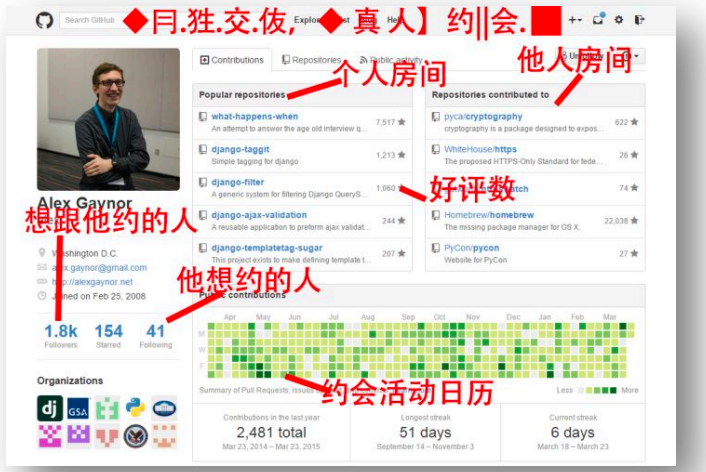
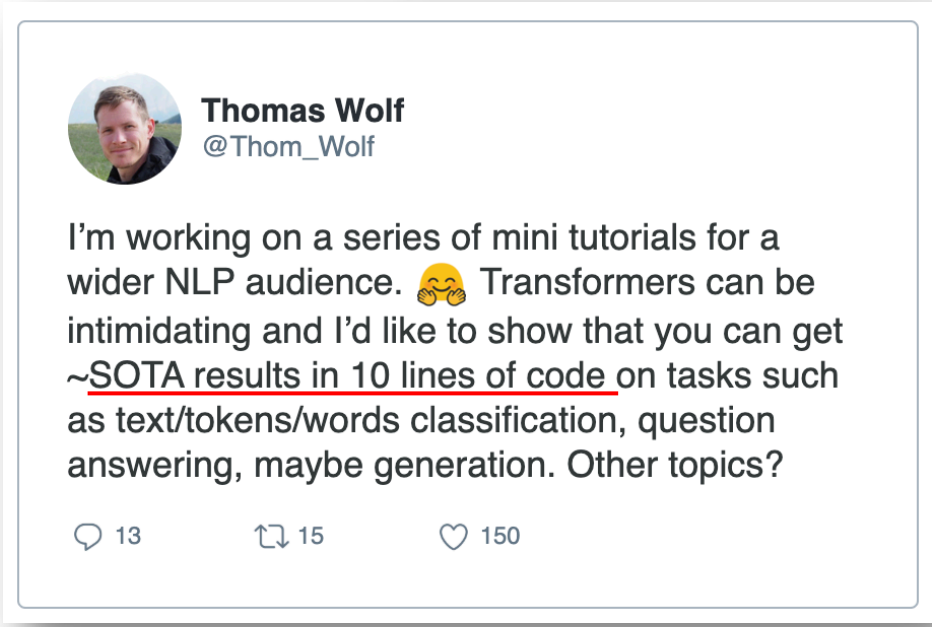
# NLPer的核心竞争力是什么？

车万翔
哈尔滨工业大学
2019-10

# NLPer的危机

- 核心工具多来自机器学习社区
  - CNN、RNN、Seq2seq、NMT、Transformer
  - 深度学习之前，我们至少还能做特征工程
- NLP的门槛逐步降低
  - 开源工具、BERT等预训练模型
  - 算力成为了主要瓶颈

# NLPer的核心竞争力

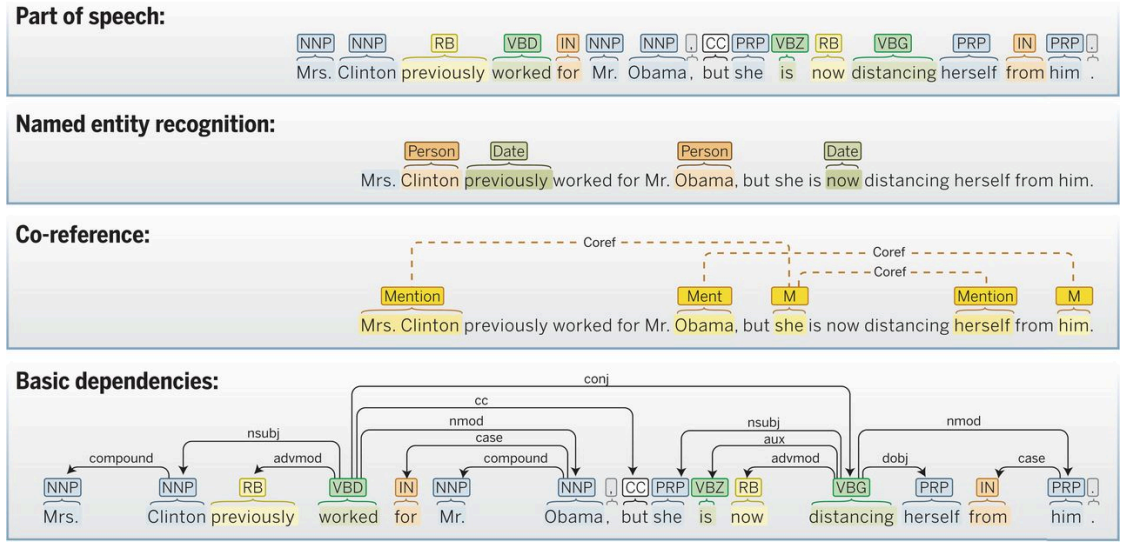
- 一个核心
  - 结构化是NLP的核心问题
- 两个能力
  - 发现问题的能力
  - 解决问题的能力
- 三个优势
  - 对基本概念理解更准确
  - 对研究有更好的品味
  - 对数据更敏感

# NLPer的核心竞争力

- 一个核心
  - 结构化是NLP的核心问题
- 两个能力
  - 发现问题的能力
  - 解决问题的能力
- 三个优势
  - 对基本概念理解更准确
  - 对研究有更好的品味
  - 对数据更敏感

# 自然语言处理的本质

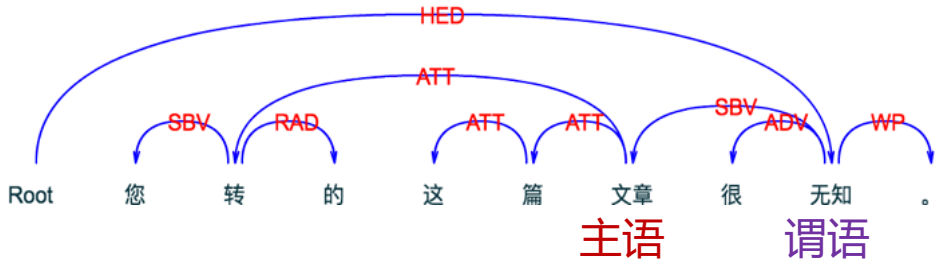
- 从无结构序列中预测有结构语义
  - 包括句法分析、命名实体识别、词性标注等任务

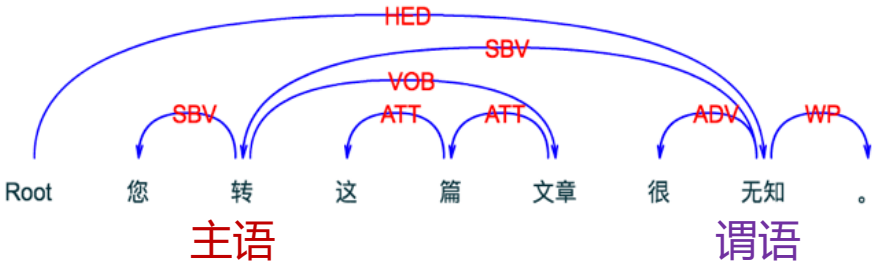



Julia Hirschberg and Christopher D. Manning. **Science** 2015

Science AAAS

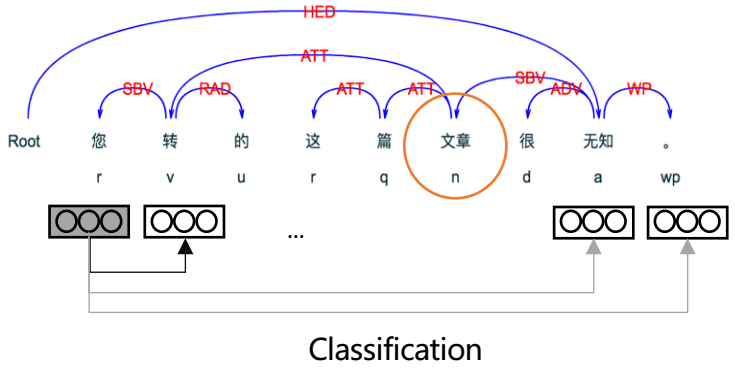
- 如句法分析
  - 您转**的**这篇文章很无知
  - 您转这篇文章很无知



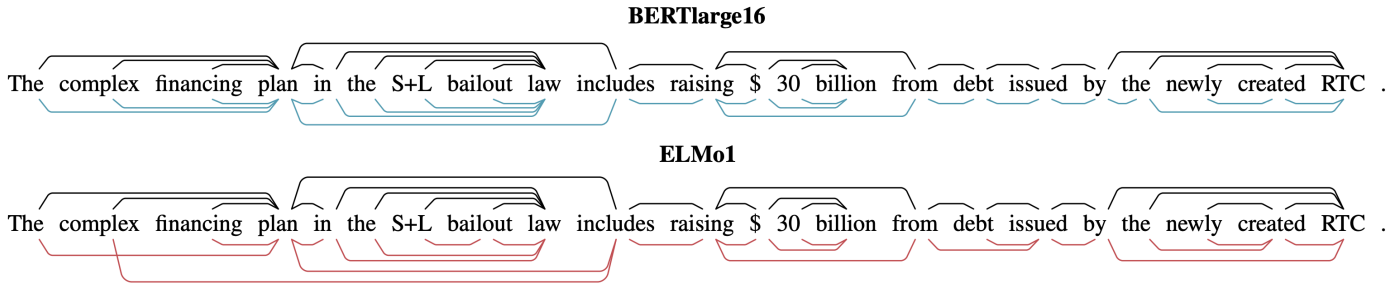



# 深度模型能力越来越强，结构是否重要？

- Encoder足够强，结构约束变的不再重要
  - Timothy Dozat and Christopher D. Manning. Deep Biaffine Attention for Neural Dependency Parsing. ICLR 2017.

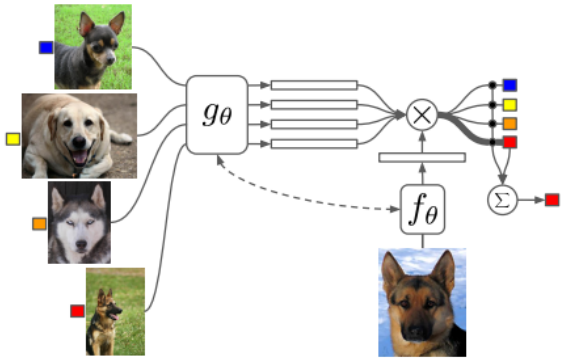


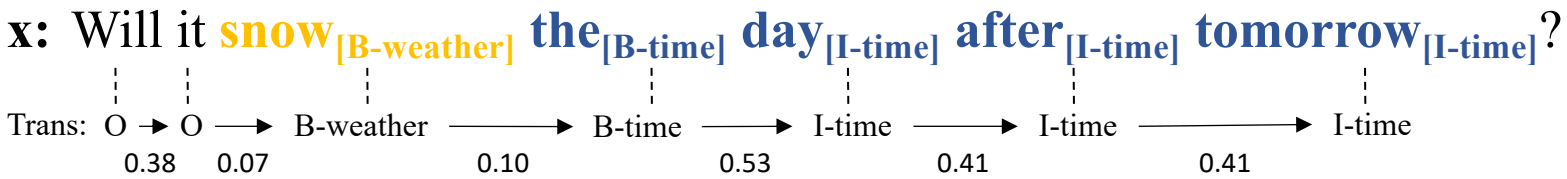

- 预训练模型蕴含句法结构信息
  - John Hewitt and Christopher D. Manning. A Structural Probe for Finding Syntax in Word Representations. NAACL 2019.



# 小样本下结构信息依然重要

- 小样本学习目前多应用于分类任务
- 如何将小样本学习应用于序列标注？
  - 标签之间互相影响，新的领域有新的标签集



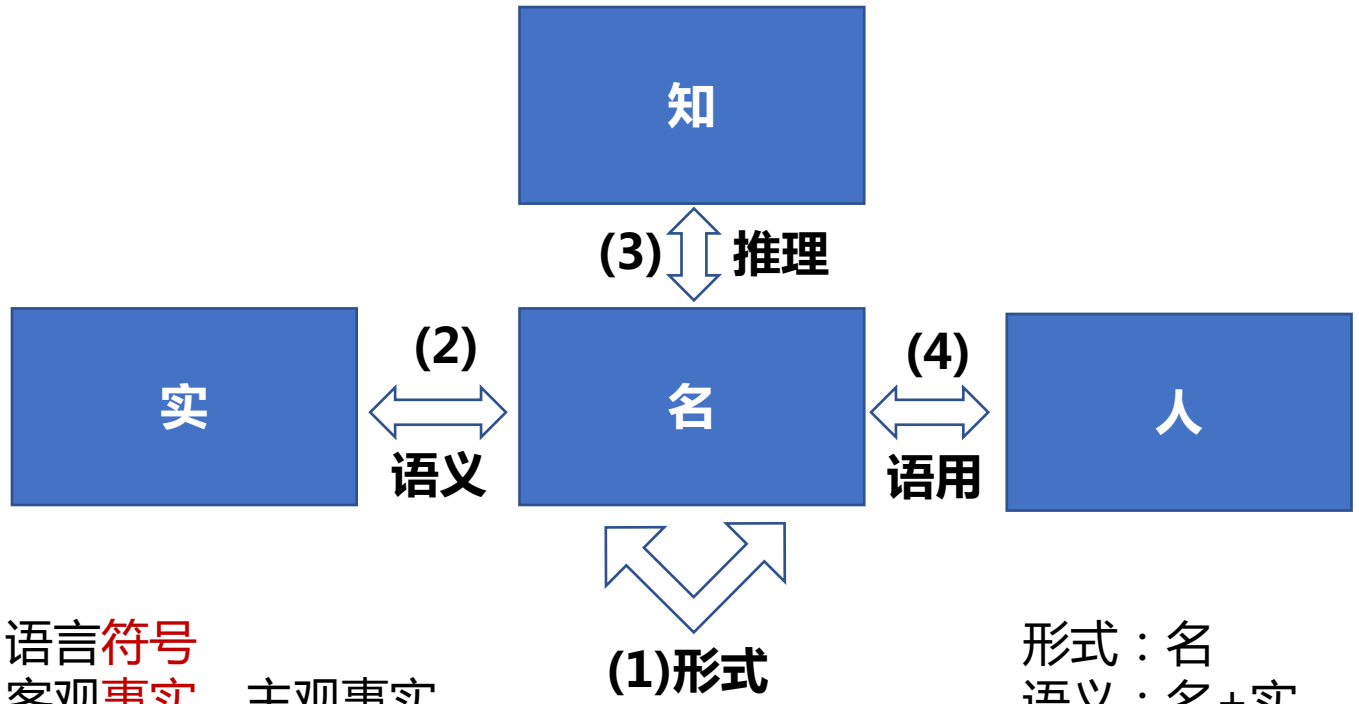

- 我们利用CRF模型来建模
  - 转移概率：提出一种回退机制，建模未见标签的转移概率
  - 发射概率：利用Pair-wise Embedding更好计算词相似度

Our Paper：https://arxiv.org/pdf/1906.08711.pdf

# NLPer的核心竞争力

- 一个核心
  - 结构化是NLP的核心问题
- 两个能力
  - 发现问题的能力
  - 解决问题的能力
- 三个优势
  - 对基本概念理解更准确
  - 对研究有更好的品味
  - 对数据更敏感

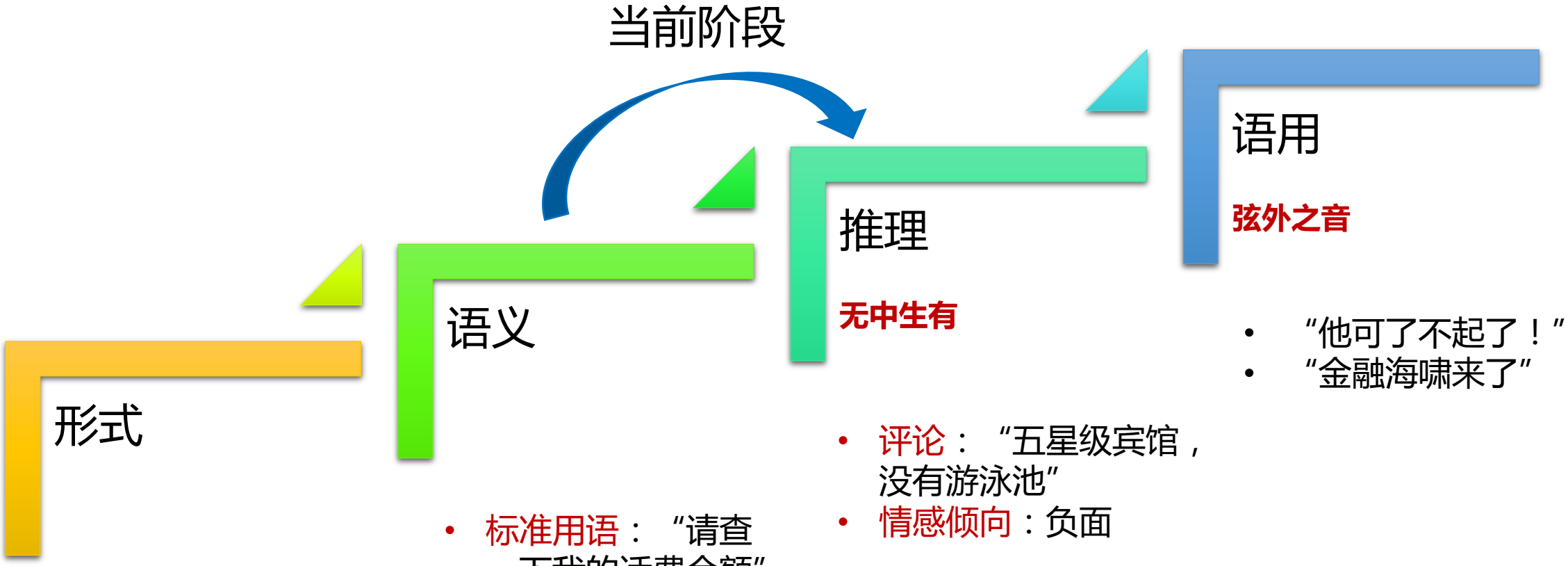
# 语言理解的四个空间




名：语言符号
实：客观事实、主观事实
知：知识
人：语言的使用者

形式：名
语义：名+实
推理：名+实+知
语用：名+实+知+人

# NLP由浅入深的四个层面



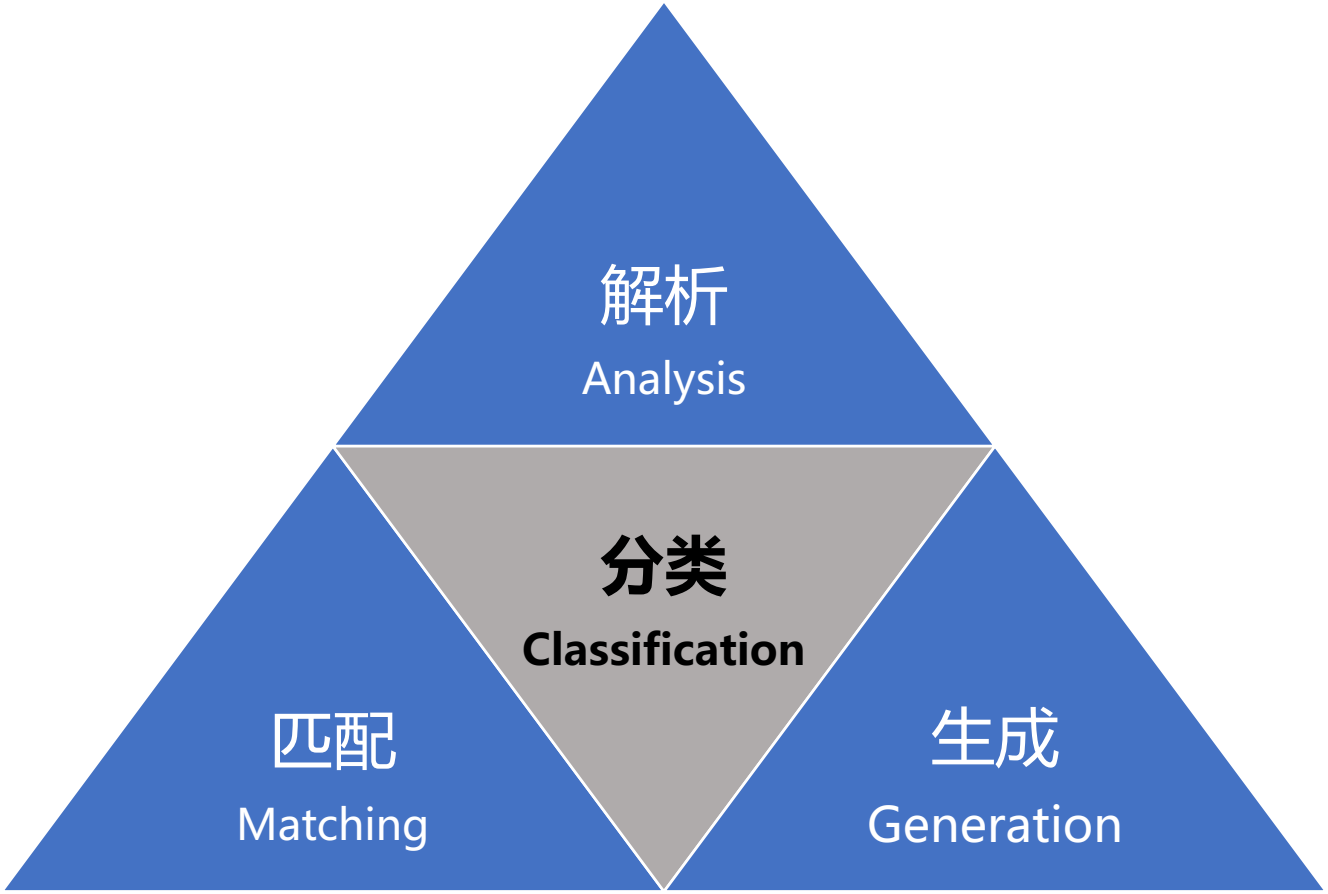
# 推理类问题


**DocRED: A Large-Scale Document-Level Relation Extraction Dataset**

**Yuan Yao**[1*], **Deming Ye**[1*], **Peng Li**[2], **Xu Han**[1], **Yankai Lin**[1], **Zhenghao Liu**[1], **Zhiyuan Liu**[1†], **Lixin Huang**[1], **Jie Zhou**[2], **Maosong Sun**[1]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
Institute for Artificial Intelligence, Tsinghua University, Beijing, China
State Key Lab on Intelligent Technology and Systems, Tsinghua University, Beijing, China
[2]Pattern Recognition Center, WeChat AI, Tencent Inc.
{yuan-yao18,ydm18}@mails.tsinghua.edu.cn

| Reasoning Types | % | Examples |
|---|---|---|
| Pattern recognition | 38.9 | [1] ***Me Musical Nephews*** is a **1942** one-reel animated cartoon directed by Seymour Kneitel and animated by Tom Johnson and George Germanetti. [2] Jack Mercer and Jack Ward wrote the script. ...<br>**Relation:** `publication_date` **Supporting Evidence: 1** |
| Logical reasoning | 26.6 | [1] "Nisei" is the ninth episode of the third season of the American science fiction television series The X-Files. ... [3] It was directed by David Nutter, and written by **Chris Carter**, Frank Spotnitz and Howard Gordon. ... [8] The show centers on FBI special agents ***Fox Mulder*** (David Duchovny) and Dana Scully (Gillian Anderson) who work on cases linked to the paranormal, called X-Files. ...<br>**Relation:** `creator` **Supporting Evidence: 1, 3, 8** |
| Coreference reasoning | 17.6 | [1] ***Dwight Tillery*** is an American politician of the Democratic Party who is active in local politics of Cincinnati, Ohio. ... [3] He also holds a law degree from the **University of Michigan Law School**. [4] ***Tillery*** served as mayor of Cincinnati from 1991 to 1993.<br>**Relation:** `educated_at` **Supporting Evidence: 1, 3** |
| Common-sense reasoning | 16.6 | [1] ***William Busac*** (1020-1076), son of William I, Count of Eu, and his wife Lesceline. ... [4] ***William*** appealed to King Henry I of France, who gave him in marriage **Adelaide**, the heiress of the county of Soissons. [5] **Adelaide** was daughter of Renaud I, Count of Soissons, and Grand Master of the Hotel de France. ... [7] ***William*** and **Adelaide** had four children: ...<br>**Relation:** `spouse` **Supporting Evidence: 4, 7** |

# 自然语言处理的四类问题



# NLP的“层面×任务”二维表

| | 分类 | 解析 | 匹配 | 生成 |
|---|---|---|---|---|
| **形式** | 文本分类 | 词性标注<br>句法分析 | 搜索 | 机械式文摘 |
| **语义** | 情感分析 | 命名实体识别<br>语义角色标注 | 问答 | 机器翻译 |
| **推理** | 隐式情感分析 | | 文本蕴含 | 写故事结尾 |
| **语用** | 反语 | | | 聊天 |

# NLPer的核心竞争力

- 一个核心
  - 结构化是NLP的核心问题
- 两个能力
  - 发现问题的能力
  - 解决问题的能力
- 三个优势
  - 对基本概念理解更准确
  - 对研究有更好的品味
  - 对数据更敏感

# 对基本概念理解更准确

- 评价方法
  - 评价指标
    - Accuracy, Precision, Recall, F1, MAP, P@XX, R@XX
  - 评价数据
    - 如：Ubuntu检索式对话任务的评价本身存在问题
    - NYT的关系抽取的测试数据使用Distance Supervision+人工Check构造
- Beam Search
  - 只在解码阶段应用存在偏置问题

# 对研究有更好的品味

- 会评判好的研究
  - 文章发表数量爆炸式增长，能够快速判断有价值的工作
  - 模型是越复杂、越炫酷越好么？
- 会做好的研究
  - 如何选择研究方向，冷门还是热门？
  - 从问题出发还是从模型出发？
  - 研究遇到瓶颈，是坚持到底还是及时止损？

# 对数据更敏感

- 观察实际数据
  - 积累直觉经验
- 对预测错误进行系统分析
  - 不仅仅关注性能指标
  - 对异常实验结果更冷静

# NLPer的核心竞争力

- 一个核心
  - 结构化是NLP的核心问题
- 两个能力
  - 发现问题的能力
  - 解决问题的能力
- 三个优势
  - 对基本概念理解更准确
  - 对研究有更好的品味
  - 对数据更敏感

谢谢！