计算语言模型中词汇多义性的表示和评估

汇报人：刘柱

专业：中国语言文学（计算语言学）

导师：刘颖教授

汇报时间：2024.07.01

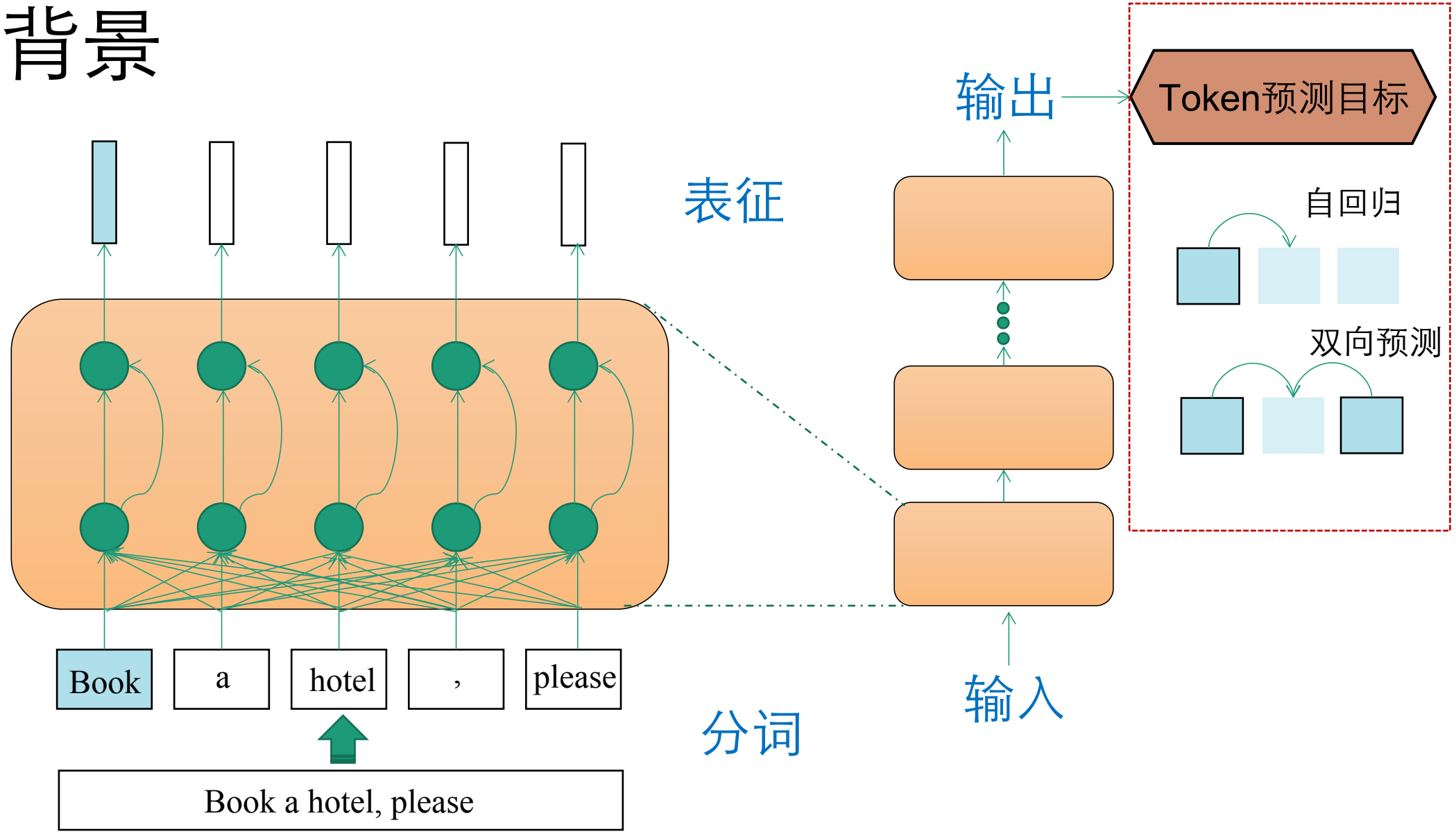
大纲

- 选题背景
- 文献综述
- 研究内容
- 已完成的工作
- 课题总结

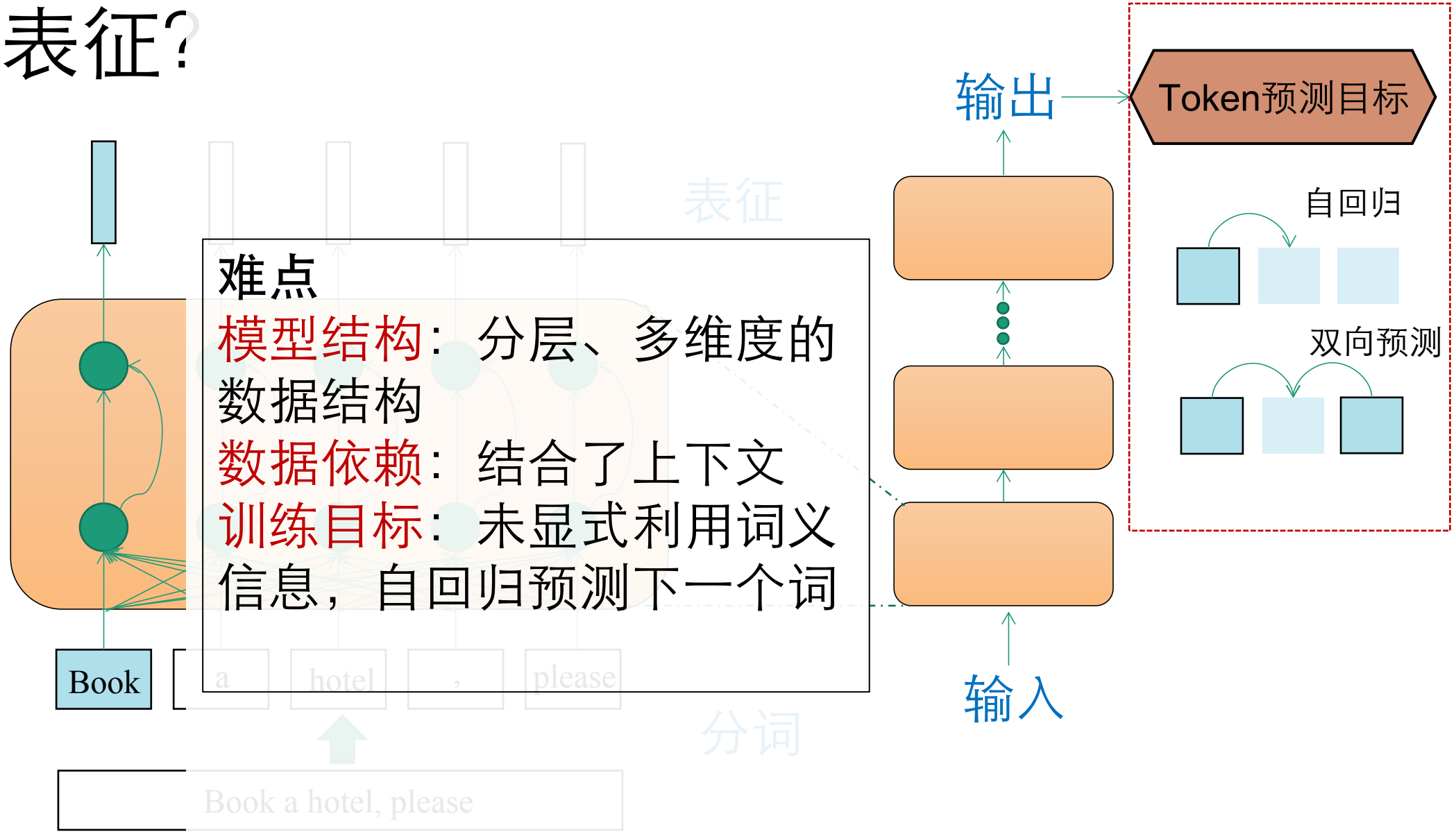
选题背景

- 语言模型具有卓越和通用的语言理解能力
  - 语义理解任务：词汇级别➡短语级别➡句子级别➡篇章级别
  - 语言模型：分层的Transformer模型；分布式假设
  - 性能卓越且通用：Foundation models
- 模型使用实值的连续向量表征来表示每个运算单元（token）
  - token：句子分词后的子词部分，运算的最小单元，粗略认为是词
  - representation（表征）：模型的中间输出，高维、连续、上下文性
- 该表征多大程度上反映token/词的上下文语义?
  - 词汇多义性的表示和评估
  - 自上而下地提供模型的可解释性

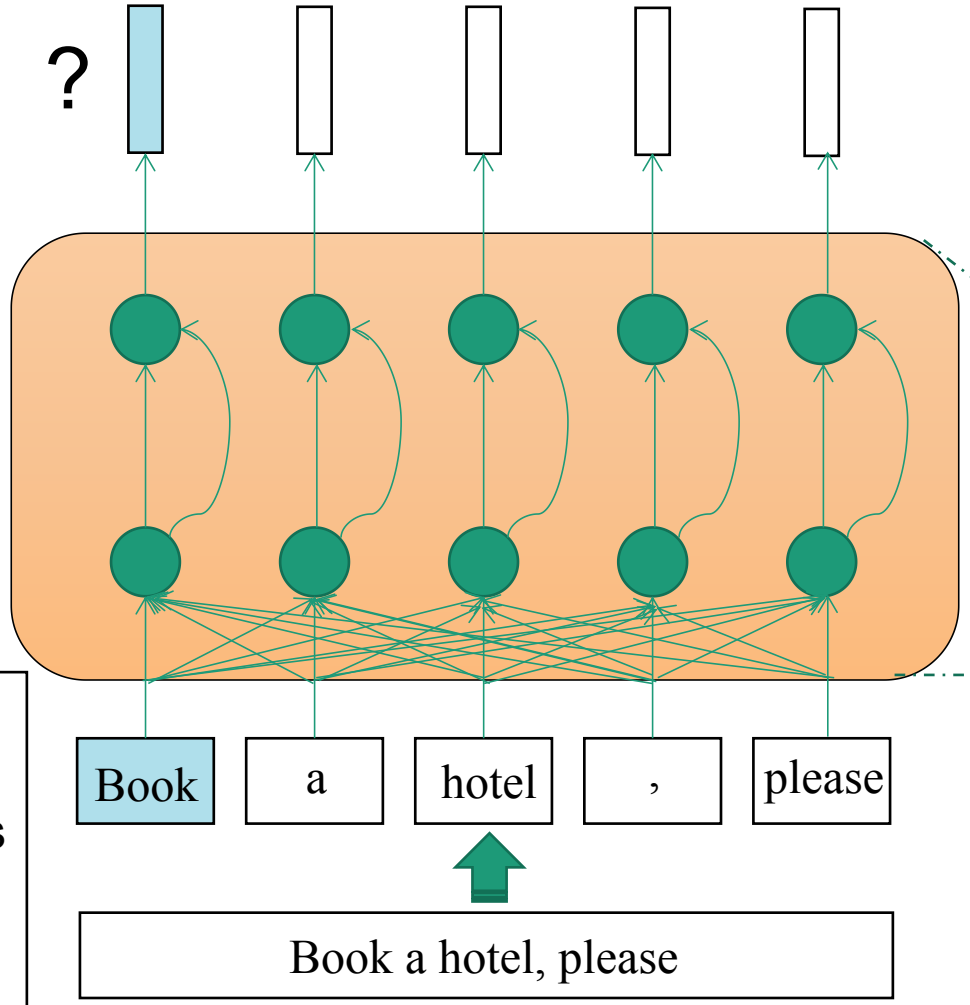
选题背景


输出
Token预测目标
表征
自回归
双向预测
Book
a
hotel
,
please
分词
输入
Book a hotel, please

如何表征?


输出
Token预测目标
自回归
双向预测
表征
难点
模型结构：分层、多维度的数据结构
数据依赖：结合了上下文
训练目标：未显式利用词义信息，自回归预测下一个词
Book
a
hotel
,
please
输入
分词
Book a hotel, please

表征如何?



Read a book
the book of Job
a book of stamps
open a book
...

模型的表征是否可以：

1）区别不同上下文的意思

2）消除歧义

3）表征词义的不确定性

……

文献综述

- 词汇多义性
- 语言模型中的知识评估
- 语言模型中的词义消歧

词汇多义性

- “一词多义”现象：词 vs. 义
- 同形异义词（homonym）和多义项词（polysemy）
  - 定义 [12,17,18]
  - 二者的区别和联系 [19,20,25,26]：词源、关联度、相似性
  - 区分特征 [21-24]：词类
- （原型）语义角色
  - 提出 [27-28]：语义角色，原型施事和原型受事
  - 特征 [13, 29-31]：连续统，程度在变化
  - 跨语言共性 [31-34,36]：格配置变动（汉语靠语序表达语义角色；其他语言格标记）

词汇多义性

表 1 不同多义类型的区分

类型	语言平面	义项关联程度	是否需独立义项	研究单元
同形异义词	词	不相关	需要	词
多义项词	义项	相关	需要	词
语义角色	用例	紧密	不需要	主谓宾结构中的词

名称	符号	词类	上下文	举例
同形异义词	$\mathcal{W}_{\mathcal{H}}$	任意	任意	“花”的植物义和花钱义
多义项词	$\mathcal{W}_{\mathcal{P}}$	实词	任意	“头”的身体义和首领义
语义角色	$\mathcal{W}_{\mathcal{V}'}$	任意	可互逆的主谓宾	“十个人”的施事性变化[27]

表 7 不同词汇多义性类别的对比

一顿饭吃十个人
十个人吃一顿饭

语言模型中的知识评估

- 评估内容
  - 语言学知识[45-59]：词法、句法、语义
  - 认知知识：推理和泛化的能力[60-64]；推理任务[65,66]的细分
  - 伦理知识[67-77]: 价值观；对齐方式
- 评估方法
  - 行为主义方式：分析输入输出的关系
  - 基于表征自上而下：无监督几何空间变换[84-87]和有监督探针[88-90]，BERTology
  - 由里及外的机械可解释性[91-94]：分析神经元、神经回路等
- 不确定性评估
  - 来源[96-98]：数据（固有的噪声）；模型（数据分布导致有偏的模型）
  - 评估和校准：MC Dropout; Model Ensemble等 [99-105]

语言模型中的词义消歧

- 语料资源
  - 语料库：SemCor[106]; OMSTI[109]; WNGC; OntoNotes; SemEval
  - 词义字典/知识库：WordNet；知网；BabelNet [114]
- 模型方法
  - 完全知识驱动：相似性匹配[132-133]；图算法[135-138]；融合语言学知识[142]
  - 监督式数据驱动算法: 分类[148-152]；语义检索[121-124]；截取式[125-127]；生成式[153-161]
  - 无/半监督式数据驱动算法: [84,162,85,130-131,163-164]
- 性能比较
  - 基准：上界(人类标注标准)、下界(随机)、强下界（MFS）

语言模型中的词义消歧

方法	SE02	SE03	SE07	SE13	SE15	ALL	名	动	形	副
ITA[118]	-	-	-	-	-	80.0				
LB_Mono[23]	-	-	-	-	-	17.4	13.5	4.5	23.7	16.3
MFS_Cop	65.6	66.0	54.5	63.8	67.1	65.5	-	-	-	-
MFS_WN1	66.8	66.2	55.2	63.0	67.8	65.2	-	-	-	-
ChatGPT	-	-	-	-	-	73.3	-	-	-	-
GAS[115]	72.2	70.5	-	67.2	72.6	70.6	72.2	57.7	76.6	85.0
GlossBERT[116]	77.7	75.2	72.5	76.1	80.4	77.0	79.8	67.1	79.6	87.4
EWISE[117]	73.8	71.1	67.3	69.4	74.5	71.8	74.0	60.2	78.0	82.1
EWISER[118]	80.8	79.0	75.2	80.7	81.8	80.1	82.9	69.4	83.6	87.3
MLWSD[119]	78.4	77.8	72.2	76.7	78.2	77.6	80.1	67.0	80.5	86.2
MLWSD*	80.4	77.8	76.2	81.8	83.3	80.2	82.9	70.3	83.4	85.5
RTWE[120]	83.4	82.9	74.5	82.1	85.3	82.7	84.9	72.8	87.7	87.9
RTWE*	85.2	83.3	77.1	83.8	86.3	84.1	85.7	75.1	90.6	88.7
BEM[121]	79.4	77.4	74.5	79.7	81.7	79.0	81.4	68.5	83.0	87.9
Z-reweight[122]	79.6	76.5	71.9	78.9	82.5	78.6	-	-	-	-
SACE[123]	82.4	81.1	76.3	82.5	83.7	81.9	84.1	72.2	86.4	89.0
SACE*	83.6	81.4	77.8	82.4	87.3	82.9	85.3	74.2	85.9	87.3
ARES[124]	78.0	77.1	71.0	77.3	83.2	77.9	80.6	68.3	80.5	83.5
ESCHER[125]	81.7	77.8	76.3	82.2	83.2	80.7	83.9	69.3	83.8	86.7
ConSec[126]	82.3	79.9	77.4	83.2	85.2	82.0	85.4	70.8	84.0	87.3
ConSec*	82.7	81.0	78.5	85.2	87.5	83.2	86.4	72.4	85.4	89.0
KELESC[127]	82.2	78.1	76.7	82.2	83.0	81.2	84.3	69.4	84.0	86.7
Generationary[129]	77.8	73.7	68.8	78.3	77.6	76.3	79.8	63.3	80.1	84.7
Lesk_ext[133]	58.4	59.4	-	-	-	-	-	-	-	-
SREF[134]	72.7	71.5	61.5	76.4	79.5	73.5	78.5	56.6	79.0	76.9
UKB[135]	59.7	57.9	41.7	-	-	-	-	-	-	-
Babelfy[136]	-	68.3	62.7	65.9	-	-	-	-	-	-
SyntagRank[137]	71.6	72.0	59.3	72.2	75.8	71.7	64.1	-	-	-
WSDG[138]	68.7	68.3	58.9	66.4	70.7	67.7	71.1	51.9	75.4	80.9

以往工作存在的缺陷

- 任务宽泛，缺乏细粒度分析
  - 例如不同词类、不同语义类、不同选择特征等都会影响词义的表达
  - 缺乏可控原则：例如同一个目标词的上下文差异太大
- 以英语为主，缺乏针对汉语的任务
  - 汉语独有的特征：词法结构、普遍的兼类用法、语素的语义稳定性等
- 仅着眼于准确率，没有考虑词义选择的连续性和不确定性
  - 词义系统是一个相互关联的网状系统
  - 不确定性来源：上下文的不充分性、模型学习的偏差
- 缺少对大语言模型中词义表示的研究
  - BERTology针对的是基于MLM的BERT模型，生成式大语言模型词义表征更难分析

研究内容

- 如何表征（提取）
  - 提取对象：如何从语言模型中提取（最佳的）表征?
- 表征如何（评估）
  - 评估任务：如何设计词汇多义性相关的评估任务?
  - 评估指标：准确性、不确定性等

研究内容

- 如何表征（提取）
  - 提取对象：如何从语言模型中提取（最佳的）表征?
- 表征如何（评估）
  - 评估任务：如何设计词汇多义性相关的评估任务?
  - 评估指标：准确性、不确定性等

方案：
- 从不同的角度对模型进行探测，需要考虑到模型的训练目标、层级、数据格式等
- 对不同类型的模型进行比较，尤其是生成式模型和双向掩码模型
- 考虑目标词与上下文之间的关联（分布式假设）

研究内容

- 如何表征（提取）
  - 提取对象：如何从语言模型中提取（最佳的）表征?
- 表征如何（评估）
  - 评估任务：如何设计词汇多义性相关的评估任务?
  - 评估指标：准确性、不确定性等

名称	符号	词类	上下文	举例
同形异义词	$\mathcal{W}_{\mathcal{H}}$	任意	任意	“花”的植物义和花钱义
多义项词	$\mathcal{W}_{\mathcal{P}}$	实词	任意	“头”的身体义和首领义
语义角色	$\mathcal{W}_{\mathcal{V}'}$	任意	可互逆的主谓宾	“十个人”的施事性变化[27]

表 7 不同词汇多义性类别的对比

研究内容

- 如何表征（提取）
  - 提取对象：如何从语言模型中提取（最佳的）表征?
- 表征如何（评估）
  - 评估任务：如何设计词汇多义性相关的评估任务?
  - 评估指标：准确性、不确定性等

方案：
- 考虑多样的评估体系，例如考虑词义选择的不确定性。
- 结合不同语言的特性

已完成工作

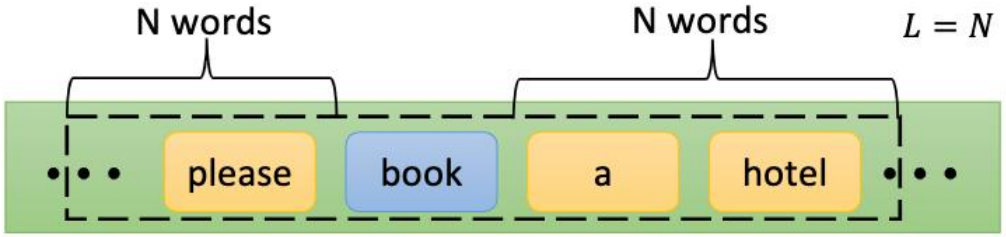
- 大语言模型表征对词义的反映
- 词义消歧中的不确定性估计
- 汉语主谓宾句主宾互易数据集构建和评估

大语言模型表征对词义的反映(C5)

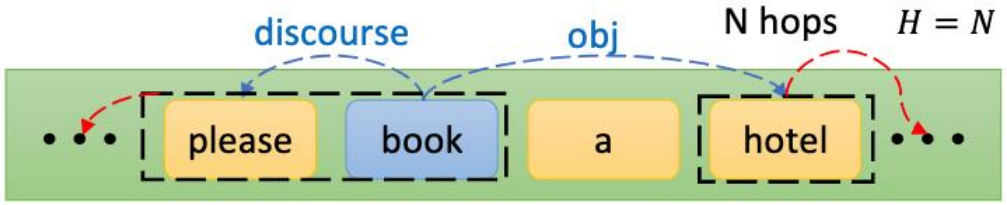
- 对大语言模型的表征如何反映词义进行了探测。
- 探测任务：WiC数据集（二分类版的词义消歧任务）
- 探测模型：Llama2；生成式模型
- 难点：自回归的模型无法看到后文；训练目标是预测下一个词
- 不同的探测手段比较
  - 句子；重复句子；prompt引导
  - 层数变化
- 结论：模型在低层进行词义预测；*高层更多反映后一个词的词义*
- 进度：当前工作被ACL'24 Findings录取，后续还在跟进

词义消歧中的不确定性估计(C4)

- 词义消歧在选择“正确”词义时候，往往存在不确定性
  - 义项之间并非完全独立，~20%的不一致率
  - 不确定性来源：数据固有的噪声（不充分的上下文等）；模型学习不足
  - 现有工作仅关注准确性，很少评估不确定性
- 评估手段
  - 数据不确定性（设计两种残缺的上下文）
  - 模型不确定性（OOD数据集）
- 结论
  - 模型处理数据不确定性更好
  - 分析了影响不确定性的词汇因素
- 进度：被ACL’23 Findings收录；计划考虑汉语特点



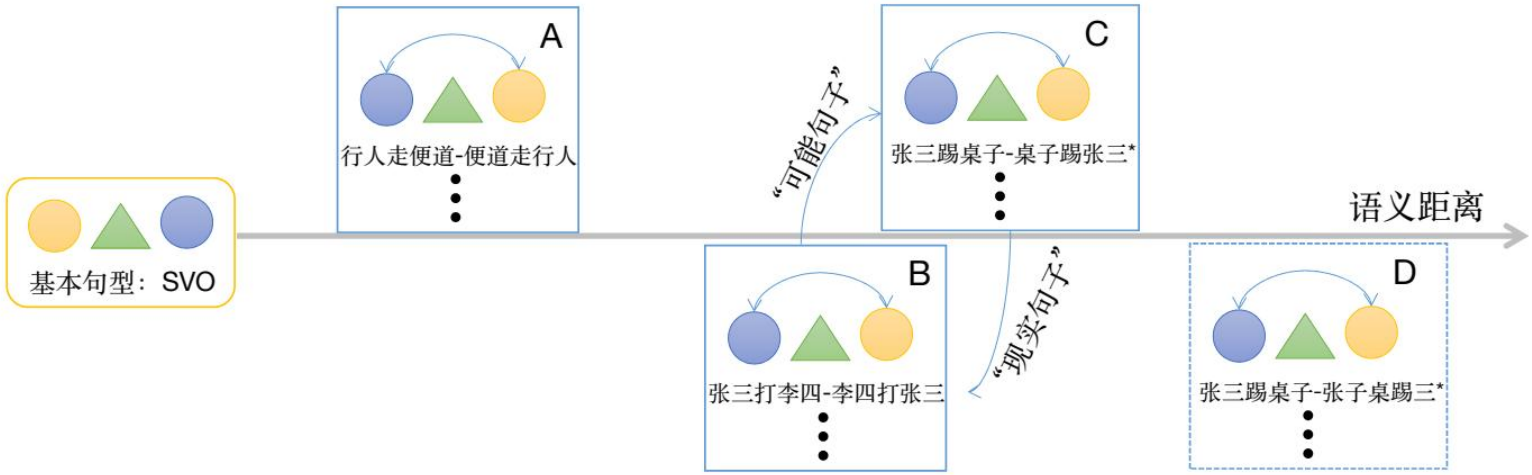
(a) window-controlled context



(b) syntax-controlled context

汉语主谓宾句主宾互易数据集构建和评估

- 考察目标
  - 模型是否可以分辨出互逆的“最小”变化语境中语义角色的变化
  - “十个人吃一顿饭”和“一顿饭吃十个人”中“十个人”施事性不变
- 数据集构建
  - 演变：SVO、OVS（可逆; 不可逆; “可能句子”; “不可能句子”）



汉语主谓宾句主宾互易数据集构建和评估

- 考察目标
  - 模型是否可以分辨出互逆的“最小”变化语境中(原型)语义角色的变化
  - “十个人吃一顿饭”和“一顿饭吃十个人”中“十个人”施事性不变
- 数据集构建
  - 演变：SVO、OVS（可逆; 不可逆; “可能句子”; “不可能句子”）
  - （难点）可逆句的构建。收集基础可逆句类型+数据扩充
  - 其他：从动词角度出发，选择不同语义特征的及物动词
- 进度
  - 数据集已经构建完。
  - 之后的工作要对模型进行分析
  - 互易数据集再完善

统计量	$\mathcal{I}$	$\mathcal{E}$	$\mathcal{U}$
句子数量	7817	13237	18056
主语数量	213	30	86
宾语数量	185	34	299
动词数量	93	126	352

课题总结

- 目标：计算语言模型中词汇多义性的表示和评估
  - 表示：如何表示；哪里表示
  - 评估
    - 评估内容：词汇在上下文中的多义性
    - 评估对象：计算语言模型，以大型语言模型为主
    - 评估指标：不确定性
- 预期创新点
  - 对语言模型中词的表征提供语言学上的可解释性以及工程上的帮助
  - 设计多样化的标准来评估语言模型
  - 设计更加复杂、适用于多语言尤其是汉语的任务

进度安排

研究目标	时间				
	2023	2024 上	2024 下	2025 上	2025 下
文献调研、整理和归纳	✓	✓	✓	✓	
实验设计一：多义性中不确定性的建模	✓				
结果分析一	✓				
写作与发表	✓				
实验设计二：模型中多义性的反映		✓			
结果分析二		✓			
写作与发表		✓			
实验设计三：主宾互易数据集收集及评测			✓	✓	
结果分析三			✓	✓	
写作与发表			✓	✓	
论文修改和写作				✓	✓

参考文献

- 文献参考开题报告中的序号

Q & A

请老师们批评指正

谢谢

Note

表 15 主宾互易后语义不变的语义类型举例说明

语义类型	举例		
混合义	两份水泥配一份沙子	⟺	一份沙子配两份水泥
依附义	名字签空格里	⟺	空格里签名字
供给义	一间屋子住五个人	⟺	五个人住一间屋子
笼罩义	大楼笼罩着晨雾	⟺	晨雾笼罩着大楼
充满义	天空布满了乌云	⟺	乌云布满了天空
进入义	暗房透进一线光	⟺	一线光透进暗房

Note

表 13 输入设定以及输入实例，加粗位置表示目标词的表征提取的位置。

设定	输入示例
base	the **bank** of the river
repeat	the bank of the river the **bank** of the river
repeat_prev	the bank of the river **the** bank of the river
prompt	In this sentence “the bank of the river”, “bank” means in one word **:**

Note

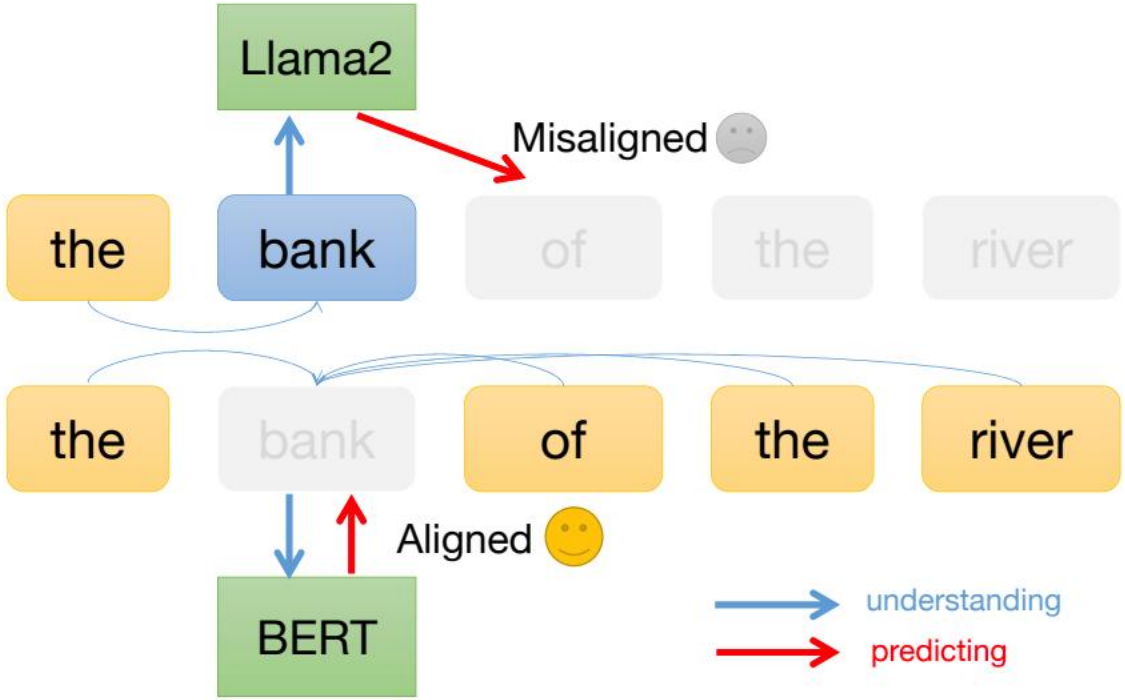


图 7 BERT 和 Llama2 语言模型的关键差异。蓝色和红色线条表示理解和预测的信息流。这里的“理解”指利用上下文捕捉词汇语义。从上下文到当前词的蓝色线条表示理解的流向。

Note

表 2 原型施事和原型受事的典型特征

原型施事	原型受事
意愿性 (violation)	经历状态变化 (change of state)
感知性 (sentience)	递增性 (incremental)
使动性 (causation)	受动性 (causally affected)
移位性 (movement)	静态性 (stationary)
独立性 (independent existence)	依存性 (existence not independent of event)