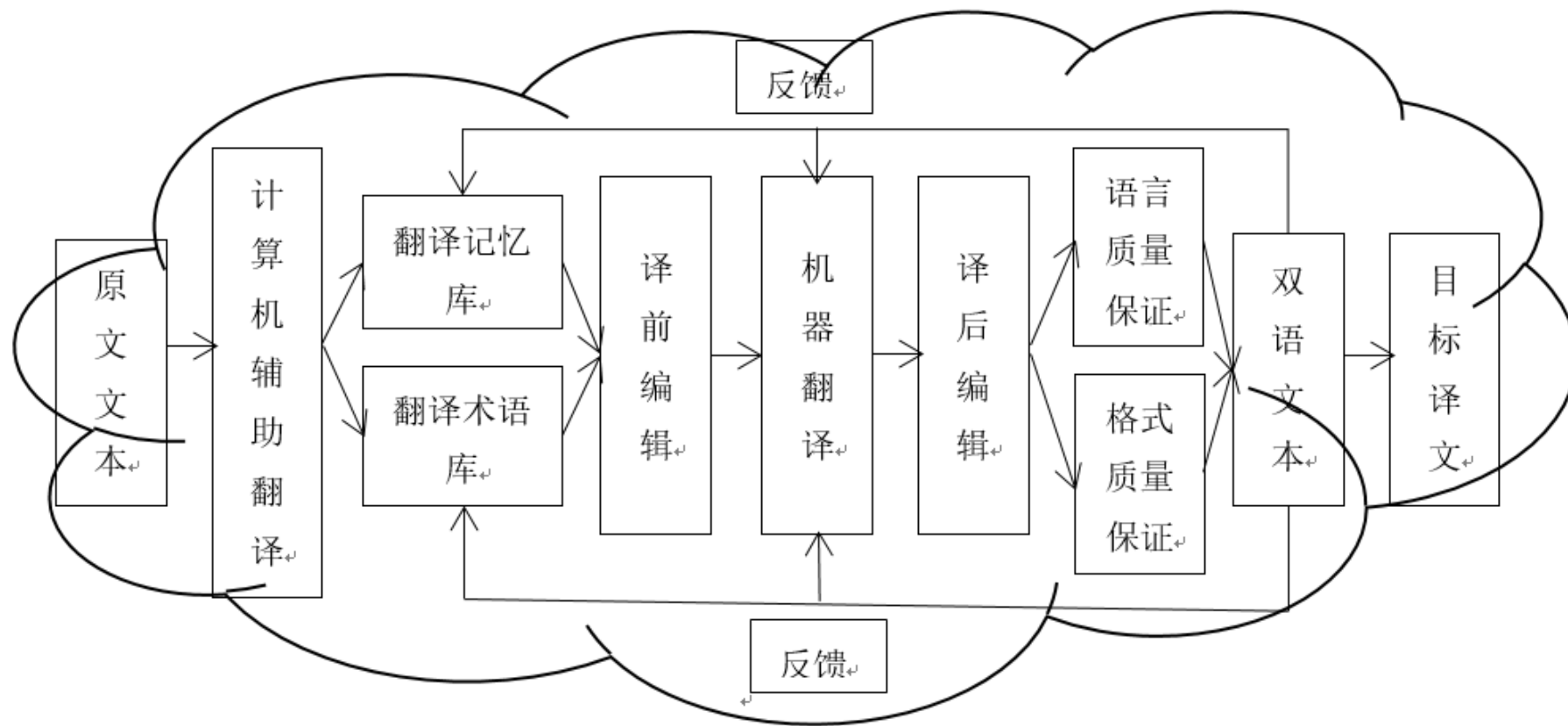
前景3: 大语言模型LLM机器翻译: ChatGPT



Microsoft Edge浏览器的“扩展”插件



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

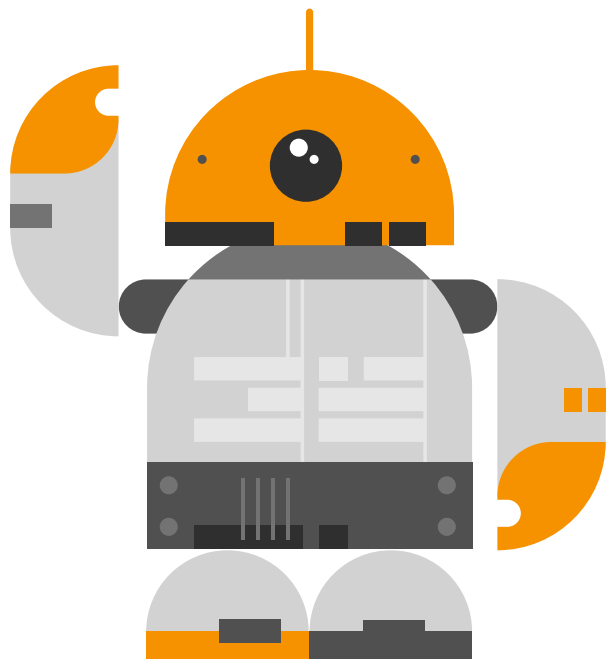
创建和对齐语料库，为机器翻译输送高质量数据

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

小组作业：CAT软件自学

- 分组自学以下CAT软件，每个组自选1种CAT工具，不要重复：
- 每个小组提交内容：
 - 1份PPTX文件，内容是与Trados Studio进行厂家对比，功能对比与演示，价格对比，使用体验对比
 - 往年小组汇报的参考文件：4_Groupwork_reference 文件夹
- 提交日期：2025年5月29日之前（2周后）
- 提交方式：各组组长发邮件到 jie.huang@uibe.edu.cn
- 小组汇报 (presentation):
 - 日期：**6月5日：全部四组**
 - 时长：每组20分钟左右，最长不要超过30分钟

小组CAT软件自学

组长	CAT软件	软件获取方式	说明
	memoQ	www.kilgray.com 下载	30天内免费
	LanguageX	https://languageX.com/ 免费注册	
	YiCAT	https://www.yicat.vip/ 免费注册	
	云译客	https://pe-x.iol8.com/ 免费注册	
	Crowdin	https://crowdin.com/	
	Phrase TMS	https://phrase.com/	
	译马网	http://www.jeemaa.com 免费注册	
	或其他自选软件	提前告知老师	

END