



# 知识指导的自然语言处理

清华大学自然语言处理实验室

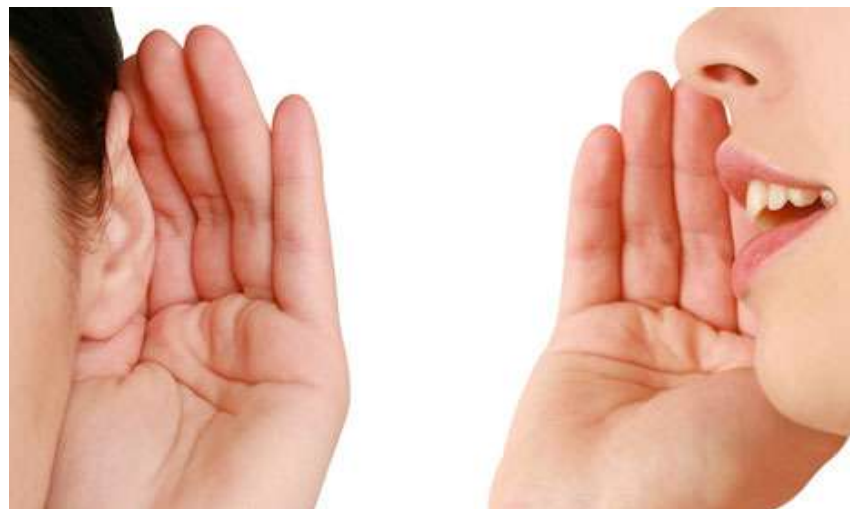
刘知远

# 自然语言

- 自然语言是人类间交流传播信息和知识的工具  
– 创新性，歧义性，>CFG

```
4  
5 int summary(void *barg,void *arg)  
6 {  
7     char *str = (char *)barg;  
8     st_board *board = (st_board *)arg;  
9     int ret = 0;  
10  
11     char *ptr_shuttercounter = ...
```

编程语言

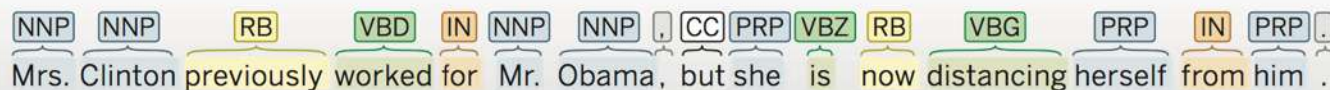


自然语言

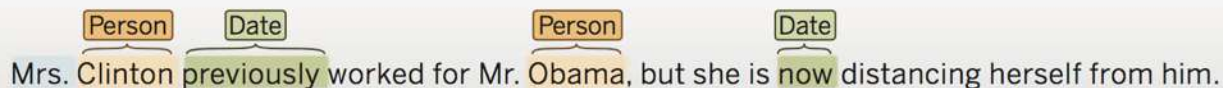
# 自然语言处理

- 自然语言处理旨在理解人类语言的语义信息
- 本质是从无结构序列中预测有结构语义

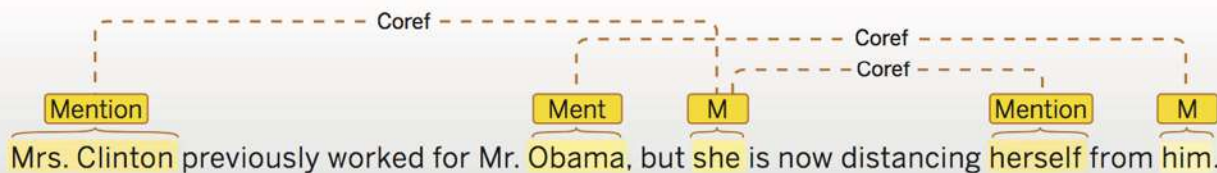
## Part of speech:



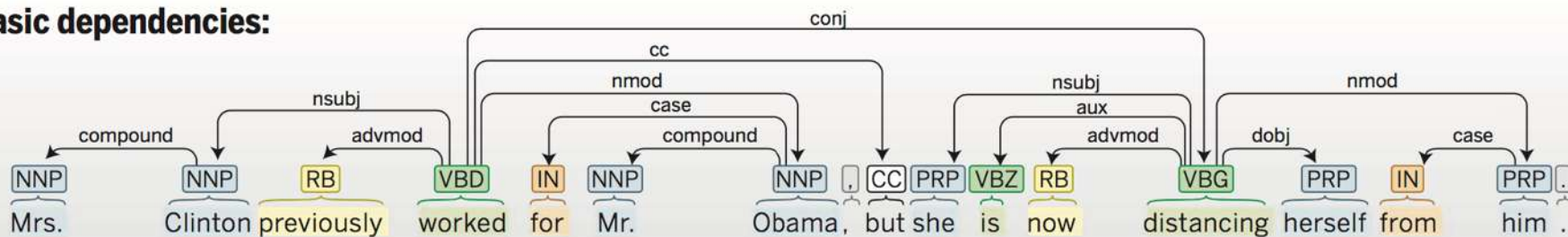
## Named entity recognition:



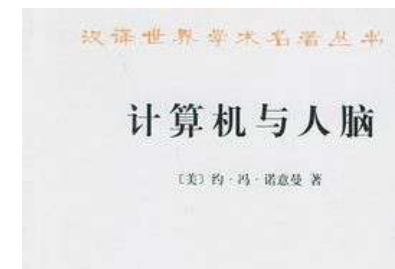
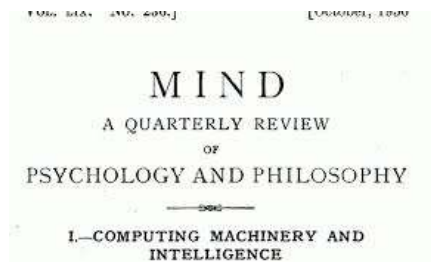
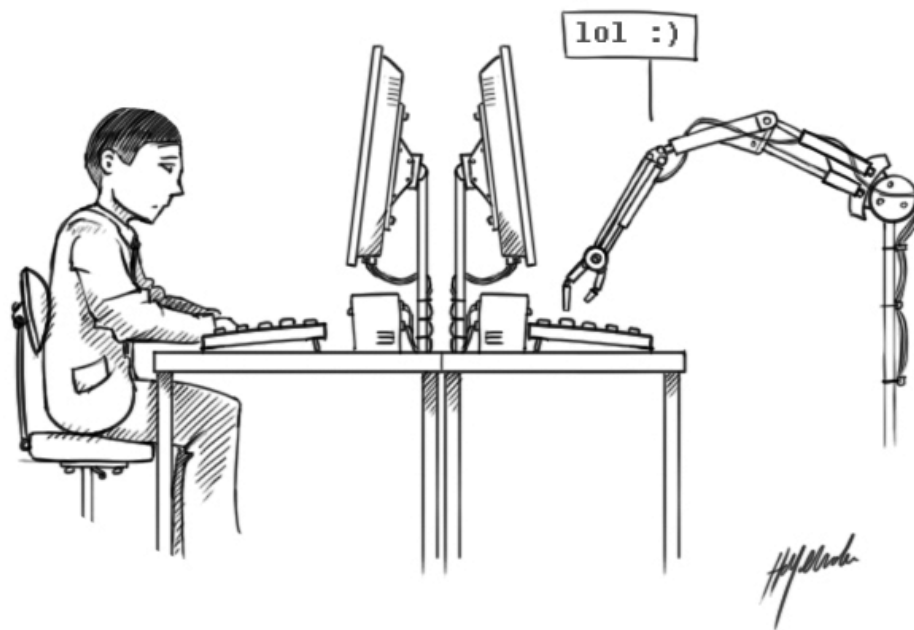
## Co-reference:



## Basic dependencies:



# 自然语言处理是AI关键问题

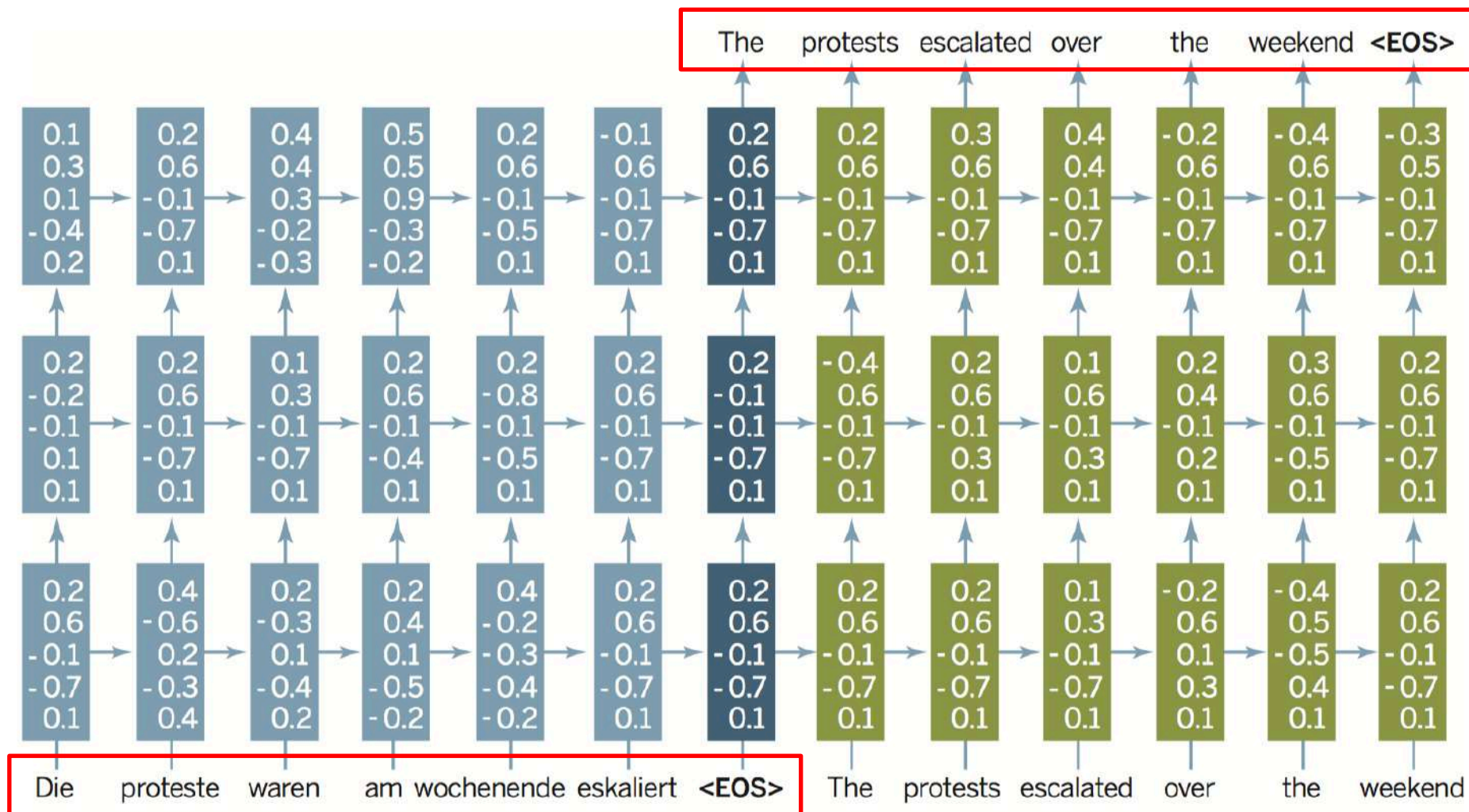


自然语言处理是实现人工智能、通过图灵测试的关键



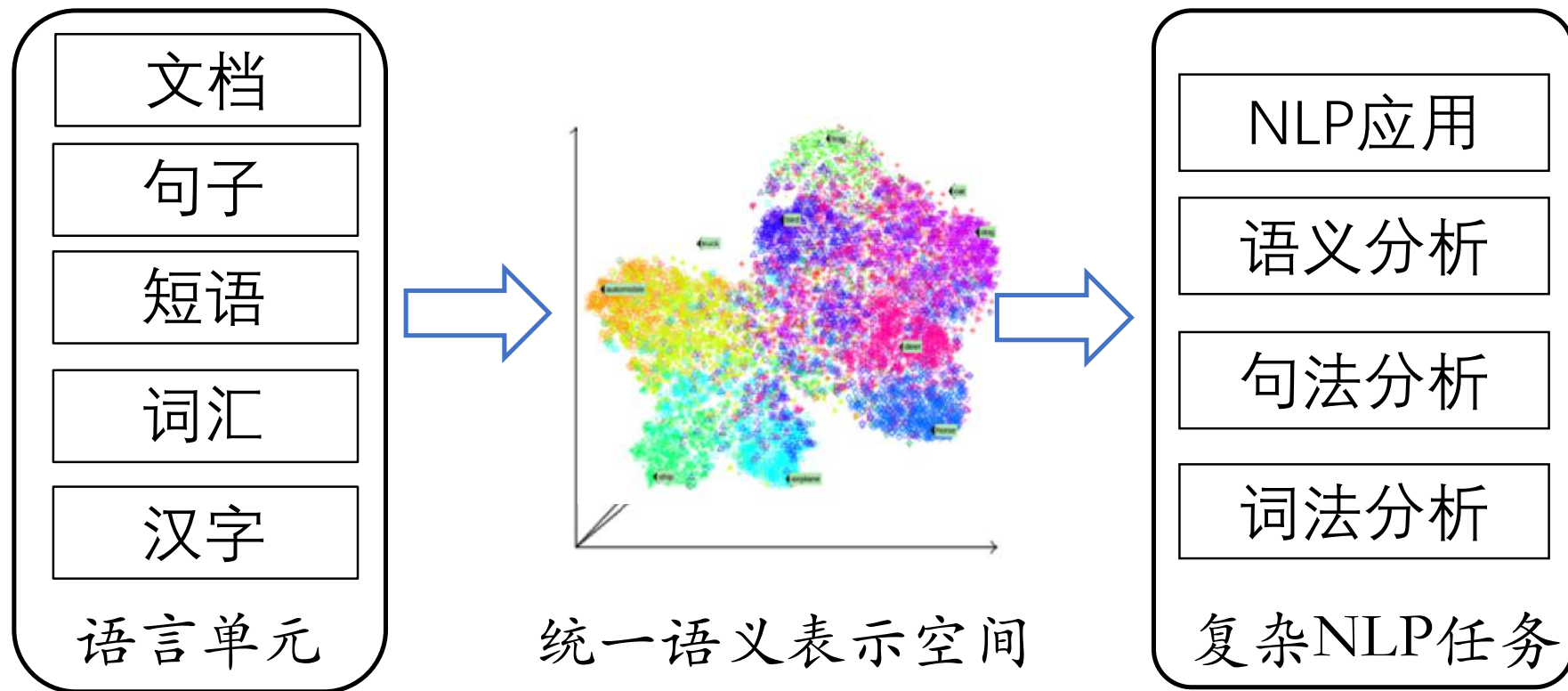
# 数据驱动的自然语言处理：深度学习

- 深度学习技术在自然语言处理取得了巨大突破



# 数据驱动的自然语言处理：深度学习

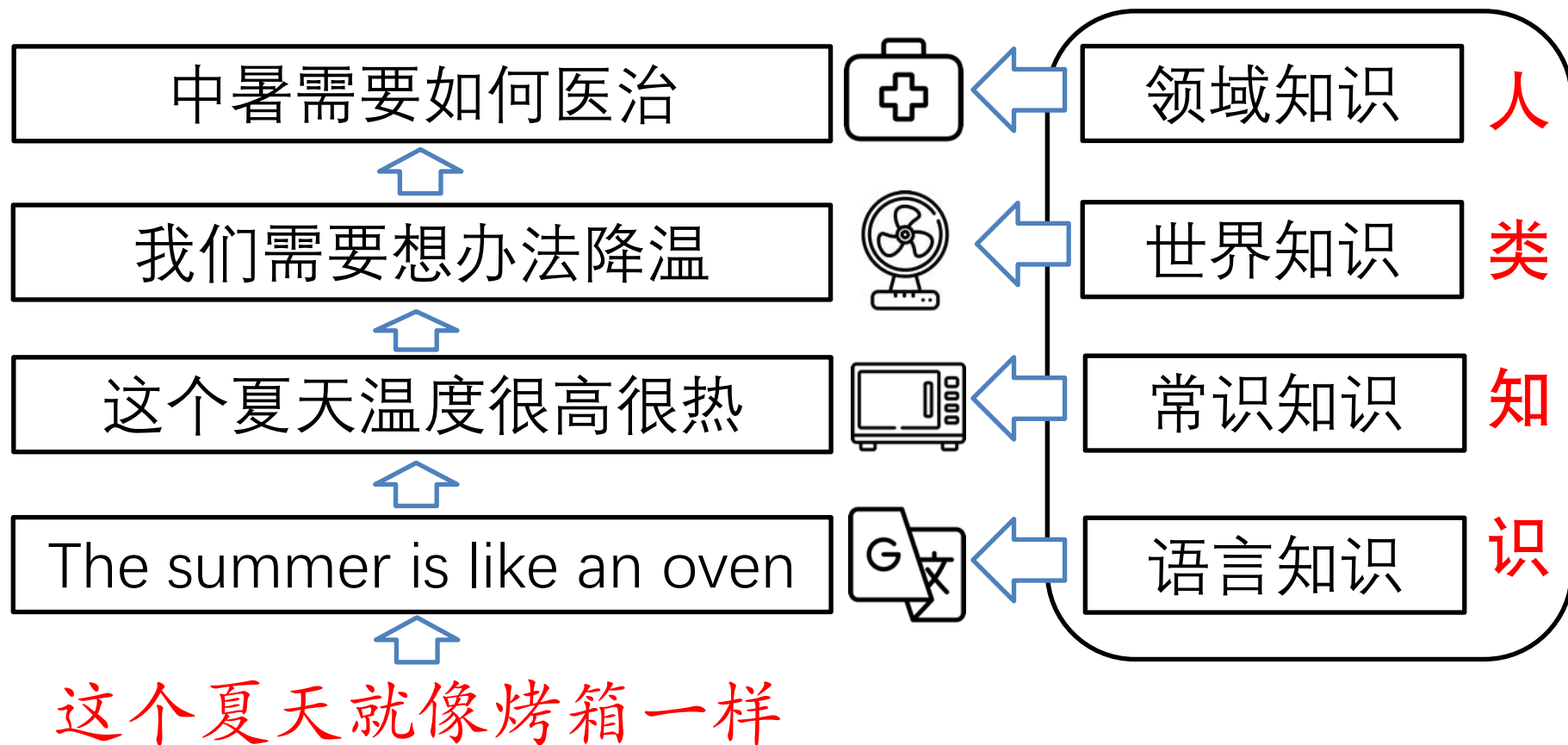
- 深度学习技术在自然语言处理取得了巨大突破



深度学习能够高效学习多粒度语言单元间复杂语义关联

# 面临挑战

- 对自然语言的深度理解需要复杂知识的支持



亟需知识支持实现NLP从字面意思到言外之意的跃迁

# 深度学习的挑战

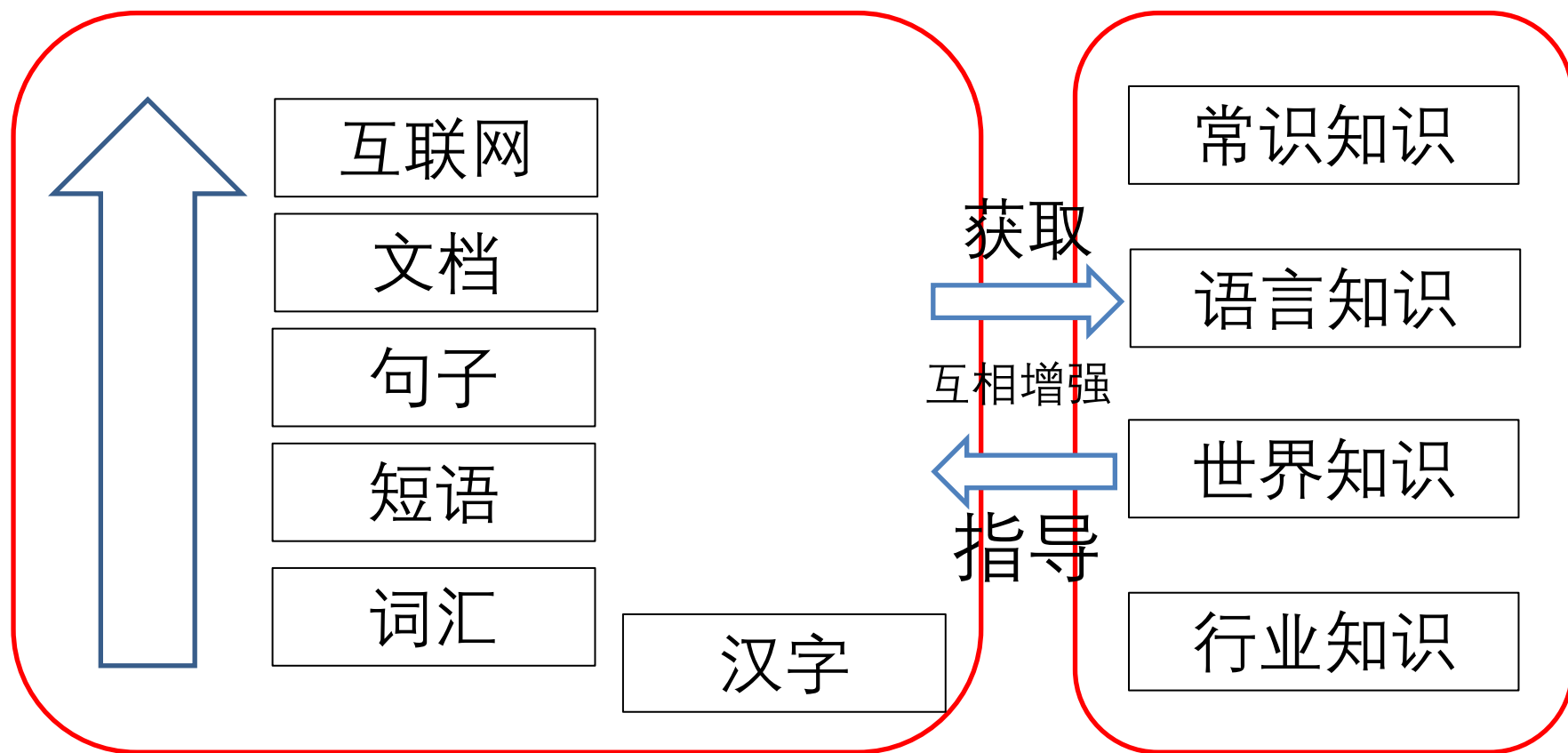


... we feel confident that more data and computation, in addition to recent advances in ML and deep learning, will lead to further substantial progress in NLP. However, the truly difficult problems of semantics, context, and knowledge will probably require new discoveries in linguistics and inference.



# 自然语言特点

- 自然语言文本蕴含丰富的语言知识和世界知识



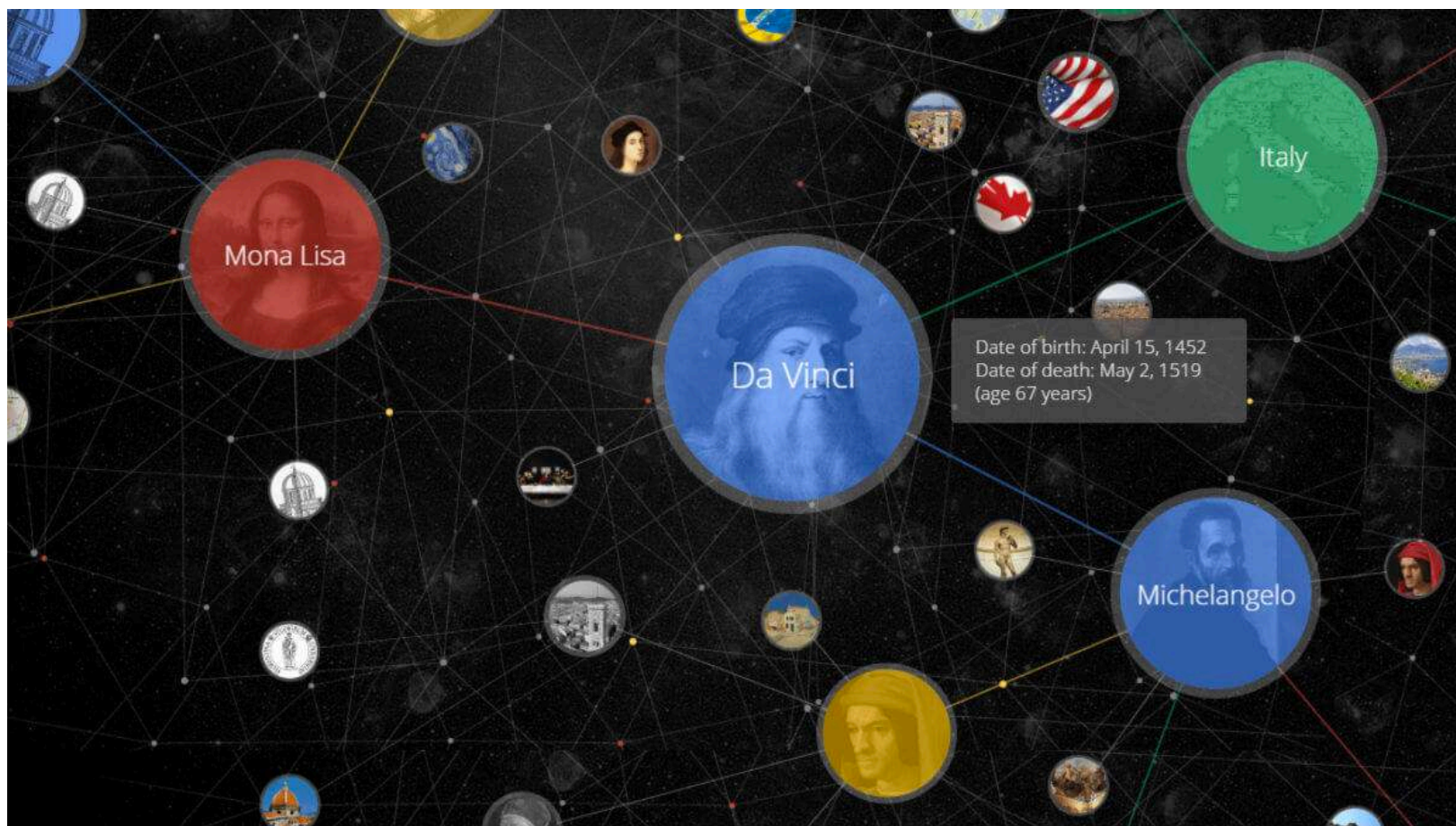
数据驱动

+

知识指导

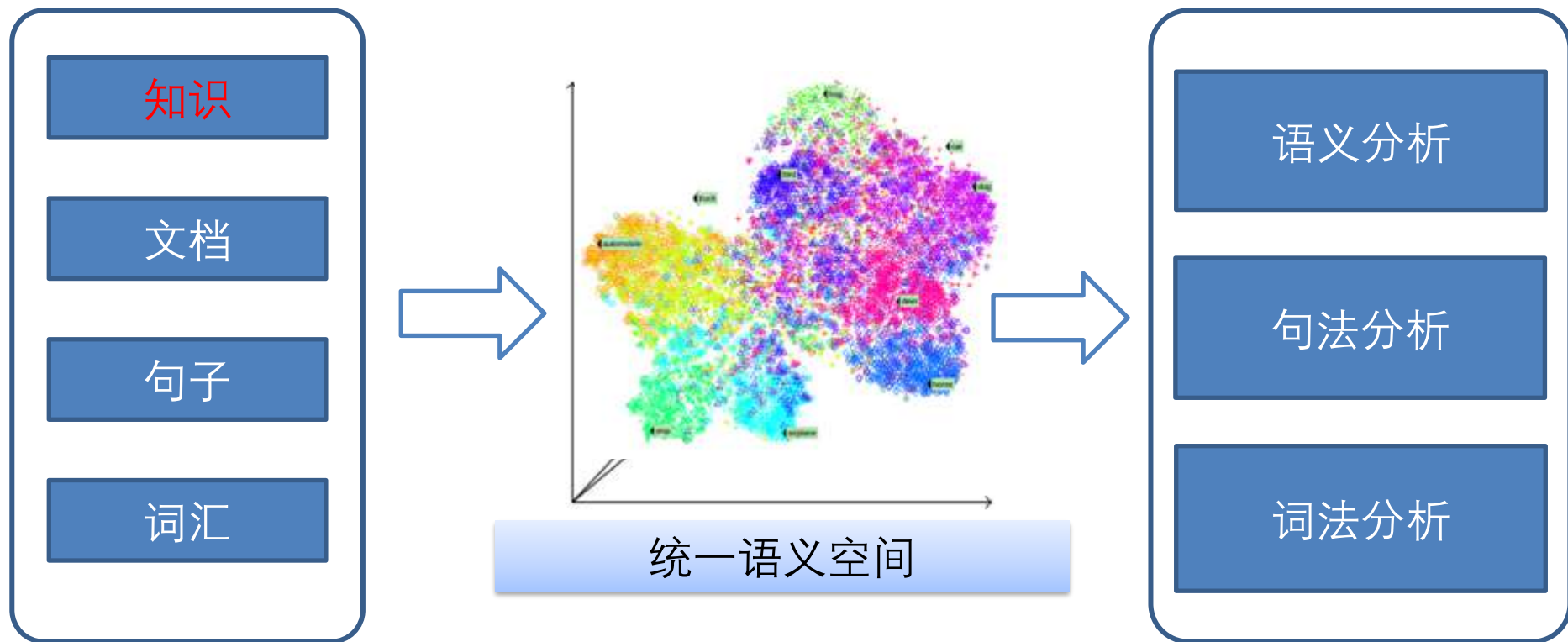
# 技术挑战

- 语言知识、世界知识均通过离散符号表示



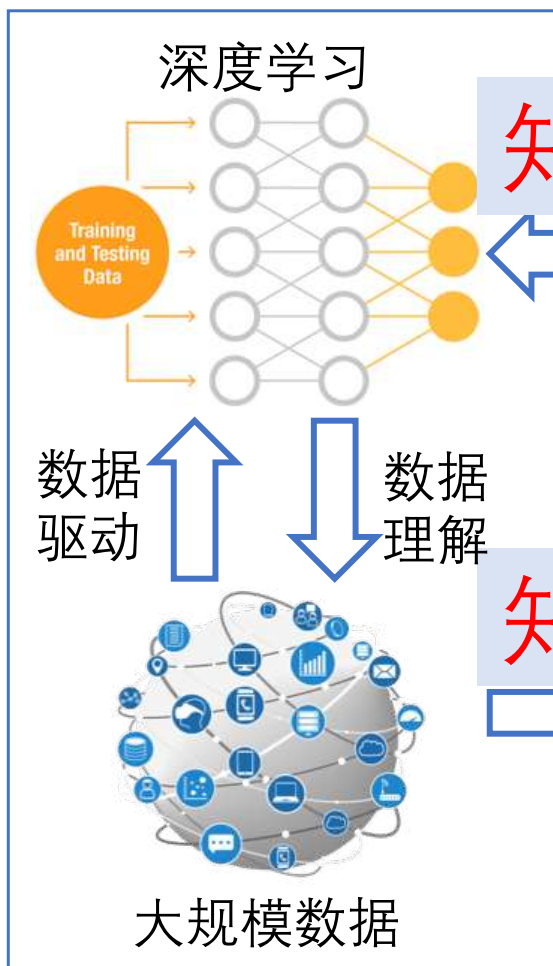
# 表示学习

- 分布式表示：实现跨粒度、跨领域、富知识的语言理解



# 研究思路

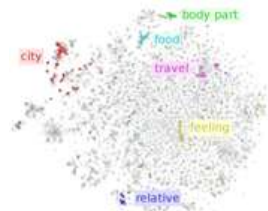
数据驱动的深度学习



知识指导

知识获取

知识表示



表示学习



知识图谱

符号表示的知识图谱

深度学习+知识图谱 双向驱动的自然语言处理技术体系

# 语言知识库



PRINCETON UNIVERSITY

**WordNet**  
A lexical database for English

基于《知网》的词汇语义相似度计算<sup>1</sup>

Word Similarity Computing Based on How-net

刘群<sup>\*</sup>、李素建<sup>\*</sup>

Qun LIU, Sujian LI

摘要

词义相似度计算在很多领域中都有广泛的应用，例如信息检索、信息抽取、文本分类、词义排歧、基于实例的机器翻译等等。词义相似度计算的两种基本方法是基于世界知识 (Ontology) 或某种分类体系 (Taxonomy) 的方法和基于统计的上下文向量空间模型方法。这两种方法各有优缺点。

《知网》是一部比较详尽的语义知识词典，受到了人们普遍的重视。不过，由于《知网》中对于一个词的语义采用的是一种多维的知识表示形式，这给词语相似度的计算带来了麻烦。这一点与 WordNet 和《同义词词林》不同。在 WordNet 和《同义词词林》中，所有同类的语义项 (WordNet 的 synset 或《同义词词林》的词群) 构成一个树状结构，要计算语义项之间的距离，只要计算树状结构中相应结点的距离即可。而在《知网》中词汇语义相似度的计算存在以下问题：

1. 每一个词的语义描述由多个义原组成；
2. 词语的语义描述中各个义原并不是平等的，它们之间有着复杂的关系，通过一种专门的知识描述语言来表示。

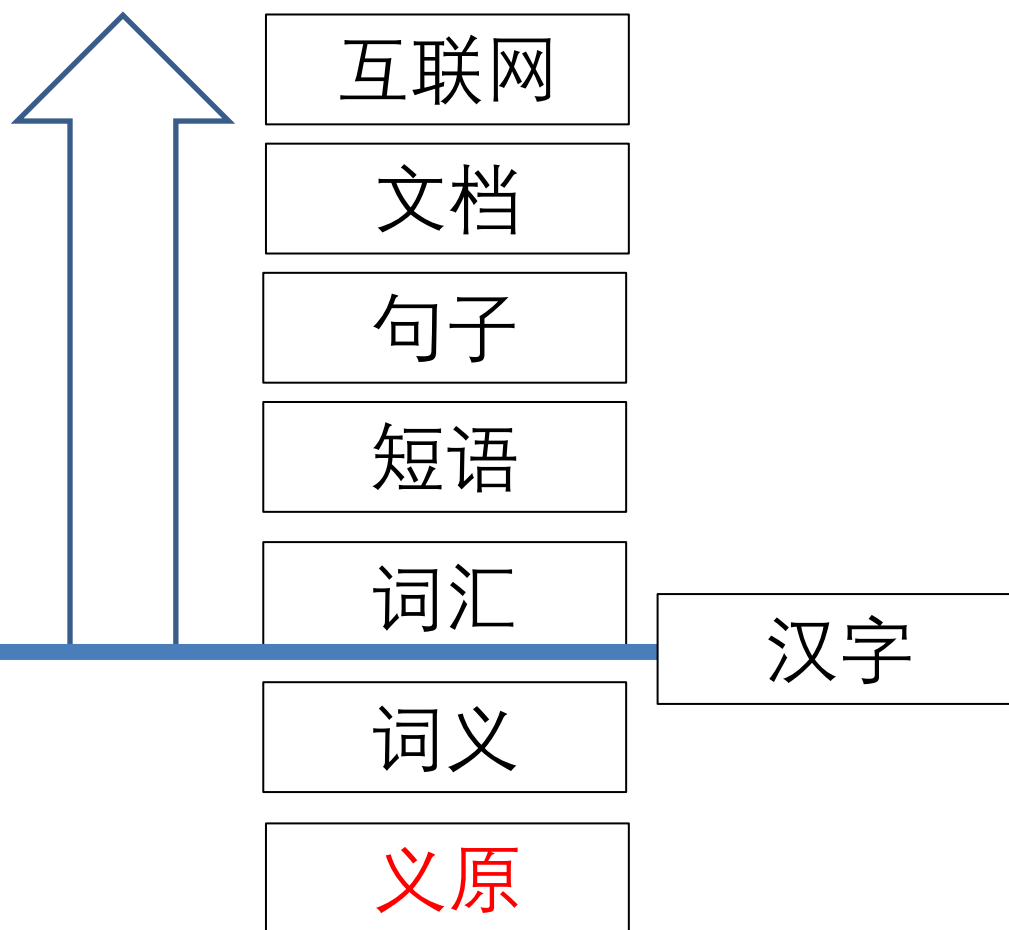
我们的工作主要包括：

1. 研究《知网》中知识描述语言的语法，了解其描述一个词义所用的多个义原之间的关系，区分其在词语相似度计算中所起的作用；我们采用一种更



# 自然语言特点

- 词汇或汉字是最小**使用单位**，但不是最小**语义单位**



# 义原知识与HowNet

- HowNet是**董振东、董强**父子毕三十年之功标注的大型语言知识库，主要面向中文的词汇与概念标注义原知识
- 秉承**还原论**思想，用义原（Sememe）标注词汇语义，义原顾名思义就是**原子语义**，即最基本的、不宜再分割的最小语义单位
- HowNet逐渐构建出一套精细的义原体系（包含约2000个义原），累计标注了数十万词汇/词义的语义信息

# HowNet—瞥

- 每个词义信息用义原标注，每个义原用 英文 | 中文 标明
- 义原之间还标记语义关系，如modifier, host, belong等

顶点#1

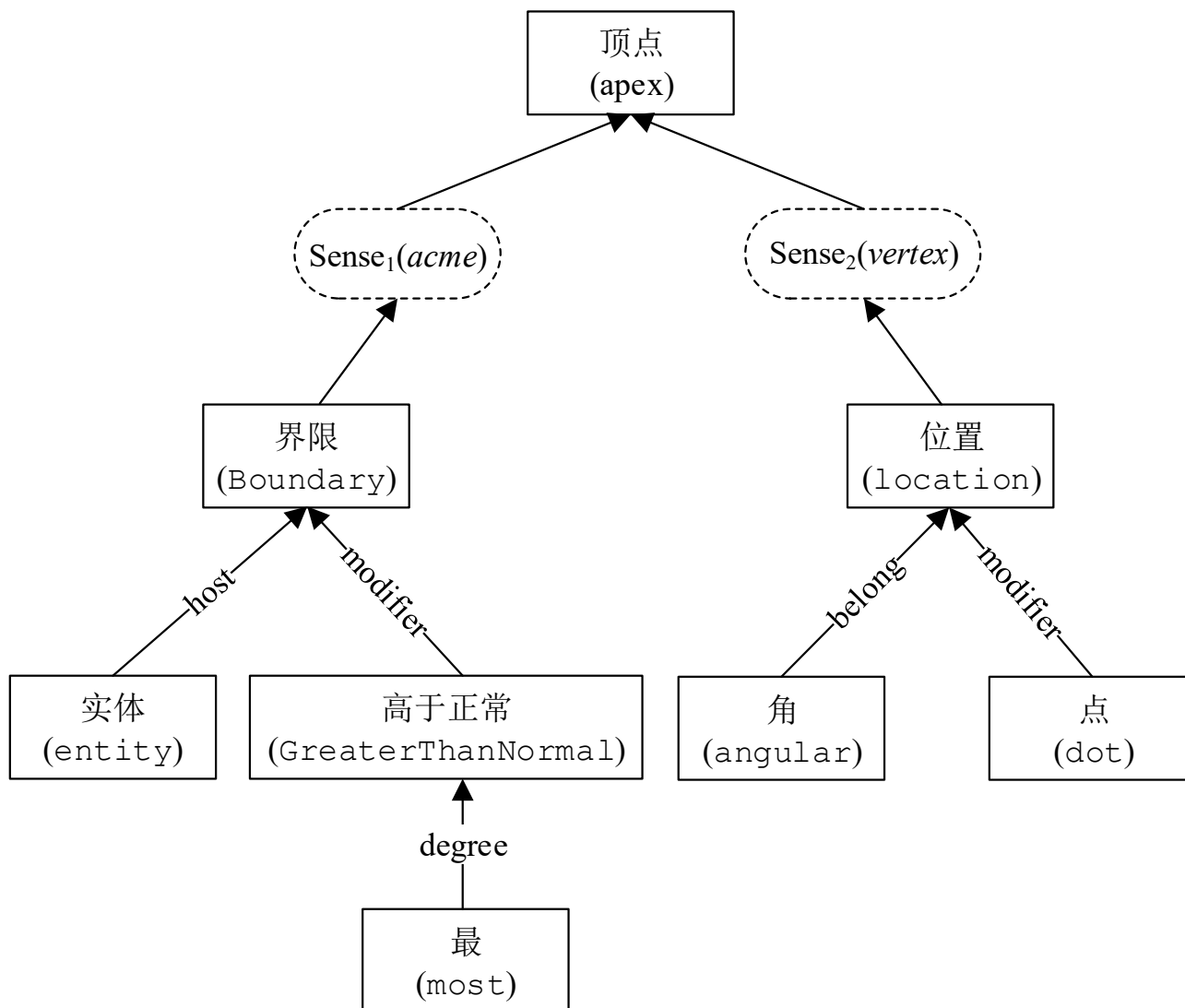
DEF={Boundary|界限:host={entity|实体},modifier={Greater Than Normal|高于正常:degree={most|最}}}

顶点#2

DEF={location|位置:belong={angular|角},modifier={dot|点}}

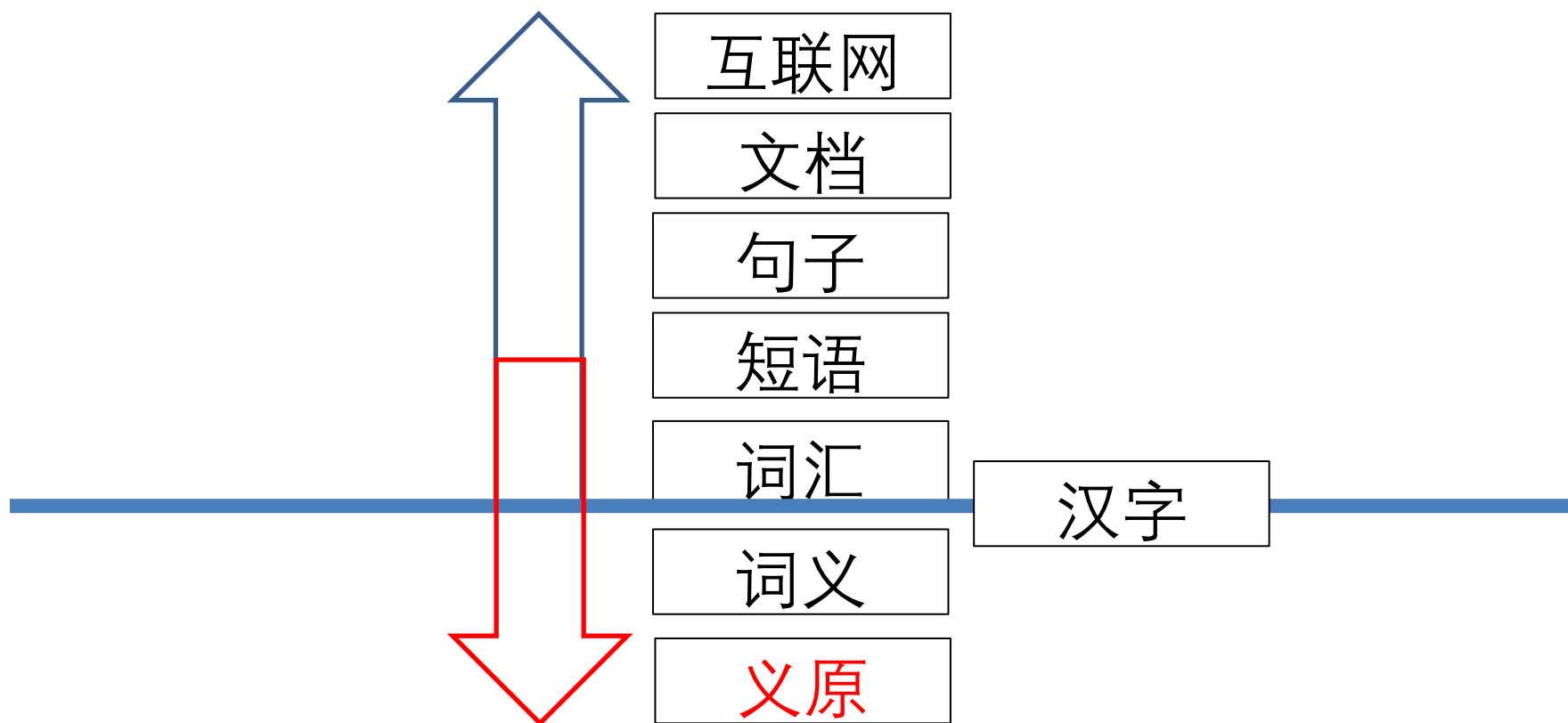
# HowNet—瞥

- 义原知识带有层次结构



# 深度学习时代HowNet的意义

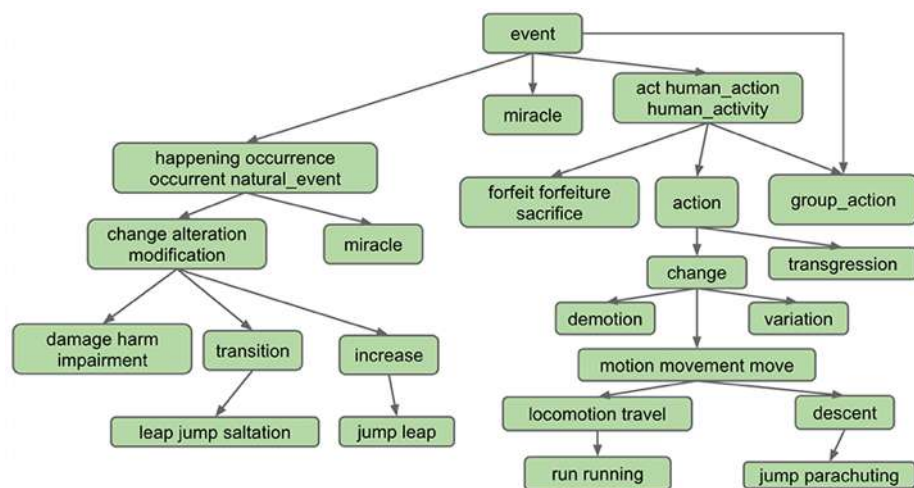
- 在自然语言理解方面，更贴近语言本质特点
  - 义原标注体系是突破词汇屏障，深入了解词汇背后丰富语义信息的重要通道



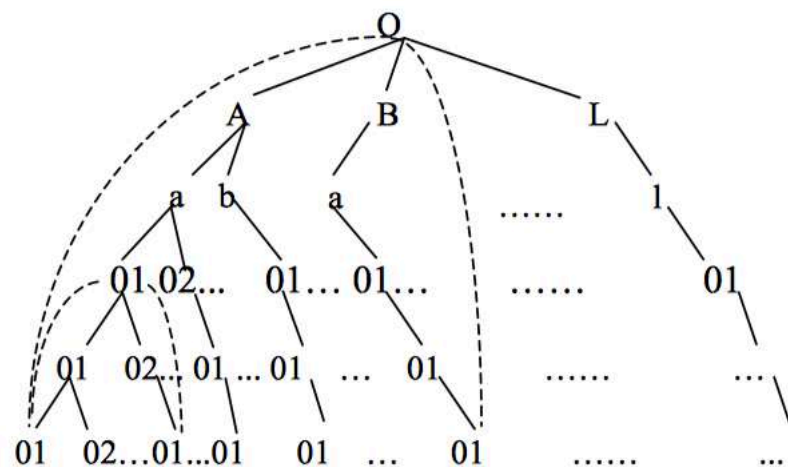


# 深度学习时代HowNet的意义

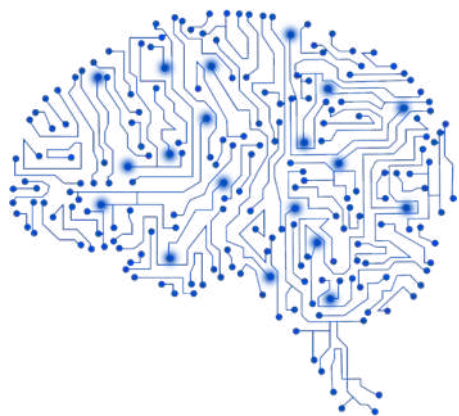
- 在融入深度学习方面，具有无可比拟优势
  - 与WordNet、同义词词林等知识库组织模式不同
  - HowNet通过统一义原标注体系直接精准刻画语义信息。每个义原**含义明确固定**，可被直接作为**语义标签**融入机器学习模型



WordNet Synset体系

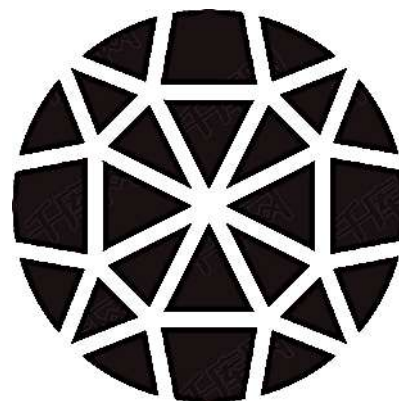
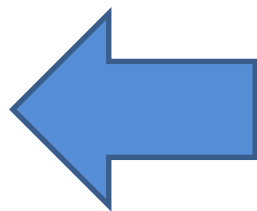


同义词词林层次类别体系



数据驱动的  
深度学习

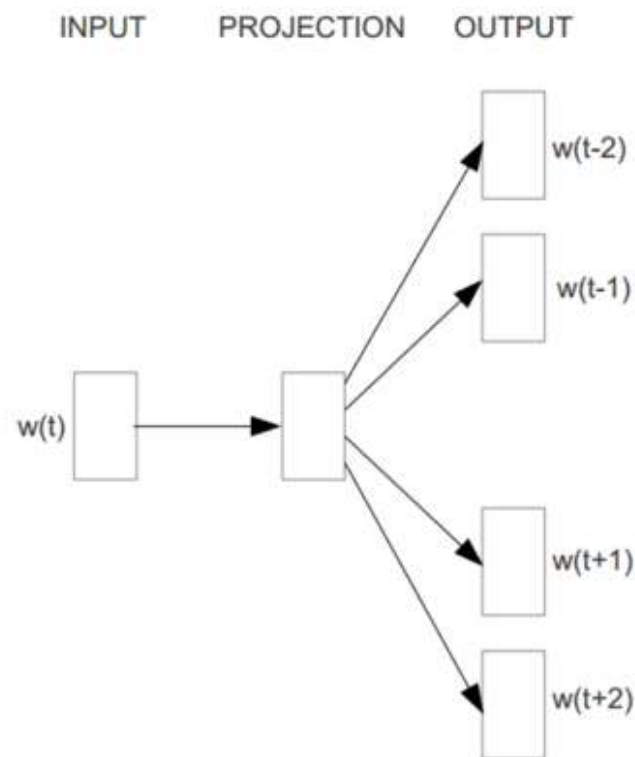
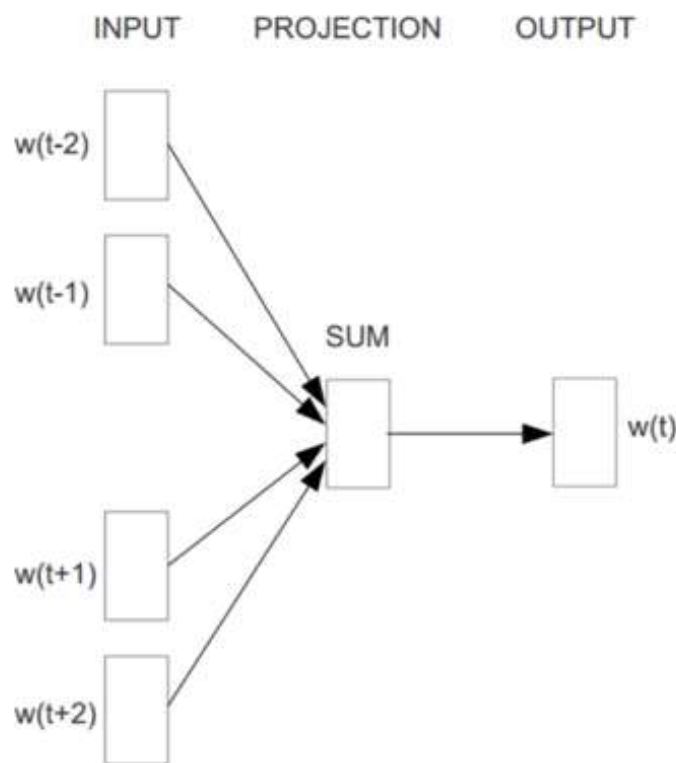
知识指导



符号表示的  
义原知识

# 词义分布式表示学习

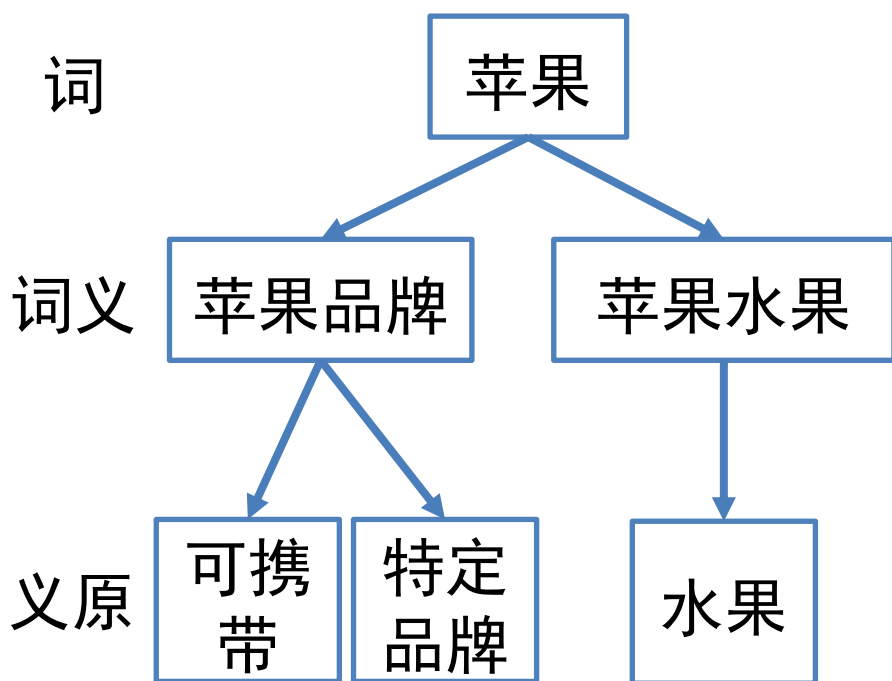
- 深度学习利用纯数据驱动方法学习语义表示



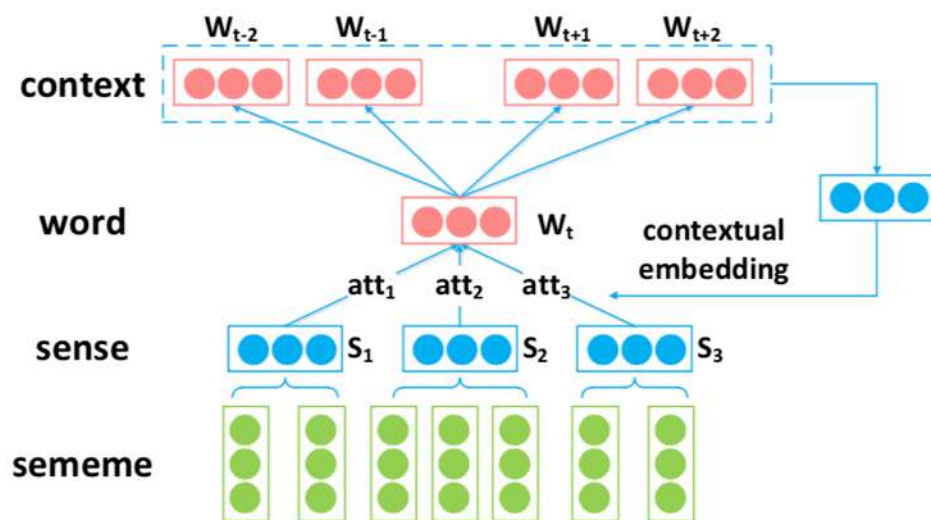
word2vec

# 融合义原知识的词义表示学习

- 考虑HowNet的词义-义原标注信息，提升词义表示性能



HowNet词义-义原标注示例



义原-词义-词汇的联合表示学习模型

# 实验结果

- 在词相似度计算和类比推理任务上的性能得到显著提升

| Model     | Accuracy    |             |              |             | Mean Rank    |             |              |             |
|-----------|-------------|-------------|--------------|-------------|--------------|-------------|--------------|-------------|
|           | Capital     | City        | Relationship | All         | Capital      | City        | Relationship | All         |
| CBOW      | 49.8        | 85.7        | <b>86.0</b>  | 64.2        | 36.98        | 1.23        | 62.64        | 37.62       |
| GloVe     | 57.3        | 74.3        | 81.6         | 65.8        | 19.09        | 1.71        | 3.58         | 12.63       |
| Skip-gram | 66.8        | 93.7        | 76.8         | 73.4        | 137.19       | 1.07        | 2.95         | 83.51       |
| SSA       | 62.3        | 93.7        | 81.6         | 71.9        | 45.74        | 1.06        | 3.33         | 28.52       |
| MST       | 65.7        | 95.4        | 82.7         | 74.5        | 50.29        | 1.05        | 2.48         | 31.05       |
| SAC       | 79.2        | 97.7        | 75.0         | 81.0        | 28.88        | 1.02        | 2.23         | 18.09       |
| SAT       | <b>82.6</b> | <b>98.9</b> | 80.1         | <b>84.5</b> | <b>14.78</b> | <b>1.01</b> | <b>1.72</b>  | <b>9.48</b> |

类比推理任务评测结果，其中SAC、SAT代表两种本工作提出的模型



# 实验结果

- 能够有效根据上下文信息实现词义消歧

| 上下文词 | 义原 “首都” | 义原 “古巴” |
|------|---------|---------|
| 古巴   | 0.39    | 0.42    |
| 俄罗斯  | 0.39    | -0.09   |
| 雪茄   | 0.00    | 0.36    |

上下文词对“哈瓦那”义原注意力值示例

| 例句                     | 词义1：概率    | 词义2：概率    |
|------------------------|-----------|-----------|
| <b>苹果</b> 素有果中王美称      | 苹果品牌：0.28 | 苹果水果：0.72 |
| <b>苹果</b> 电脑无法正常启动     | 苹果品牌：0.87 | 苹果水果：0.13 |
| 八支 <b>队伍</b> 进入第二阶段团体赛 | 团体：0.90   | 部队：0.10   |
| 公安基层 <b>队伍</b> 组织建设    | 团体：0.15   | 部队：0.85   |

根据上下文消歧结果示例

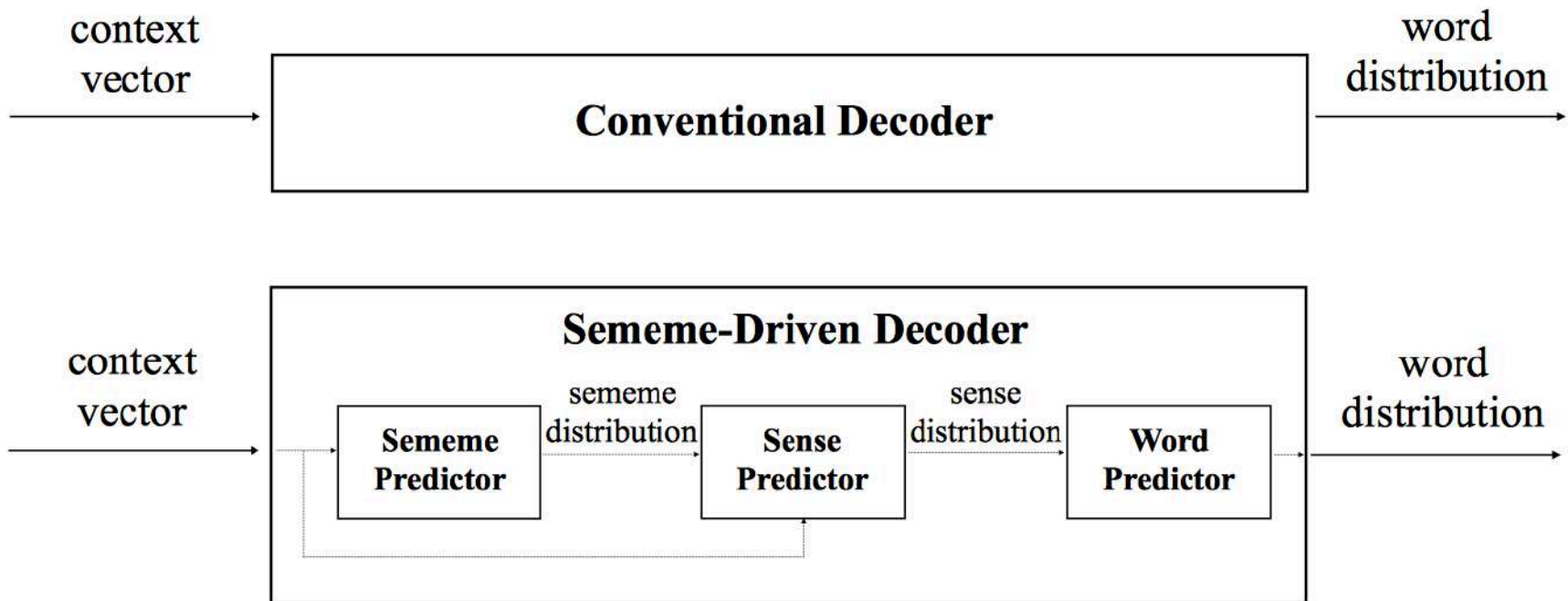
# 神经语言模型

- 语言模型是自然语言处理的核心任务
- N-Gram是前深度学习时代的代表语言模型，深度学习框架CNN、RNN即用来学习语言模型
- 马尔科夫性：当前词出现的概率，依赖于上下文出现的词

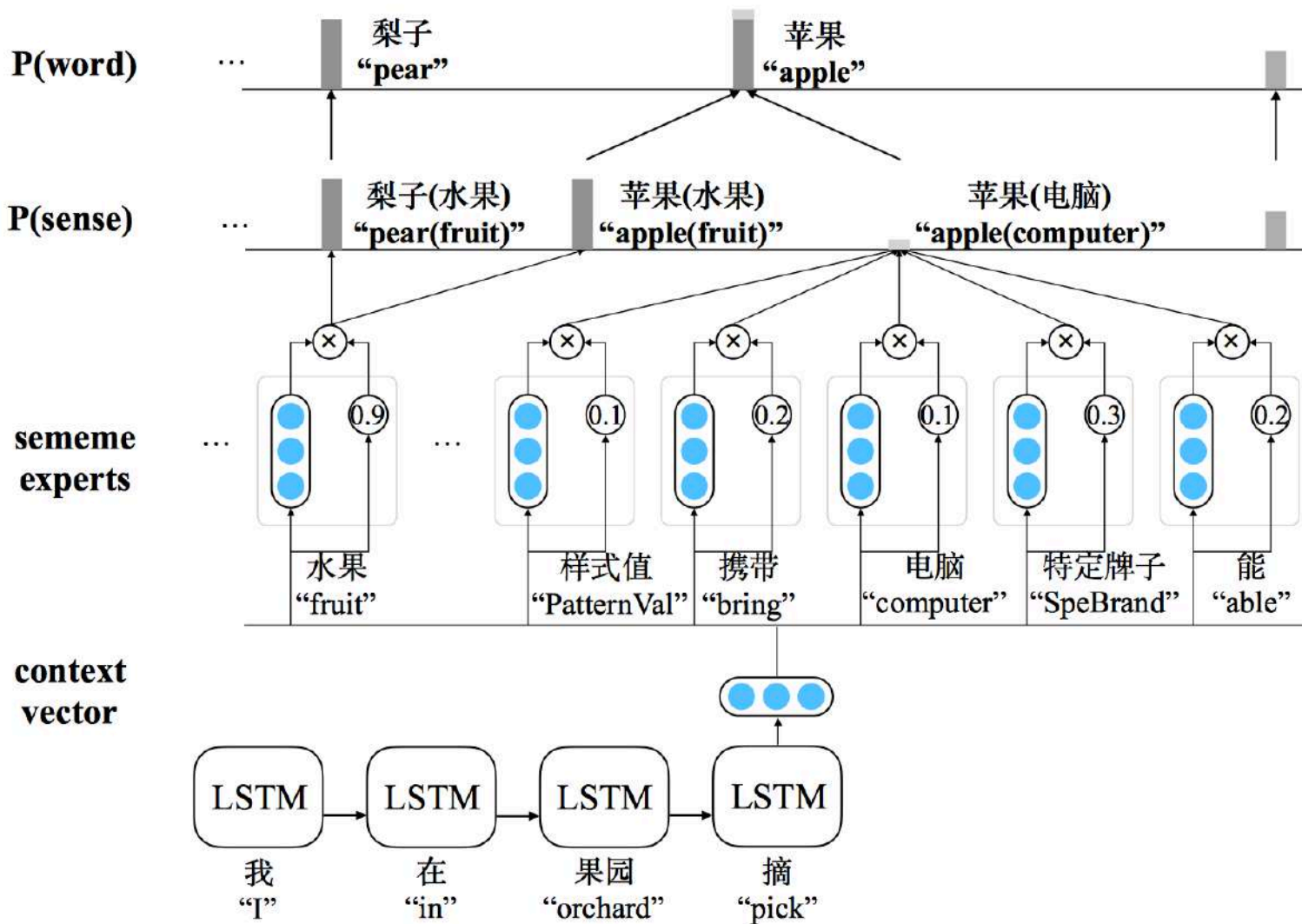
*The U.S. trade deficit last year is initially  
estimated to be 40 billion \_\_\_\_\_.*

# 融合义原知识的神经语言模型

- 传统深度学习语言模型是纯数据驱动模型
- 目标：建立义原知识驱动的语言模型



# 融合义原知识的神经语言模型



# 实验结果

- 义原驱动的语言模型（SDLM）普遍优于已有复杂语言模型

| Model                 | #Paras | Validation   | Test         |
|-----------------------|--------|--------------|--------------|
| LSTM (medium)         | 24M    | 116.46       | 115.51       |
| + cHSM                | 24M    | 129.12       | 128.12       |
| + tHSM                | 24M    | 151.00       | 150.87       |
| Tied LSTM (medium)    | 15M    | 105.35       | 104.67       |
| + cHSM                | 15M    | 116.78       | 115.66       |
| + MoS                 | 17M    | 98.47        | 98.12        |
| + SDLM                | 17M    | <b>97.75</b> | <b>97.32</b> |
| LSTM (large)          | 76M    | 112.39       | 111.66       |
| + cHSM                | 76M    | 120.07       | 119.45       |
| + tHSM                | 76M    | 140.41       | 139.61       |
| Tied LSTM (large)     | 56M    | 101.46       | 100.71       |
| + cHSM                | 56M    | 108.28       | 107.52       |
| + MoS                 | 67M    | 94.91        | 94.40        |
| + SDLM                | 67M    | <b>94.24</b> | <b>93.60</b> |
| AWD-LSTM <sup>4</sup> | 26M    | 89.35        | 88.86        |
| + MoS                 | 26M    | 92.98        | 92.76        |
| + SDLM                | 27M    | <b>88.16</b> | <b>87.66</b> |



# 实验结果

## Example (1)

去年 美国 贸易逆差 初步 估计 为 <N> \_\_\_\_\_ 。

The U.S. trade deficit last year is initially estimated to be <N> \_\_\_\_\_ .

### Top 5 word prediction

美元 “**dollar**” , “.” 。 “.”  
日元 “yen” 和 “and”

### Top 5 sememe prediction

商业 “**commerce**” 金融 “**finance**” 单位 “**unit**”  
多少 “amount” 专 “proper name”

## Example (2)

阿 总理 \_\_\_\_\_ 已 签署 了 一 项 命令 。

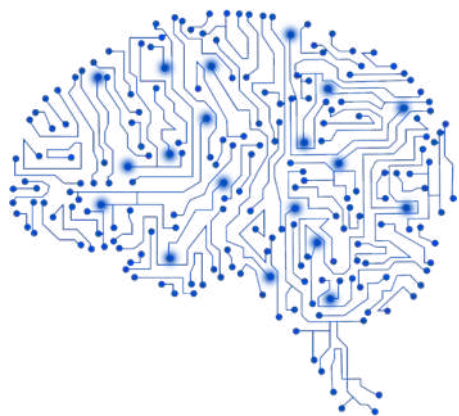
Albanian Prime Minister \_\_\_\_\_ has signed an order.

### Top 5 word prediction

内 “inside” <unk> 在 “at”  
塔 “tower” 和 “and”

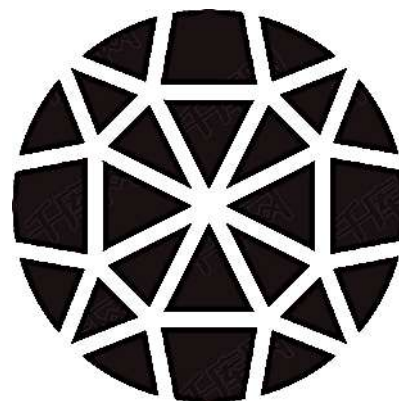
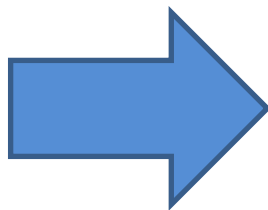
### Top 5 sememe prediction

政 “**politics**” 人 “**person**” 花草 “flowers”  
担任 “**undertake**” 水域 “waters”



数据驱动的  
深度学习

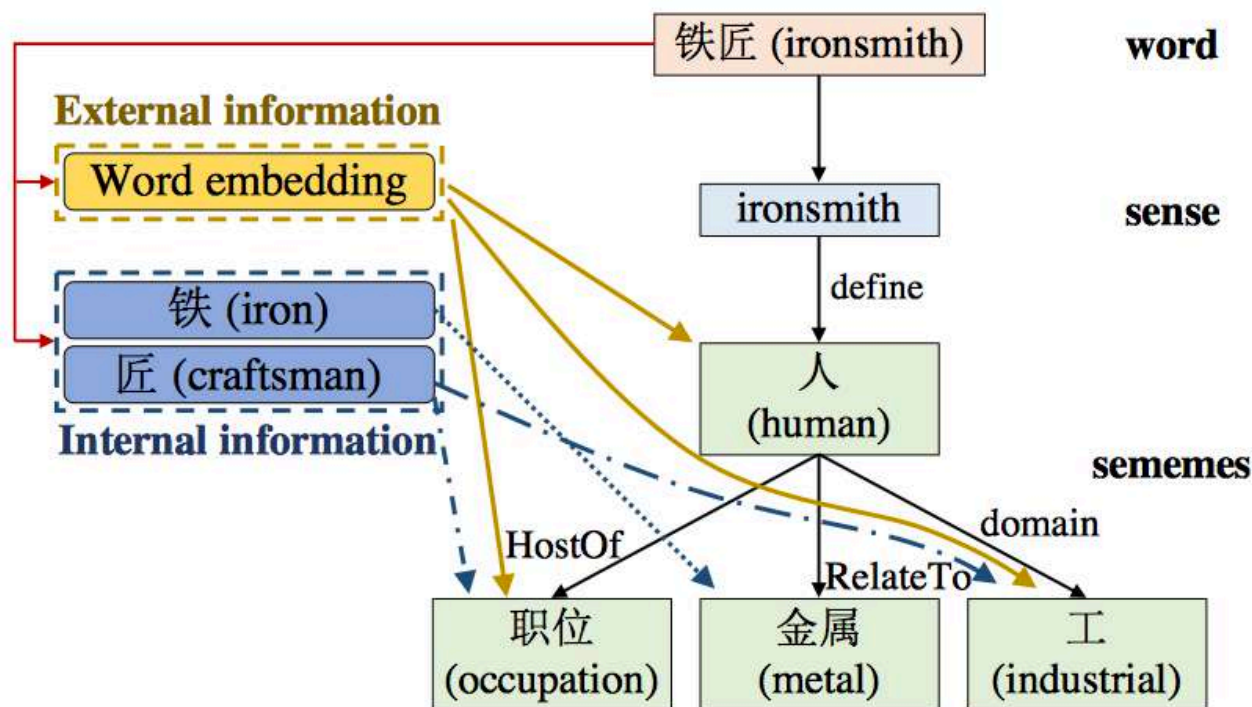
知识获取



符号表示的  
义原知识

# 基于语义表示学习的义原推荐

- 义原自动推荐：实现义原知识库与时俱进，提升标注一致性



Ruobing Xie, Xingchi Yuan, Zhiyuan Liu, Maosong Sun. Lexical Sememe Prediction via Word Embeddings and Matrix Factorization. IJCAI 2017.

Huiming Jin, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, Leyu Lin. Incorporating Chinese Characters of Words for Lexical Sememe Prediction. ACL 2018.

# 实验结果

- 将两种方法相融合，能够显著提升义原推荐效果。词性、词频有显著影响。

| Method            | MAP          |
|-------------------|--------------|
| SPSE              | 0.554        |
| SPASE             | 0.506        |
| GloVe+LR          | 0.662        |
| SPWE              | 0.676        |
| <b>SPWE+SPASE</b> | <b>0.683</b> |
| <b>SPWE+SPSE</b>  | <b>0.713</b> |

义原推荐效果

| POS       | number of words | MAP   |
|-----------|-----------------|-------|
| adverb    | 136             | 0.568 |
| adjective | 808             | 0.544 |
| verb      | 1,867           | 0.583 |
| noun      | 3,556           | 0.747 |

不同词性的词汇义原推荐效果

| word frequency | number of words | MAP   |
|----------------|-----------------|-------|
| <800           | 1,659           | 0.817 |
| 800 - 3,000    | 1,494           | 0.736 |
| 3,001 - 15,000 | 1,672           | 0.690 |
| >15,000        | 1,311           | 0.596 |

不同词频的词汇义原推荐效果

| words              | Top 5 sememes prediction   |
|--------------------|--|
| 网迷(webaholic)      | 人(human), 因特网(internet), 经常(frequency), 利用(use), 喜欢(fond of)             |
| 专递(express mail)   | 邮寄(post), 信件(letter), 快(fast), 事情(fact), 车(landvehicle)                  |
| 电影业(film industry) | 事务'affairs), 艺(entertainment), 表演物(shows), 拍摄(take picture), 制造(produce) |
| 漂流(rafting)        | 船(ship), 旅游(tour), 游(swim), 水域(waters), 消闲(whileaway)                    |
| 公羊(ram)            | 牲畜(livestock), 男(male), 女(female), 走兽(beast), 饲养(foster)                 |

# 义原知识计算相关论文

<https://github.com/thunlp/SCPapers>

- Fanchao Qi, Junjie Huang, Chenghao Yang, Zhiyuan Liu, Xiao Chen, Qun Liu, Maosong Sun. **Modeling Semantic Compositionality with Sememe Knowledge**. ACL 2019.
- Yihong Gu, Jun Yan, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin and Leyu Lin. **Language Modeling with Sparse Product of Sememe Experts**. EMNLP 2018.
- Fanchao Qi, Yankai Lin, Maosong Sun, Hao Zhu, Ruobing Xie, Zhiyuan Liu. **Cross-lingual Lexical Sememe Prediction**. EMNLP 2018.
- Huiming Jin, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, Leyu Lin. **Incorporating Chinese Characters of Words for Lexical Sememe Prediction**. ACL 2018.
- Xiangkai Zeng, Cheng Yang, Cunchao Tu, Zhiyuan Liu, Maosong Sun. **Chinese LIWC Lexicon Expansion via Hierarchical Classification of Word Embeddings with Sememe Attention**. AAAI 2018.
- Ruobing Xie, Xingchi Yuan, Zhiyuan Liu, Maosong Sun. **Lexical Sememe Prediction via Word Embeddings and Matrix Factorization**. IJCAI 2017.
- Yilin Niu, Ruobing Xie, Zhiyuan Liu, Maosong Sun. **Improved Word Representation Learning with Sememes**. ACL 2017.



# 世界知识库

- 以Google Knowledge Graphs为代表的世界知识库，用三元组形式记录知识

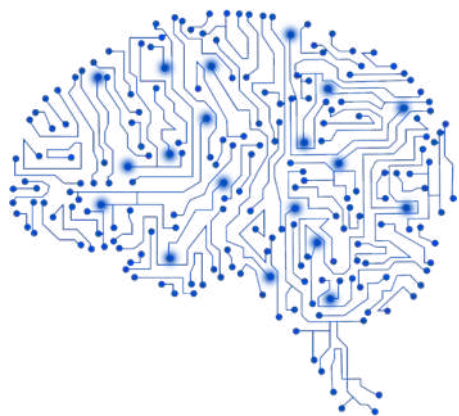


莎士比亚

写作

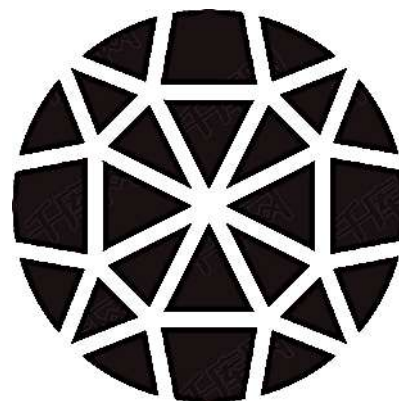
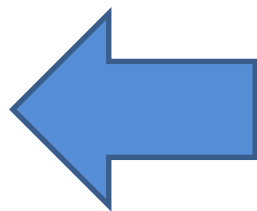


罗密欧与朱丽叶



数据驱动的  
深度学习

知识指导

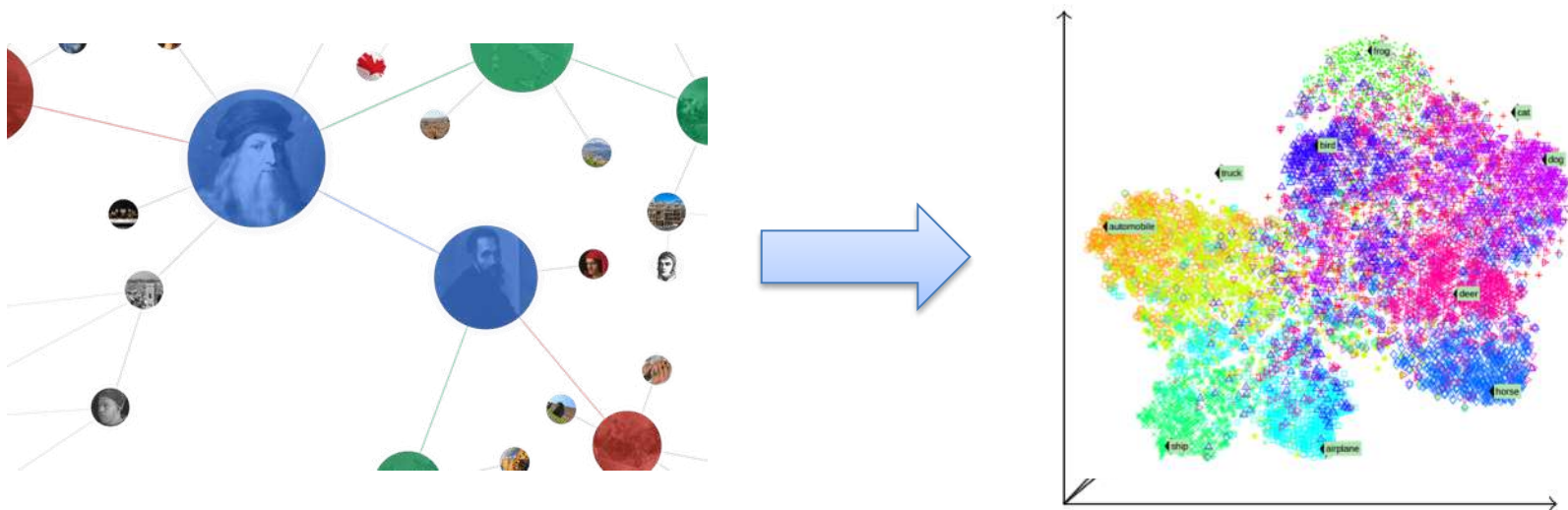


符号表示的  
世界知识



# 知识表示学习

- 基于知识图谱的知识表示学习





# 评价任务：链接预测

WALL-E    \_has\_genre    ?



# 评价任务：链接预测

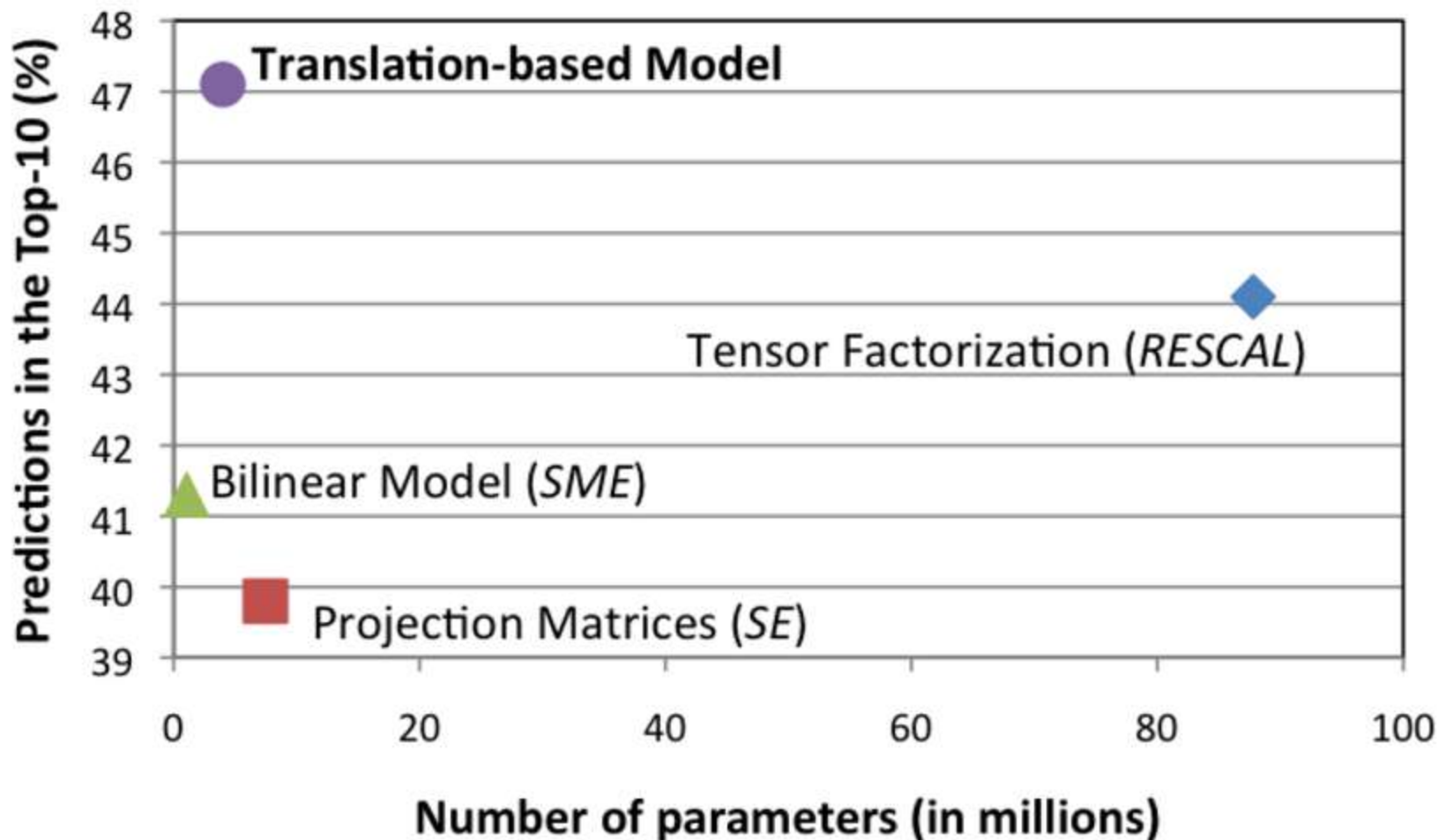
WALL-E      \_has\_genre



Animation  
Computer animation  
Comedy film  
Adventure film  
Science Fiction  
Fantasy  
Stop motion  
Satire  
Drama  
Connecting

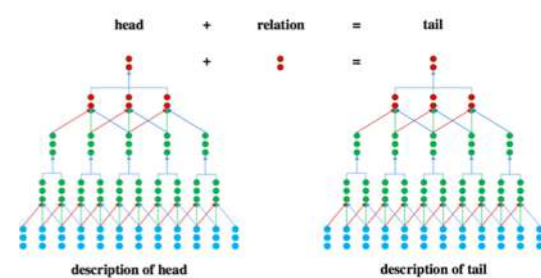
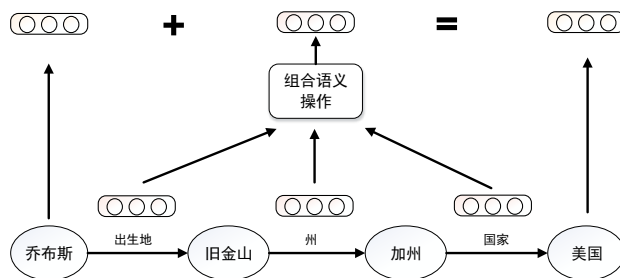
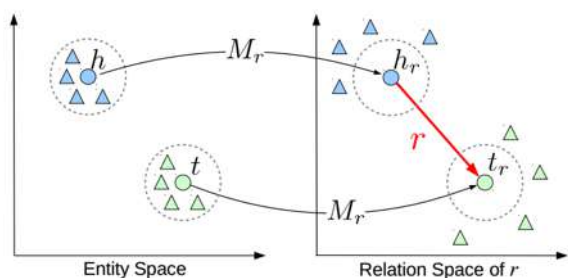
# 链接预测性能比较

Freebase15K



# 世界知识的分布式表示学习

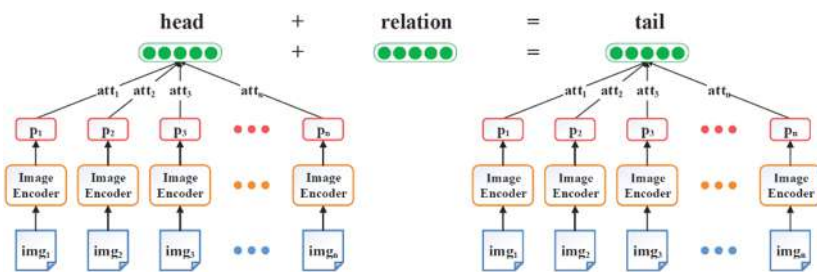
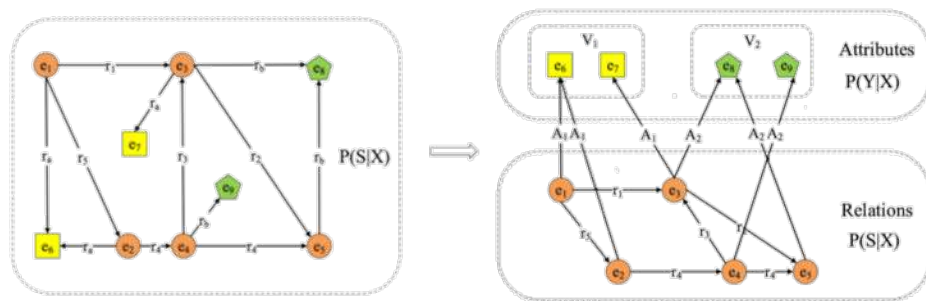
- 利用知识图谱和实体描述、类别和图像等外部信息，实现高效知识表示学习



考虑复杂关系类型的知识表示  
TransR (AAAI 2015)

考虑关系路径的知识表示  
PTransE (EMNLP 2015)

考虑实体描述信息的知识表示  
DKRL (AAAI 2016)

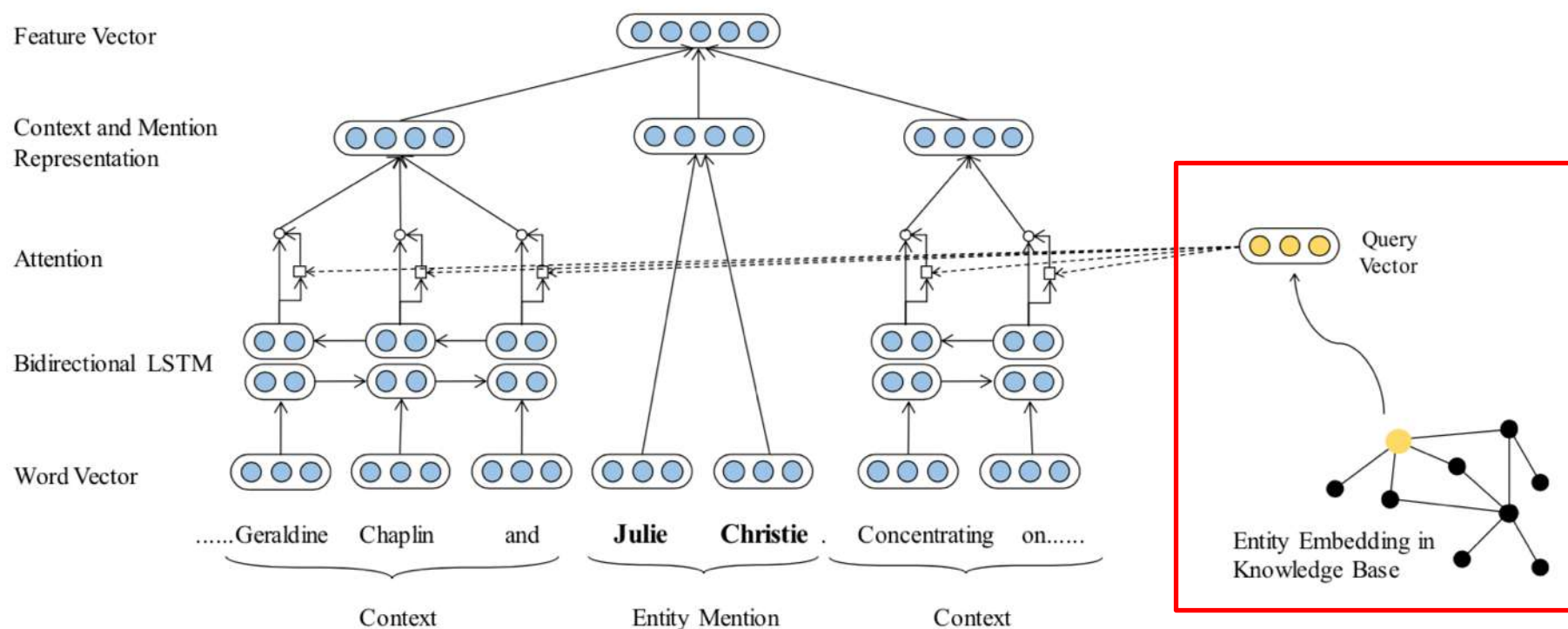


综合考虑实体、属性与关系的知识表示  
KR-EAR (IJCAI 2016)

考虑实体图像信息的知识表示  
IKRL (IJCAI 2017)

# 知识指导的实体细粒度分类

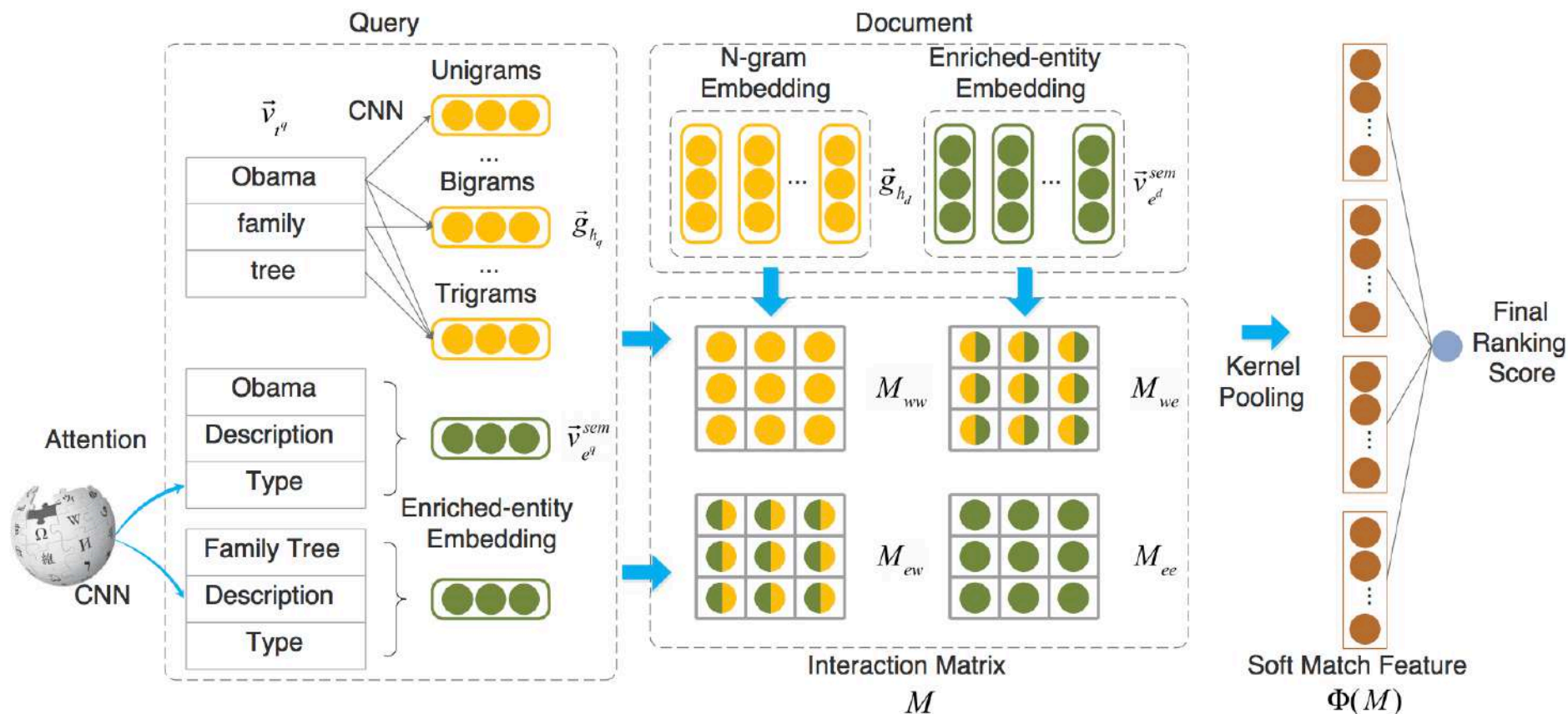
- 对文本实体进行细粒度分类，助力深度分析
- 充分利用KG实体表示，提出知识注意力机制，建立对上下文的高效建模





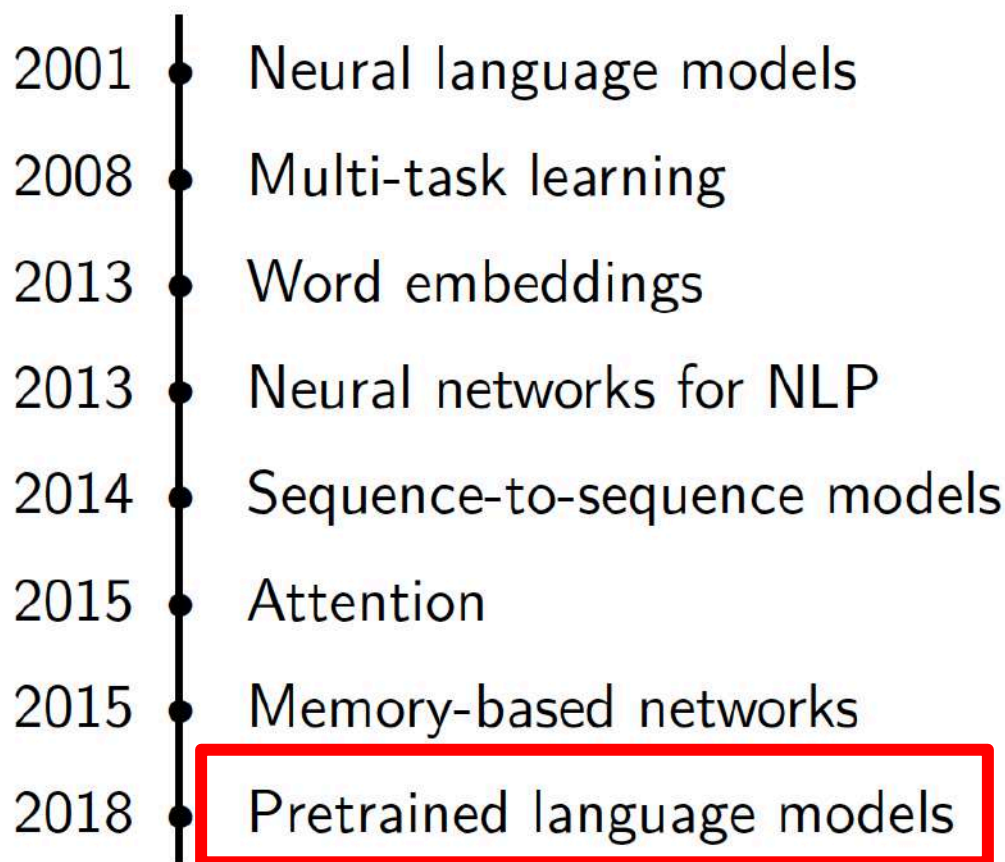
# 知识指导的神经网络文档排序

- 在利用神经网络学习查询-文档匹配关系模型 (KNRM) 中, 引入KG世界知识




# 知识指导的预训练语言模型

- 深度学习对大规模无监督数据建模的最新进展



# 预训练语言模型

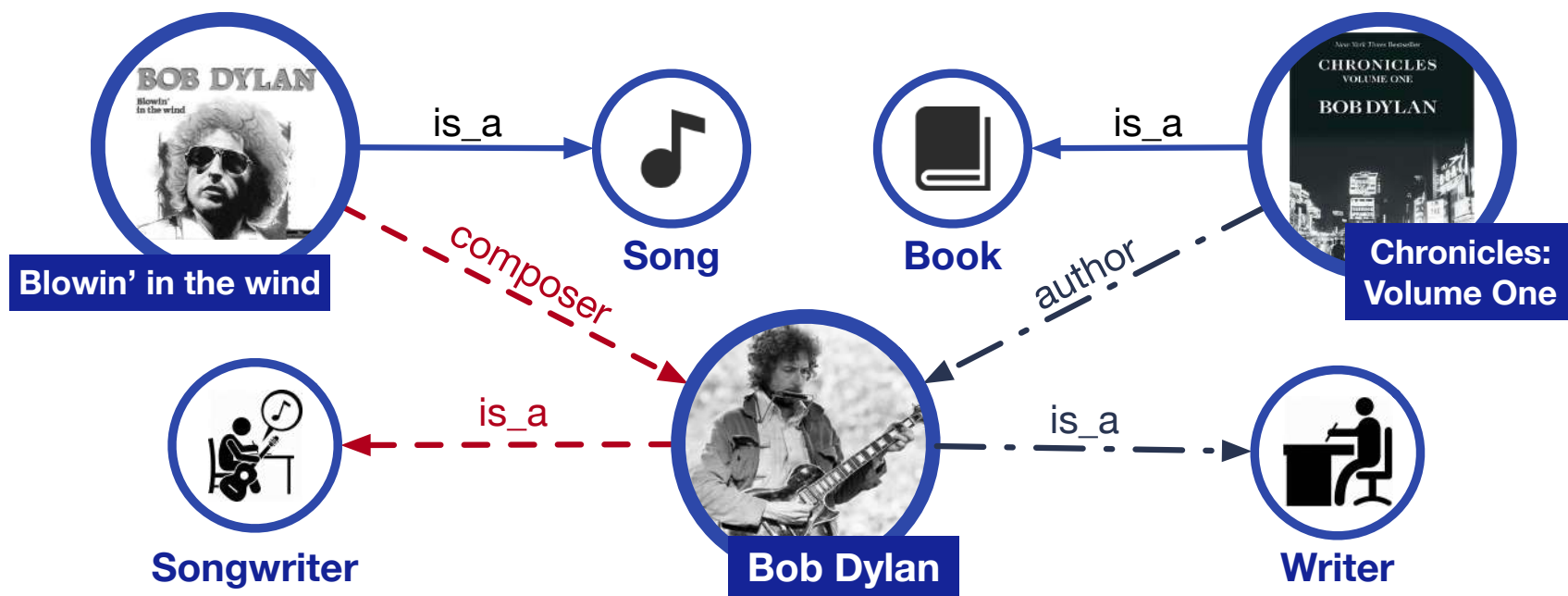
- 主要预训练模型
  - ELMo横扫6项NLP任务
  - GPT刷新9项NLP任务榜单
  - BERT再次刷新11项NLP任务

| Rank | Name           | Model                             | URL   | Score |
|------|----------------|-----------------------------------|---|-------|
| 1    | bigbird he     | Microsoft D365 AI & MSR AI        |   | 81.9  |
| – 2  | Jacob Devlin   | BERT: 24-layers, 1024-hidden, 16  |    | 80.4  |
|      |                | BERT: 12-layers, 768-hidden, 12-l |    | 78.3  |
| 3    | Jason Phang    | GPT on STILTs                     |  | 76.9  |
| 4    | Alec Radford   | Singletask Pretrain Transformer   |  | 72.8  |
| + 5  | Samuel Bowman  | BiLSTM+ELMo+Attn                  |  | 70.5  |
| 6    | GLUE Baselines | BiLSTM+ELMo+Attn                  |  | 68.9  |

Leaderboard of GLUE benchmark (2019.1)

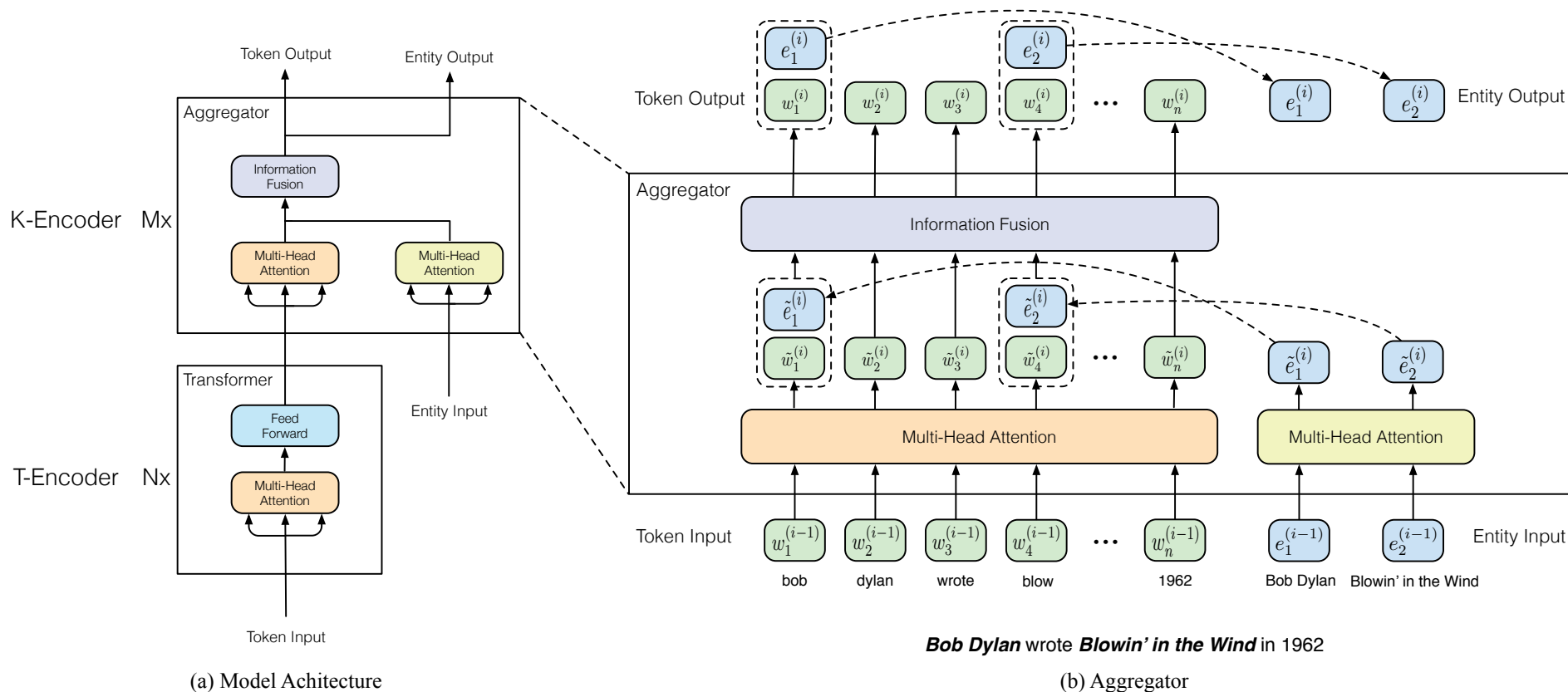
# 知识指导的预训练语言模型

- 预训练模型未考虑知识图谱中的结构化知识
- 结构化知识可以有效的提升模型对于文本中的低频实体的理解能力



**Bob Dylan** wrote **Blowin' in the Wind** in 1962, and wrote **Chronicles: Volume One** in 2004.

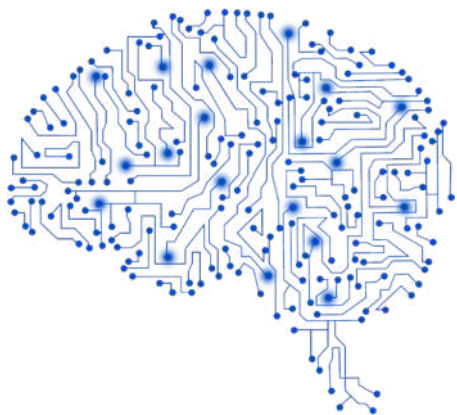
# 知识指导的预训练语言模型



- 左边是模型的总体框架，下面N层是文本编码层，上面M层是增加知识信息的混合编码层，左边是混合编码层的具体结构
- 混合编码层同时输入文本序列和实体序列，对两个序列分别进行自注意力机制
- 按照文本和实体的对应关系，对自注意力层的输出进行组合，输入信息融合层
- 在文本、知识双向融合后，产生下一层的文本和实体输入

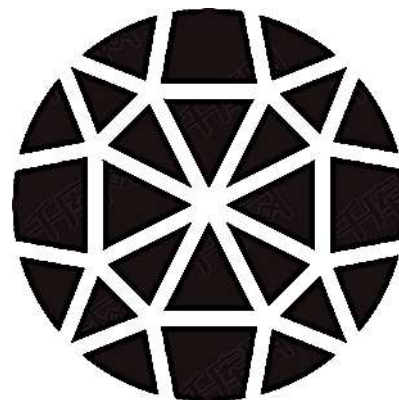
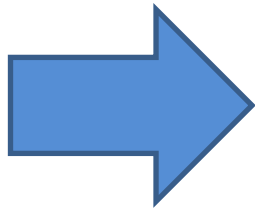
# 世界知识指导NLP相关论文

- Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, Qun Liu. **ERNIE: Enhanced Language Representation with Informative Entities**. ACL 2019.
- Zhenghao Liu, Chenyan Xiong, Maosong Sun, Zhiyuan Liu. **Entity-Duet Neural Ranking: Understanding the Role of Knowledge Graph Semantics in Neural Information Retrieval**. ACL 2018.
- Ji Xin, Yankai Lin, Zhiyuan Liu, Maosong Sun. **Improving Neural Fine-Grained Entity Typing with Knowledge Attention**. AAAI 2018.
- Hao Zhu, Ruobing Xie, Zhiyuan Liu, Maosong Sun. **Iterative Entity Alignment via Joint Knowledge Embeddings**. IJCAI 2017.
- Yankai Lin, Zhiyuan Liu, Maosong Sun. **Knowledge Representation Learning with Entities, Attributes and Relations**. IJCAI 2016.



数据驱动的  
深度学习

知识获取

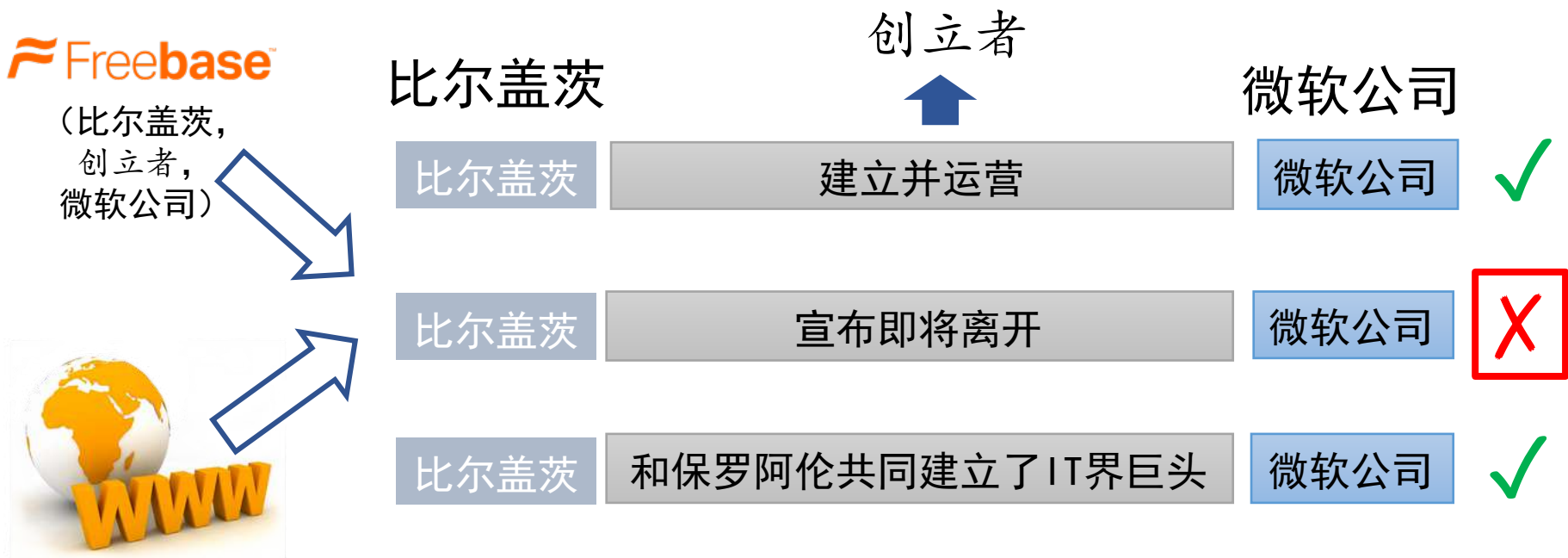


符号表示的  
世界知识



# 知识获取

- 基于已有知识和海量文本信息获取结构化知识
- 解决标注数据噪音，融合多源信息



# 神经网络知识获取技术

- 采用神经网络对句子进行语义理解
- 使用大规模自动标注训练数据学习

Freebase

(Bill Gates,  
Founder,  
Microsoft)

Bill Gates

Founder

Microsoft

Bill Gates

was the co-founder and CEO of

Microsoft

Bill Gates

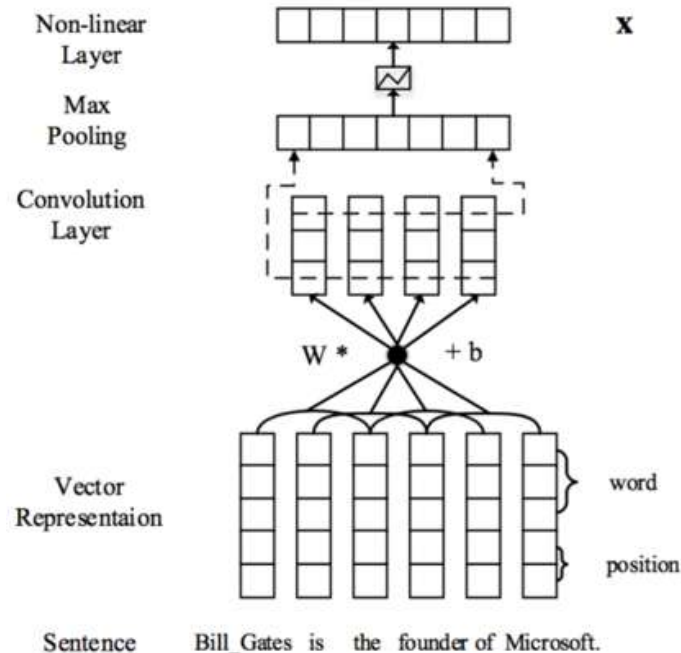
announced to retire from

Microsoft

Bill Gates

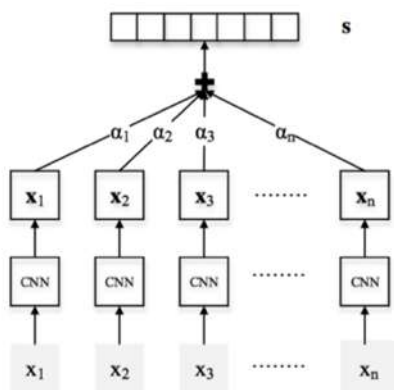
and Paul Allen co-founded the IT giant

Microsoft

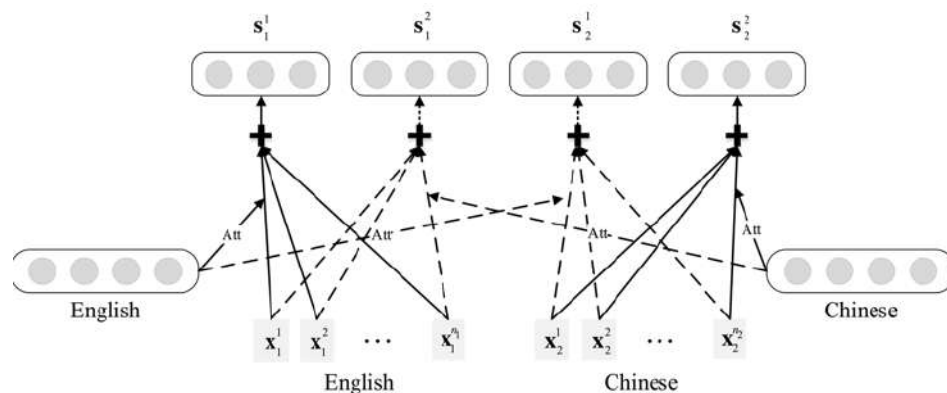


# 高效鲁棒的知识获取技术

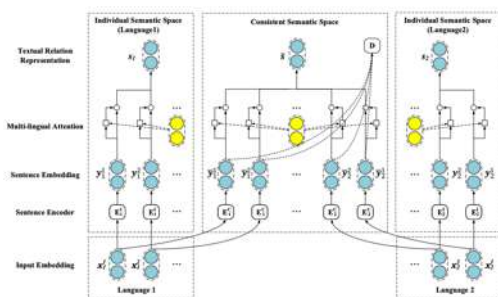
- 提出**选择注意力**机制自动降噪并整合多源信息



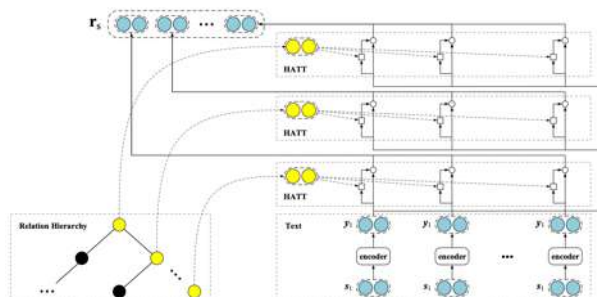
基于**句级注意力**的远程监督  
神经网络关系抽取(ACL 2016)



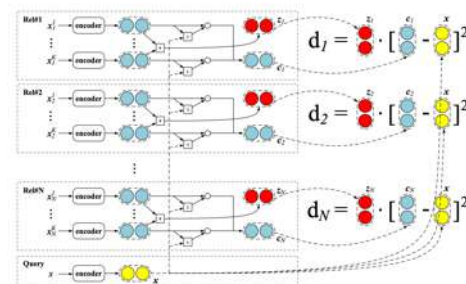
基于**跨语言注意力**的  
神经网络关系抽取(ACL 2017)



基于**对抗注意力**的神经网络  
关系抽取(COLING 2018)



基于**层次注意力**的神经网络  
关系抽取(EMNLP 2018)



基于**混合注意力**的  
少次关系抽取(AAAI 2019)

# 开源工具

- 义原计算、知识表示、知识获取等相关算法工具均在全球最大开源社区GitHub发布，获得超过**17000+**星标关注

<https://github.com/thunlp>

- THULAC : 中文词法分析
- THUCTC : 中文文本分类
- THUTAG : 关键词抽取与社会标签推荐
- OpenKE : 知识表示学习
- OpenNRE : 神经网络关系抽取
- OpenNE : 网络表示学习
- OpenQA : 开放域自动问答



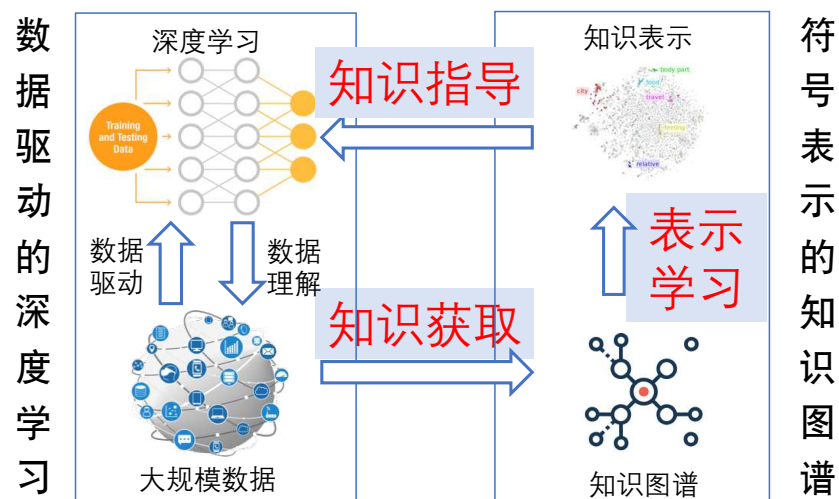
The screenshot displays the GitHub profile for the 'thunlp' repository. It includes the repository's logo, which is a circular seal of Tsinghua University with 'NLP' and 'Natural Language Processing' text. Below the logo, the repository name 'thunlp' is shown in a blue font, followed by location tags for 'Beijing' and 'China'. There are links to 'Tweet your ranking' and a 'Refresh your profile' button. To the right, two ranking tables are visible: one for 'c++ ranking' and one for 'python ranking'. Each table lists rankings for 'Beijing', 'China', and 'Worldwide', along with the number of repositories ('Repos') and stars.

| Ranking Type | Location       | Rank    | Count   |
|--------------|----------------|---------|---------|
| c++ ranking  | Beijing        | 12      | 2 413   |
|              | China          | 30      | 9 212   |
|              | Worldwide      | 519     | 251 037 |
|              | Repos :        | 11      |         |
|              | Stars :        | 822     |         |
|              | python ranking | Beijing | 33      |
|              | China          | 91      | 12 113  |
|              | Worldwide      | 2 045   | 419 419 |
|              | Repos :        | 6       |         |
|              | Stars :        | 529     |         |

# 总结展望

- 义原语言知识突破词汇屏障，对语言理解极具重要意义，具有极佳融合深度学习的特性
- 世界知识对于富知识文本深度理解具有重要意义，知识表示学习是目前较好的解决方案
- 深度学习自然语言处理技术反过来可以帮助从大规模文本中获取知识

## 五个更加



1. 更加全面的知识类型
2. 更加复杂的知识结构
3. 更加有效的知识获取
4. 更加强大的知识指导
5. 更加精深的知识推理

# 感谢各位!

<http://nlp.csai.tsinghua.edu.cn/~lzy/>  
[liuzy@Tsinghua.edu.cn](mailto:liuzy@Tsinghua.edu.cn)