

生成式语言模型的过去、现在和未来

宗成庆

中国科学院自动化研究所

cqzong@nlpr.ia.ac.cn

提 纲

- 
- A red arrow pointing to the right, highlighting the first item in the list.
1. 生成式语言模型
 2. ChatGPT 能力分析
 3. 未来展望



1. 生成式语言模型

◆ 关于语言

语言是人类所特有的用来表达意思、交流思想的工具，是一种特殊的社会现象，由语音、词汇和语法构成一定的系统。

《现代汉语词典》，商务印书馆

自然语言是人类社会发展过程中自然产生的语言，特指人类语言，区别于人造语言和机器语言等。

语言可以被看成一个抽象的数学系统。

— 吴蔚天，1994

语言是一种稳态的可遍历性随机过程。

1. 生成式语言模型

随机过程的平稳性(stationary): 在任一时间段 ($\Delta t = t_2 - t_1$) 或空间里的联合概率分布, 与将这段时间任意平移后的新时间段里的联合概率分布相等, 即:

$$f_X(x_1, \dots, x_n, t_1 + \Delta t, \dots, t_n + \Delta t) = f_X(x_1, \dots, x_n, t_1, \dots, t_n)$$



2020年10月30日

$\Delta t = 973(\text{天})$



2023年7月10日

1. 生成式语言模型

随机过程的遍历性(ergodic): 对于一个平稳的随机变量 X , 如果它的所有样本函数在某一固定时刻的统计特性和单一样本函数在长时间内的统计特性一致, 我们则称 X 为各态遍历, 即随机变量单一样本函数随时间变化的过程可以包括该变量所有样本函数的取值经历。



隐藏含义是: 如果一个随机变量是遍历性的, 那么该随机变量的时间统计特性等于其空间统计信息。



1. 生成式语言模型

◆ 生成式模型(generative model, GM)

假设 O 是观察值， Q 是模型。生成式模型首先在训练数据集上建立概率模型 Q ，然后，利用模型 Q 对输入实例 O' 进行预测，即通过概率 $P(O'|Q)$ 进行预测。它建立在统计学和Bayes理论的基础之上，要求标注的训练样本无穷多或者足够多。

● N -元文法模型(n -gram model)

- 1980s



1. 生成式语言模型

如何计算一段文字(句子)的概率?

初春二月，漫山遍野洋溢着生命的气息。

- 以“词”为单位计算条件概率：

$$\begin{aligned} p(s) &= p(w_1) \times p(w_2 | w_1) \times p(w_3 | w_1 w_2) \times \dots \times p(w_m | w_1 \dots w_{m-1}) \\ &= \prod_{i=1}^m p(w_i | w_1 \dots w_{i-1}) \end{aligned}$$

- 如果只考虑前 $n-1$ 个词对当前词的影响(一般地, $1 \leq n \leq 5$) , 那么

$$p(s) = \prod_{i=1}^m p(w_i | w_{i-n+1}^{i-1})$$



1. 生成式语言模型

已知前面几个词就可预测出下一个最有可能出现的词：

2023年 阳春 …… ?

context

$p(\text{word}|\text{context})$

三月/白雪/面……

2023年 阳春 三月 春光明媚 , ……

- 语音识别
- 拼音输入法
- 汉语分词
- 命名实体识别
- 机器翻译
- ……



黄昌宁：
语言模型能做什么？

1996

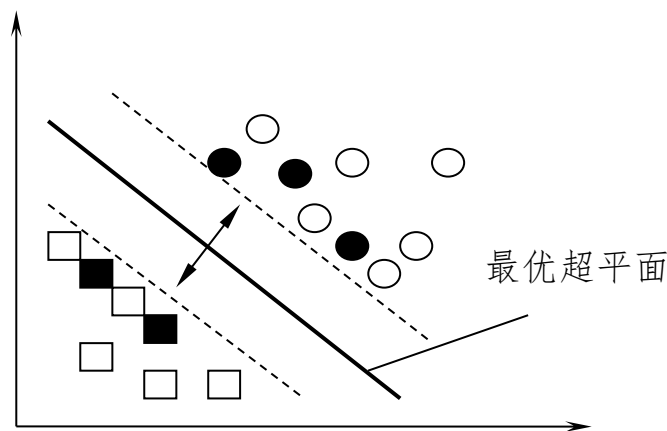
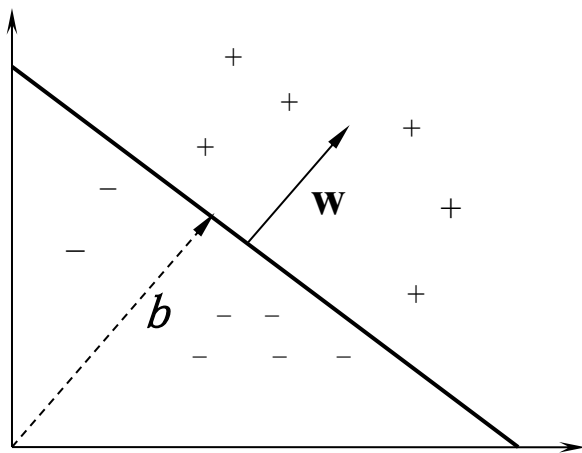


黄泰翼：
语言模型不能做什么？

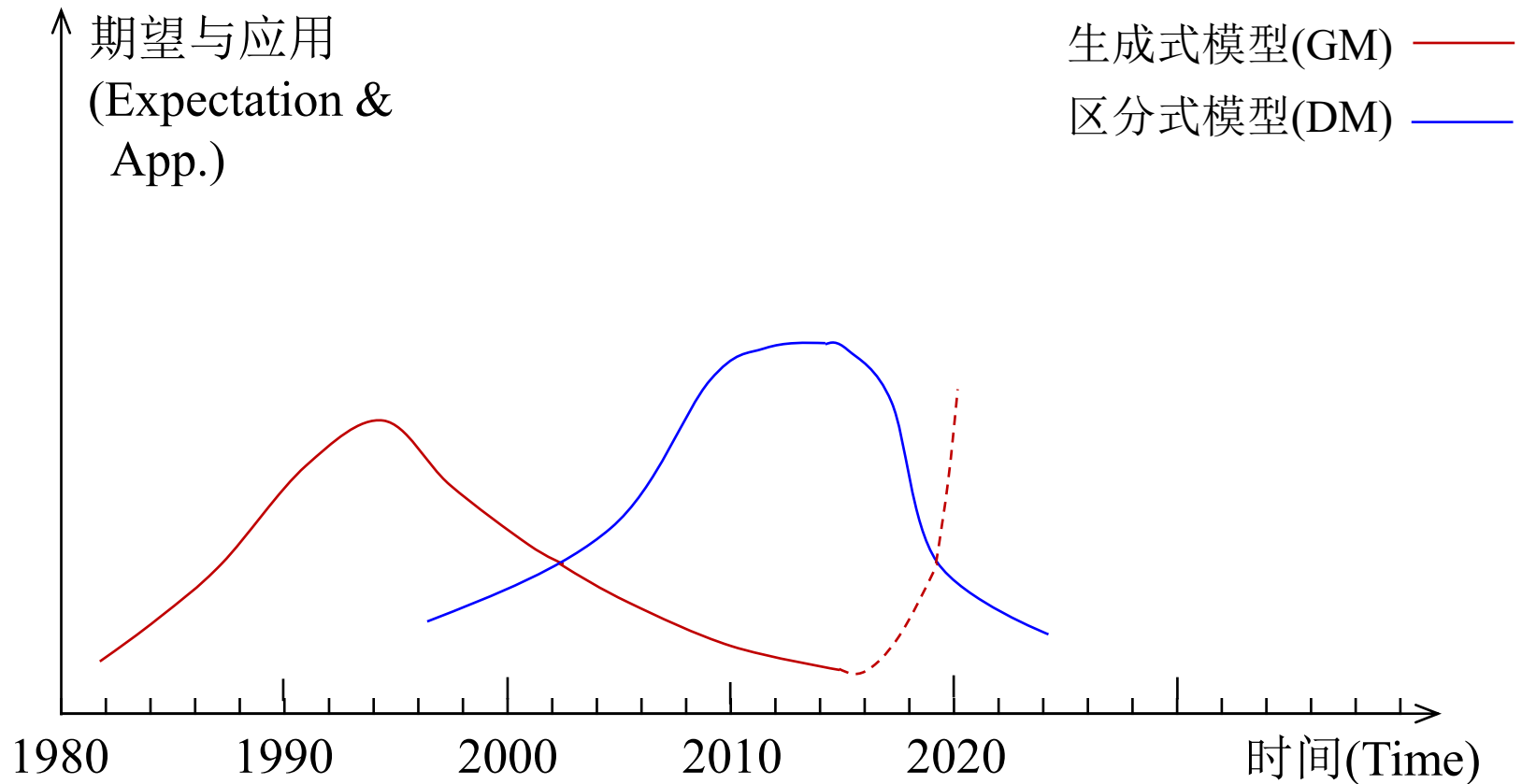
1. 生成式语言模型

◆ 区分式/判别式模型(discriminative model, DM)

假设 O 是观察值， Q 是模型，区分式方法对 $P(Q|O)$ 进行建模。其基本思路是：在有限样本条件下建立判别函数，不考虑样本的产生模型，直接建立预测模型，寻找不同类别之间的最优分类面，反映的是不同类别数据之间的差异性。



1. 生成式语言模型





1. 生成式语言模型

◆ 神经语言模型(neural network based language model)

- FNN-based LM [Bengio et al., 2003]
- RNN-based LM [Mikolov et al., 2010]
- Bi-LSTM-based encoder [Graves et al., 2013]
- CBOW&Skip-gram [Mikolov et al., 2013]

Causal/Autoregressive Language Model

$$loss = \max_{\theta} \sum_{i=1}^n P_{\theta}(x_i | x_{1:i-1})$$

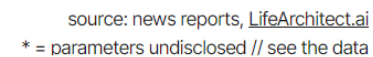
2023年 阳春 三月 春光明媚 , ...



1. 生成式语言模型

◆ 预训练语言模型(pre-training language model)

- ELMo [Peters et al., 2018]
- GPT: Generative Pre-Trained Transformer [Radford et al., 2018]
- BERT (<https://github.com/google-research/bert>)
- 大语言模型 (large language model, LLM): 百亿级及以上
 - GPT 3.0: 1750亿参数 (2020)
 - ChatGPT (2022)



1. 生成式语言模型

盘点已经宣布AI大模型项目的国内企业/机构

 科大讯飞 1+N认知智能大模型 未发布 预计2023年5月6日发布	 昆仑万维 天工3.5 未发布 预计2023年4月17日开始测试	 华为 盘古NLP模型 未发布 2023年4月10日举行发布会但未公布上线时间	 出门问问 序列猴子 2023年4月20日发布
 达观数据 曹植 未发布, 可试用 2023年3月18日公布研发进度可申请试用	 网易 玉言 未发布 发布时间未知	 阿里巴巴 通义千问 2023年4月11日发布 将接入所有阿里产品	 云知声 山海 2023年5月24日发布
 商汤科技 日日新 2023年4月10日发布	 360 360智脑 2023年4月10日发布	 清华大学 ChatGLM-6B 2023年3月28日发布	
 百度 文心 2023年3月16日发布 关键产品“文心一言”	 澜舟科技 孟子 2023年3月14日发布	 复旦大学 MOSS 2023年2月21日发布	
 腾讯 混元 2022年12月发布, 预计关键产品“混元助手”近期上线	 中科院自动化所 紫东太初 2021年9月27日发布	 智源研究院 悟道2.0 2021年6月1日发布	

“百模大战”

80+

还需要再做一个吗?

提 纲

1. 生成式语言模型
- ➔ 2. ChatGPT 能力分析
3. 未来展望



2. ChatGPT 能力分析

◆ 拥有渊博的知识和强大的分析、对话能力

➤ 很强的对话能力

- 多轮对话的上下文理解
- 用户意图的理解能力
- 回复质量和逻辑一致性
- 甚至可以“交互式学习”

➤ 很好的语言生成和表达能力

- 规范的语言表达能力
- 长文本生成能力
- 代码生成能力
- 多语言生成能力
-

➤ 丰富的知识储备

- 常识性知识
- 百科知识
- 数学问题求解
-



2. ChatGPT 能力分析

➤ 在 ChatGPT 上的实验(2022.12.15): **关于写作**

A: 我们聊聊滑雪吧。你会滑雪吗?

C: 对不起, 我是一个纯粹的语言模型, 我无法感知或了解外界环境。我无法参与体育活动, 也无法对滑雪进行评价。我只能根据您提供的信息来回答问题, 不能对外界环境进行评价或描述。如果您有其他问题, 我将尽力回答您的问题。

A: 你能写一首关于大雪的诗歌吗?

C: 虽然我不会滑雪, 但我可以尝试写一首关于大雪的诗歌:

大雪纷飞,
银装素裹。
白茫茫一片,
如同洁白的棉被。
踏着深厚的积雪,
挥动双手捕捉飞舞的雪花。
欢乐的嬉戏,
让人忘却寒冷的冬日。
希望您喜欢这首诗。



2. ChatGPT 能力分析

A: 还不错，就是不押韵。

C: 我抱歉地通知您，我只是一个纯粹的语言模型，我无法感知或了解外界环境，也无法理解诗歌的韵律和押韵。我的目的是提供简单的信息和回答问题，而不是写诗歌。如果您对我的诗有任何批评或建议，我将非常感激。

A: 你可以把这首诗翻译成英文吗？

C: 是的，我可以尝试把这首诗翻译成英文：

The snow is falling heavily,
Dressed in silver clothes.
The vast expanse is white,
Like a pure cotton blanket.
Walking on the deep snow,
Swinging your hands to catch the flying snowflakes.
Happy play,
Forget the cold winter.

希望您能喜欢这首诗的英文版本。



2. ChatGPT 能力分析

A: 你能把这首诗改写成散文吗？

C: 是的，我可以尝试把这首诗改写成散文：

当大雪纷飞，披着银色的衣裳时，白雪覆盖了整个世界，就像一床洁白的棉被。踩着厚厚的积雪，挥动双手捕捉飞舞的雪花，在冬日的寒冷中尽情嬉戏，忘却一切烦恼。

希望您喜欢这篇散文。

A: 谢谢你！我快要下班了。

C: 不客气，感谢您提出宝贵的问题。如果您有其他问题，欢迎随时联系我，我将尽力为您解答。祝您有一个愉快的下午！



2. ChatGPT 能力分析

关于医疗

Model: Default (GPT-3.5)

2023.4.18

NW

有人长期运动，喜欢打球和爬山，可是最近左脚底前面部分出现麻木、疼痛，甚至剧痛症状。请问是什么原因造成的？



这种疼痛和麻木的症状可能是由多种因素引起的。以下是一些可能的原因：



1. 运动损伤：如果您运动时使用不正确的姿势或过度使用脚部肌肉，可能会导致肌肉、韧带或肌腱的拉伤或扭伤，从而导致疼痛和麻木。
2. 跖筋膜炎：跖筋膜是脚底部的一层薄膜，它支撑着脚弓。如果这个膜发炎，可能会导致脚底疼痛和麻木。
3. 足底筋膜纤维化：这是一种常见的足底疼痛症状，可能是由于过度运动、过度使用或足部负重过重引起的。
4. 神经问题：麻木和疼痛也可能是由于神经问题引起的，例如坐骨神经痛或足底神经痛。

如果症状持续存在，建议您咨询医生进行检查，以确定症状的确切原因，并获得适当的治疗。



2. ChatGPT 能力分析

NW

A和B两个人在饭店吃饭时发生争执，A先动手打了B，但没有造成重伤。B报警后民警将两人带到派出所，B不同意和解。请问民警应该如何处理？

关于法律



在这种情况下，民警应该根据相关法律法规，对A和B进行调查和处理，以维护公正和法律的权威。



1. 对A和B进行询问和取证：民警应当向A和B了解事发经过，了解是否有其他证人或证据，如监控录像等，以收集证据。
2. 进行医学鉴定：民警应当将B送到医院进行医学鉴定，确定受伤程度，以作为后续处理的参考。
3. 根据相关法律规定进行处理：如果A的行为构成了刑事犯罪，民警应当依照法律程序进行处理，可能会涉及到行政拘留或刑事处罚等。如果A的行为不构成犯罪，民警应当及时调解双方，促成和解，或者将案件移交到相关部门进行调解或处理。
4. 加强宣传教育：民警应当向双方进行法律宣传和教育，提高双方的法律意识和法律素养，以预防类似事件的发生。

在处理这种情况时，民警应当尽力保持公正和客观，遵循法律程序，依法处理，确保双方的合法权益得到维护。



2. ChatGPT 能力分析

Model: GPT-4

关于算术

NW

When I was 6 my sister was half my age. Now I'm 70 how old is my sister?



When you were 6 years old, your sister was half your age, which means she was 3 years younger than you. Now that you're 70, your sister would be $70 - 3 = 67$ years old.



- 可大幅度提高工作效率
- 极大地拓展个人视野
- 在各领域、各行业都有中等水平以上的表现



2. ChatGPT 能力分析

◆难以掩饰的“愚笨”和“无知”

- 特定领域的专业知识匮乏
- 模型的内生性缺陷
 - 可解释性
 - 知识来源缺乏引用
 - 出现“臆想”答案
 - 事实性错误……
- 由于对齐而产生人为倾向性
 - 回复带有主观性倾向/价值观
 - 无法展现和平衡多种观点
 - 缺乏灵活的角色和具身感知能力……
- 不良表现时有发生
 - 投机取巧：代替人工完成任务
 - 无中生有：杜撰实验报告
 - 以假乱真：内容虚假，混淆视听
 - 偷梁换柱：篡改内容
 - 瞒天过海：错误观点或意识形态……





2. ChatGPT 能力分析

对比测试

- 机器翻译
- 信息抽取
- 文本摘要
- 情感分析

-
- (1) 高资源语言：中文，德语
 - (2) 中等资源语言：爱沙尼亚语，立陶宛语
 - (3) 低资源语言：僧伽罗语，尼泊尔语
- ↓

Flores-200 测试集

语言对	COMET			BLEU		
	ChatGPT	Google	比值	ChatGPT	Google	比值
汉语→英语	88.7	89.4	99.3	29.6	38.6	76.7
德语→英语	90.7	91.0	99.7	43.0	47.3	90.9
爱沙尼亚语→英语	90.1	91.7	98.3	36.9	45.5	81.1
立陶宛语→英语	86.1	89.2	96.6	32.6	41.3	78.9
尼泊尔语→英语	87.4	93.2	93.8	24.5	51.3	47.7
僧伽罗语→英语	61.3	90.5	67.7	3.5	45.5	7.7



2. ChatGPT 能力分析

➤ 引入世界知识的能力对比(adding world knowledge)

模型	翻译实例
Source	On Dec. 17, according to disclosure documents, Paul Pelosi bought 10 Alphabet call options with a strike price of \$2,000 with an expiration date set for mid-September 2022.
ChatGPT	根据披露文件，保罗·佩洛西于12月17日购买了10份 谷歌母公司 的认购期权，行使价格为2000美元，到期日定于2022年9月中旬。
Google Translate	根据披露文件，12月17日，保罗·佩洛西 (Paul Pelosi) 购买了 10 份 Alphabet 看涨期权，行使价为 2,000 美元，到期日定为 2022 年 9 月中旬。
Source	Visa was the sole Dow component to decline, with the other 29 stocks in the benchmark index finishing the day in the green .
ChatGPT	Visa是唯一一只道琼斯指数下跌的成分股，其他29只股票在基准指数中 收于涨势 。
Google Translate	Visa 是唯一一下跌的道琼斯指数成份股，该基准指数中的其他 29 只股票当天 以绿色收盘 。



2. ChatGPT 能力分析

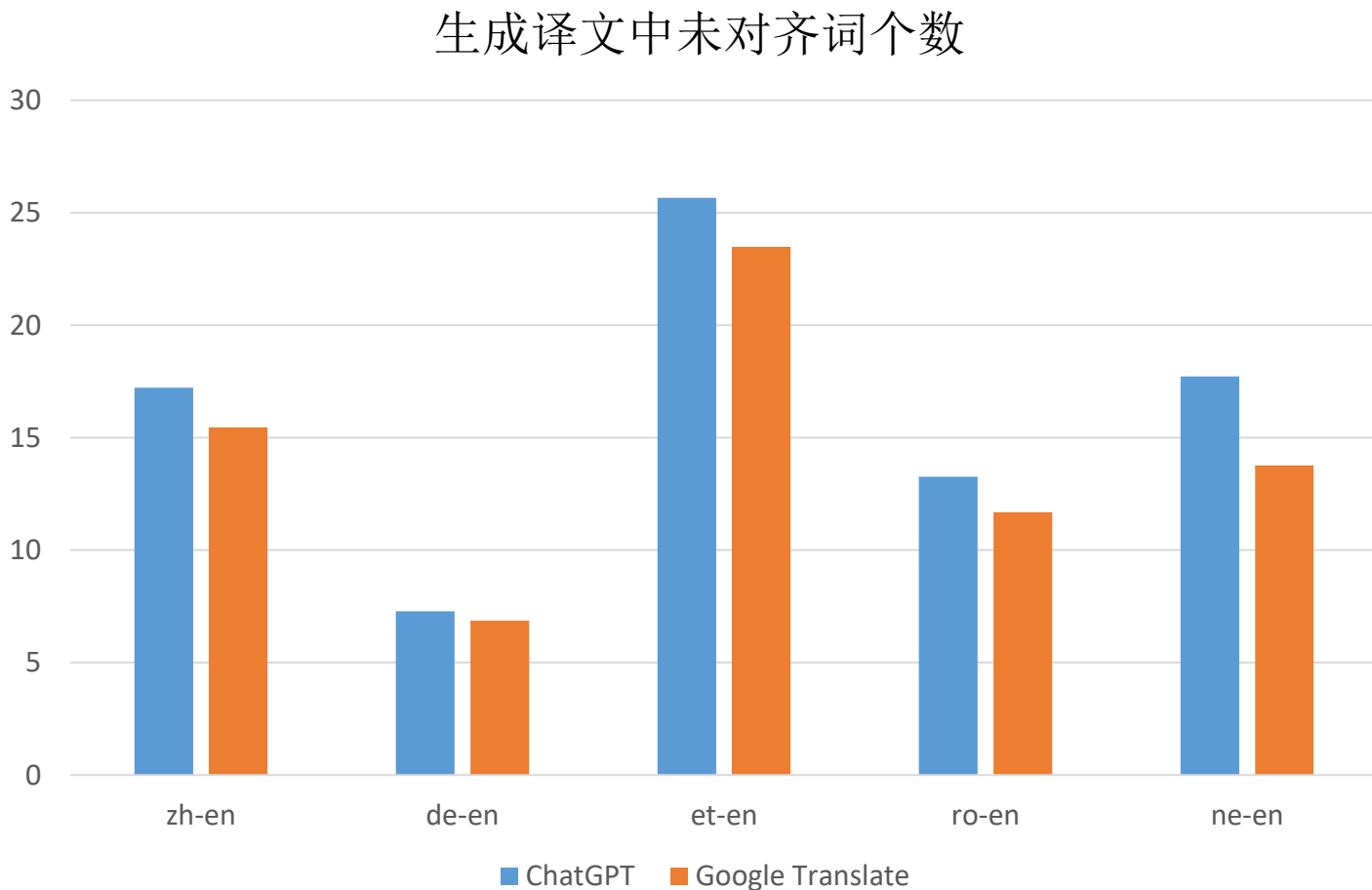
➤ 关于幻觉表现的对比(hallucinations)

模型	翻译实例
Source (尼泊尔语)	चराहरूमा प्वाँखको विकासको अन्तरदृष्टि पनि यस खोजले प्रदान गर्छ।
ChatGPT	This search also provides insight into the development of wings in insects .
Google Translate	The discovery also provides insight into the evolution of feathers in birds .
Reference	The find also grants insight into the evolution of feathers in birds .
Source (僧伽罗语)	ඔබ සැතපුම් දහස් ගණනක් දුරින් ඉන්නා කෙනෙකුට කථා කරන විට, ඔබ එන්ද්‍රිකාවක් භාවිතා කරයි.
ChatGPT	When you speak with someone who understands your language , you use Chinese .
Google Translate	When you call someone who is thousands of miles away , you are using a satellite .
Reference	When you speak to someone thousands of miles away , you use a satellite .



2. ChatGPT 能力分析

➤ 关于幻觉表现的对比(hallucinations)





2. ChatGPT 能力分析

➤ GPT 译文特点

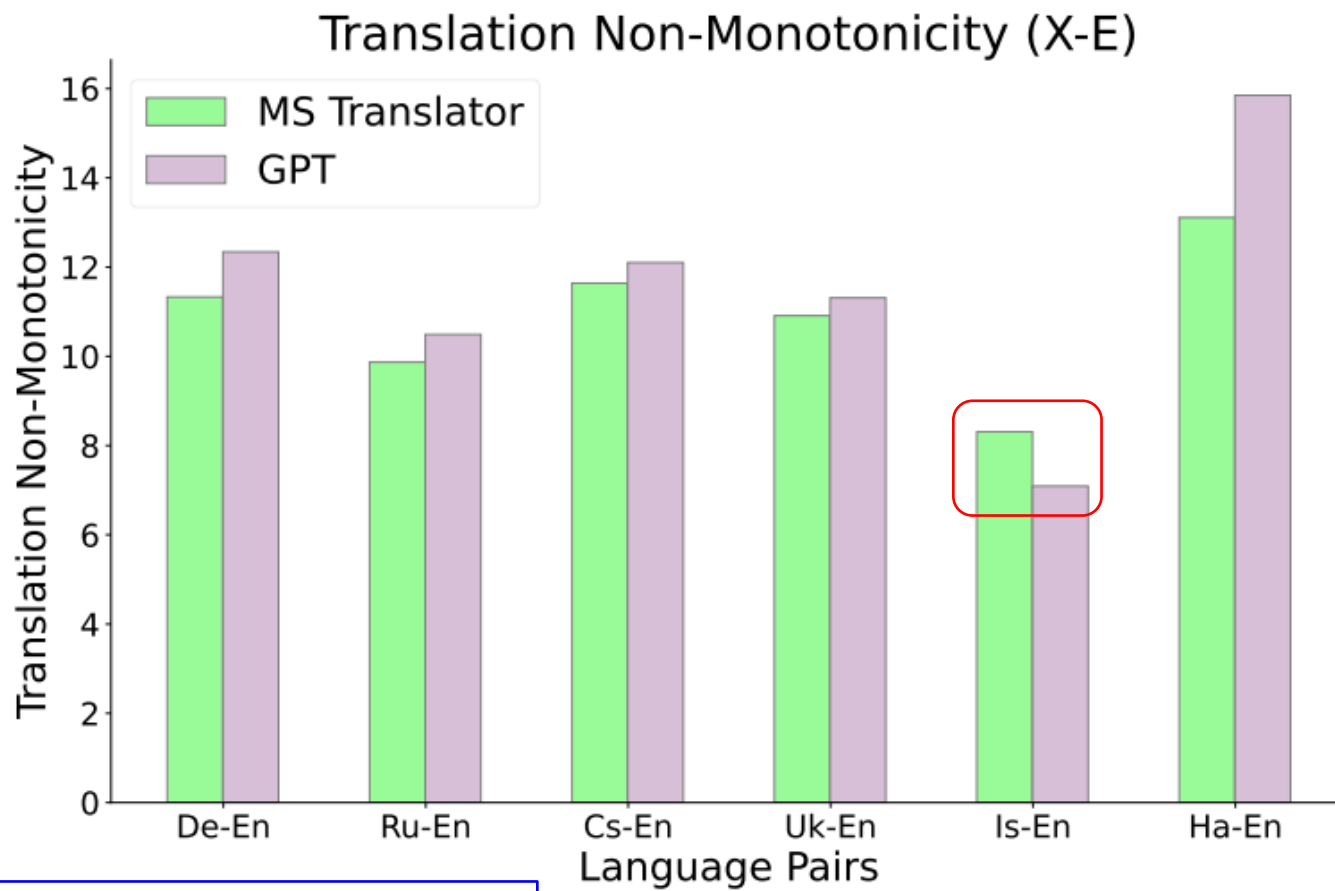
Sequence Type	Translation Instance	Phenomenon
Source MS Translator GPT	Bis auf die E 95 02 wurden <u>alle Lokomotiven zerlegt</u> . With the exception of E 95 02, all locomotives were dismantled . <u>All locomotives were dismantled</u> except for the E 95 02.	Non-Monotonicity (NM)
Source MS Translator GPT	Oder ist sie ganz aus dem Sortiment genommen? Or is it completely removed from the range? <u>Or has it been</u> completely removed from the range?	Fluency (F)
Source MS Translator GPT	Sehen Sie bitte im Screenshot was der Kollege geschrieben hat Please see in the screenshot what the colleague wrote Please see the screenshot for what the colleague wrote.	Punctuation Insertion (PI)
Source MS Translator GPT	Die Email zur Stornierung wurde am 26.12. <u>#NUMBER#</u> versendet. The cancellation email was sent on 26.12. <u>#NUMBER#</u> . The cancellation email was sent on December 26th.	Dropped Content (USW)
Source MS Translator GPT	"We won't accept the <u>CAA</u> and that is for sure. “我们不会接受CAA，这是肯定的。 “我们不会接受 <u>《公民法》</u> ，这是肯定的。	Inserted Content (UTW)

Amr Hendy et al., Good Are GPT Models at Machine Translation? A Comprehensive Evaluation, <https://arxiv.org/abs/2302.09210>



2. ChatGPT 能力分析

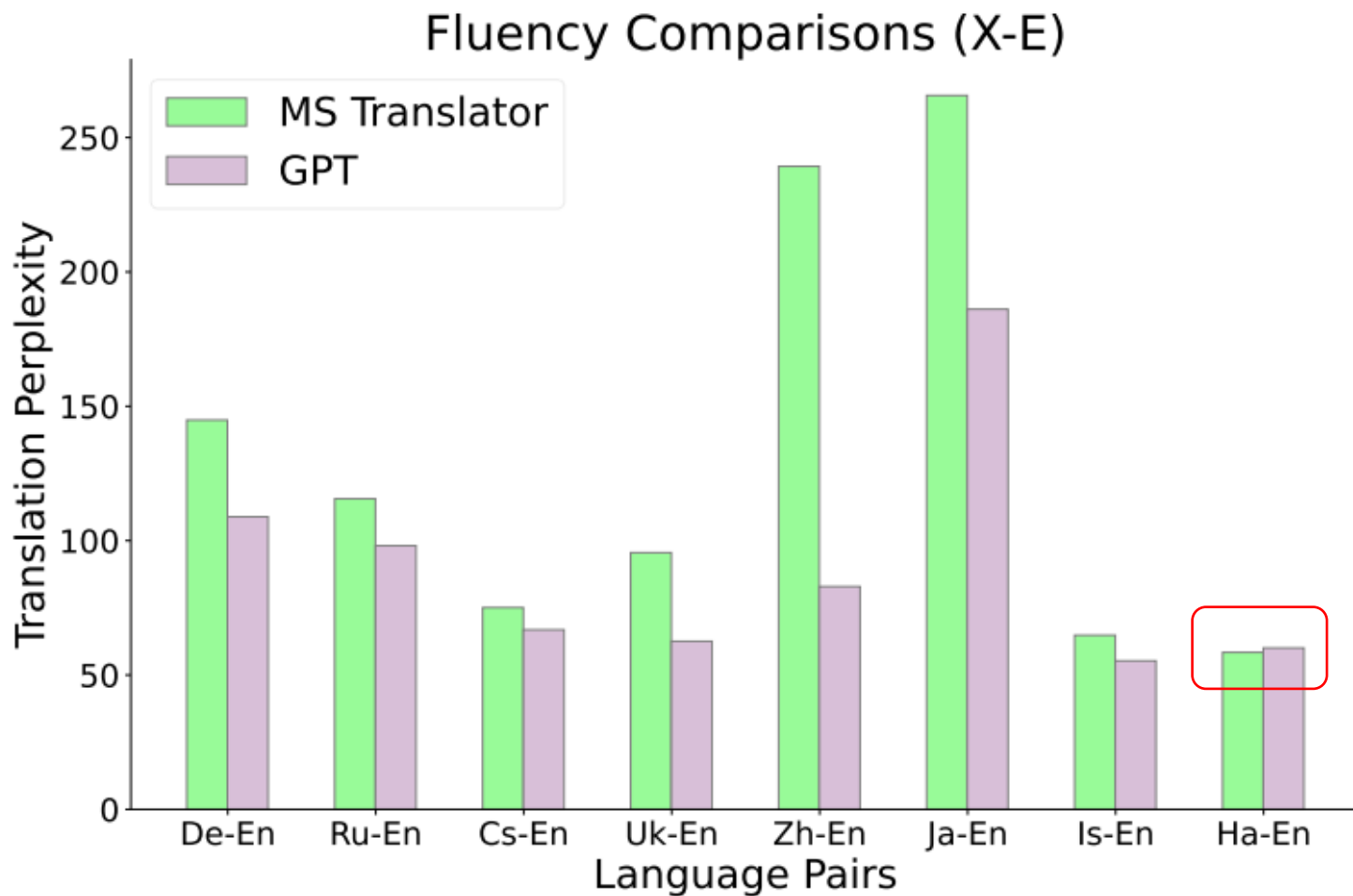
➤ 非单调性对比(non-monotonicity)



De- 德语 Ru- 俄语 CS- 捷克语
Uk- 乌克兰语 Is- 冰岛语 Ha- 豪萨语

2. ChatGPT 能力分析

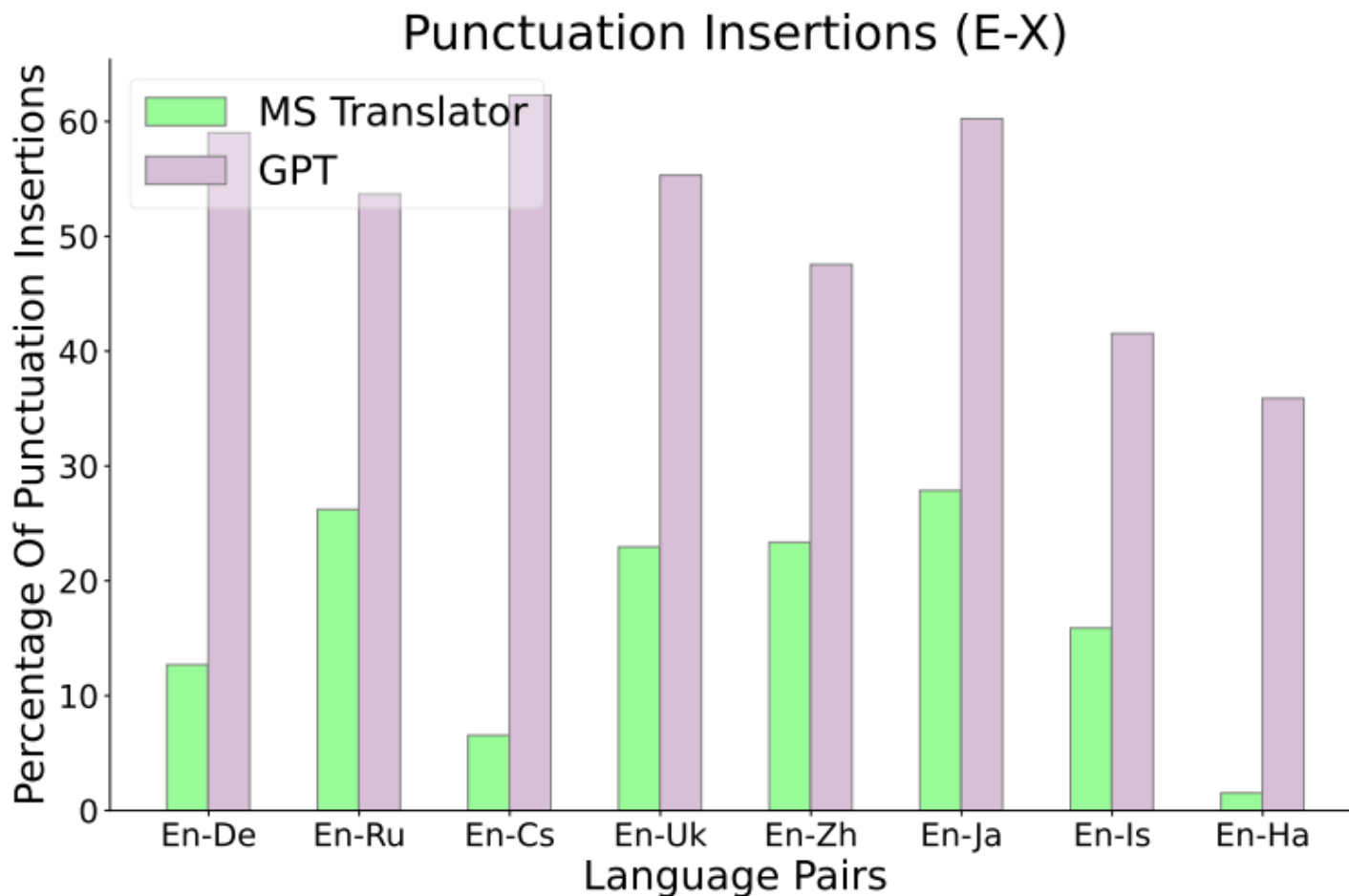
➤ 流畅性对比(fluency)





2. ChatGPT 能力分析

➤ 标点插入能力对比(punctuation Insertions)





2. ChatGPT 能力分析

● Is GPT a good foundation model?

➤ 模型特点(Bias)

- 以牺牲充分性(adequacy)为代价更喜欢选择流畅性好的译文
- 有更强的标点符号插入能力
- 对于缩略语有扩展解释能力
- 能够引入世界知识
- 常产生幻觉译文

➤ 约束大模型的行为以适应具体任务的需求并非易事

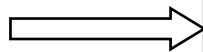
- 提示学习(prompt learning)/ 上下文学习(in-context learning) 好用，但并不完美
- 指令调试(instruction tuning) / RLHF 可能会引入新的偏见



2. ChatGPT 能力分析

对比测试

- 机器翻译
- **信息抽取**
- 文本摘要
- 情感分析



在命名实体识别、关系抽取和事件抽取三项任务上, ChatGPT的性能最高只能达到专门训练出来的最优模型性能的63.5%、43.0%和35.3% [Han et al., 2023]。

CoNLL04: 人工评价, 仍不如现有最好的系统。



2. ChatGPT 能力分析

对比测试

● 机器翻译

● 信息抽取

● 文本摘要 $\longrightarrow$ 在CSDS中文对话数据集上测试:

● 情感分析



SemEval-2014
情感极性三元组

Model	ROUGE-2	ROUGE-L	BERTScore
ChatGPT	17.5	39.1	70.4
Fast-RL	41.4	47.1	79.8

约为该数据集上最优模型BDTS的 **58.6%**[Zhang et al., 2022]。

提 纲

1. 生成式语言模型
2. ChatGPT 能力分析
- ➡ 3. 未来展望

3. 未来展望

◆ 大语言模型会成为NLP研究的终结者吗？

- 不会！
- ChatGPT 的“成功”只是证明了生成式语言模型可能是实现“通用AI”的一条可行之路，但是，它还远没有成功！
- 生成式语言模型未必是解决 NLP 问题的唯一之路。



3. 未来展望

◆ 面对大语言模型，NLP研究如何开展？

- 企业有企业的做法，学术界应该有学术界的使命和担当；
- 以 ChatGPT 为代表的大语言模型对学术界带来的冲击是不可逆转的；
- 如果学术界和企业界做同样的事情，或者学术界只是为企业培养调参工人，甚至沦为企业的打工仔，那么，学术界就没有存在的必要；
- 对于学术界，除了培养人才以外，找准切入点，做能做的事，做有用的事，做创新的事，才是立足的根本。

3. 未来展望

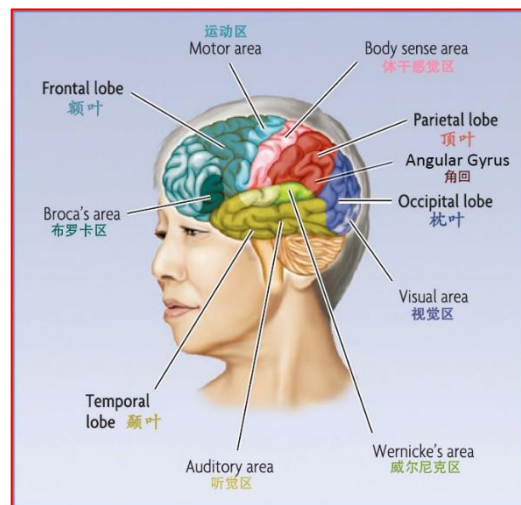
◆ 未来的NLP研究



3. 未来展望

◆ 大模型之外

- 面向垂直领域和特定任务的创新方法探索
话说NLP大事：符号逻辑转统计学习；统计学习转神经网络；神经网络 + 符号逻辑？
- 基于小数据甚至零样本、少量人工的小模型
- 基于人机交互的强化学习和模型自主进化
- 可解释性、可溯源的推理模型
- 特定语言相关的应用技术研发
- 评测方法/标准与数据集
- 受脑认知过程启发的NLP模型
-



3. 未来展望

◆ 回首假设

- 语言的“稳态性” $\implies$ 语言是动态发展和变化的
- 语言的“可遍历性” $\implies$ 每个个体的语言风格各不相同

◆ 如何建立类脑的更加便捷、灵活的语言模型

◆ 如何模拟人学习语言的知识获取过程

◆ 如何构建具有丰富联想和复杂推理能力的语言模型

◆ 如何实现针对不同文化背景、不同用户和不同任务的个性化语言处理

NLP 仍将是AI皇冠上那颗闪亮的明珠！



谢谢!
Thanks!



参考文献

- [Bengio et al., 2003] Y. Bengio et al. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(3): 1137-1155
- [Graves, 2013] Graves A. 2013. Generating sequences with recurrent neural networks. Arxiv Preprint Arxiv: 1308. 0850
- [Han et al., 2023] R. Han et al., Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors. (2023-05-23)/[2023-06-21]. <https://arxiv.org/abs/2305.14450>
- [Hendy et al., 2023] A. Hendy et al. 2023. Good Are GPT Models at Machine Translation? A Comprehensive Evaluation, <https://arxiv.org/abs/2302.09210>
- [Mikolov et al., 2010] T. Mikolov. 2010. Recurrent neural network based language model. *Proc. Interspeech*, pp. 1045-1048
- [Mikolov et al., 2013a] T. Mikolov. 2013. Efficient estimation of word representations in vector space. *Proc. ICLR Workshop Track*
- [Peters et al., 2018] Matthew E. Peters et al. 2018. Deep Contextualized Word Representations. *Proc. NAACL'2018*, pages 2227–2237
- [Radford et al., 2018] A. Radford et al. 2018. Improving language understanding by generative pre-training[Z/OL]. [https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper.pdf](https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language%20understanding%20paper.pdf)
- [Zhang et al., 2022] Y. Zhang et al. 2022. Boundary-Driven Table-Filling for Aspect Sentiment Triplet Extraction. *Proc. EMNLP'2022*, pp. 6485–6498