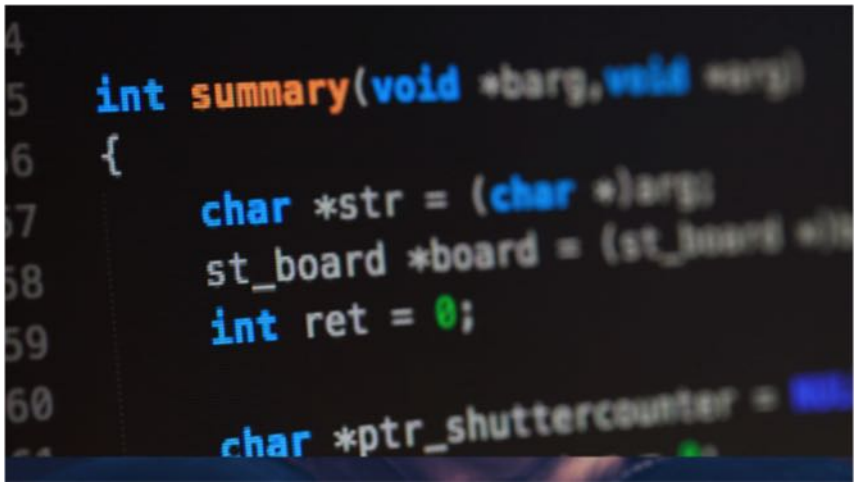
知识指导的自然语言处理

自然语言处理实验室

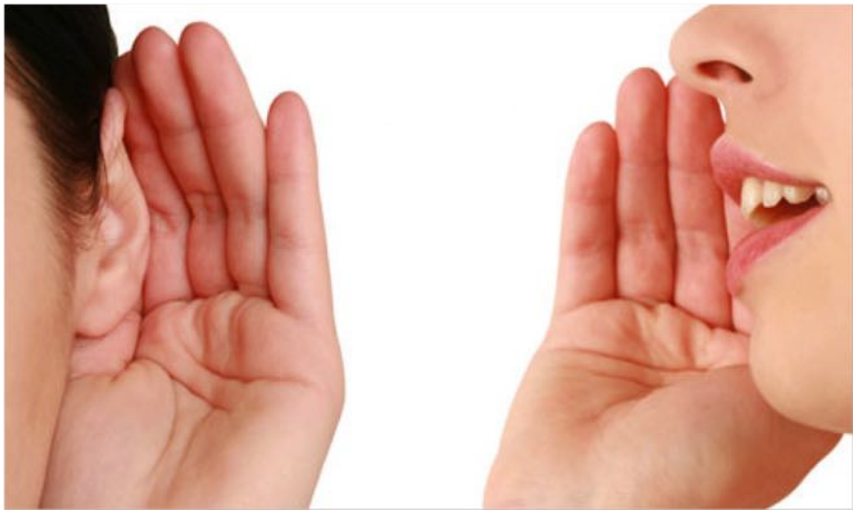
刘知远

自然语言

- 自然语言是人类间交流传播信息和知识的工具



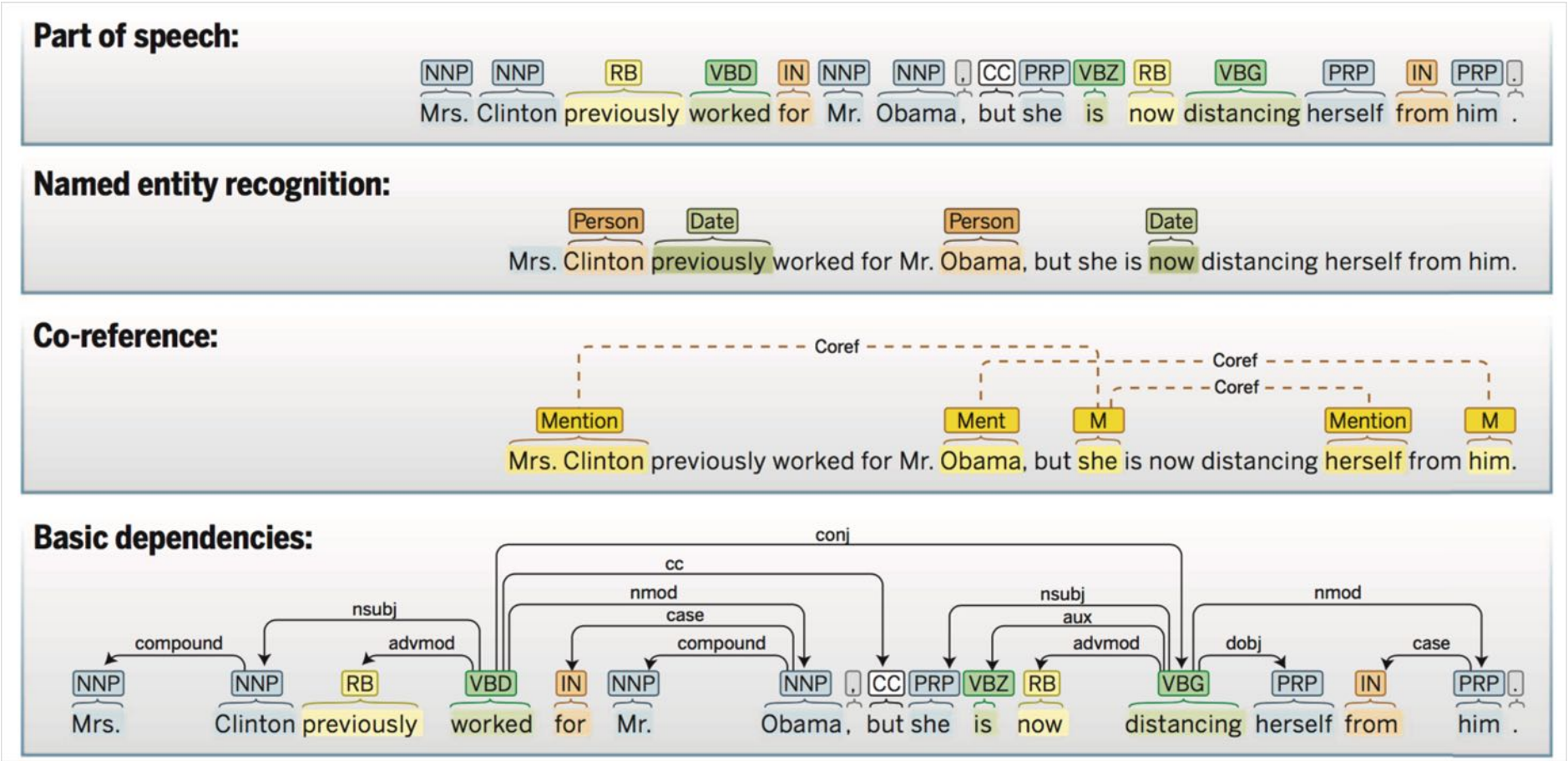
编程语言



自然语言

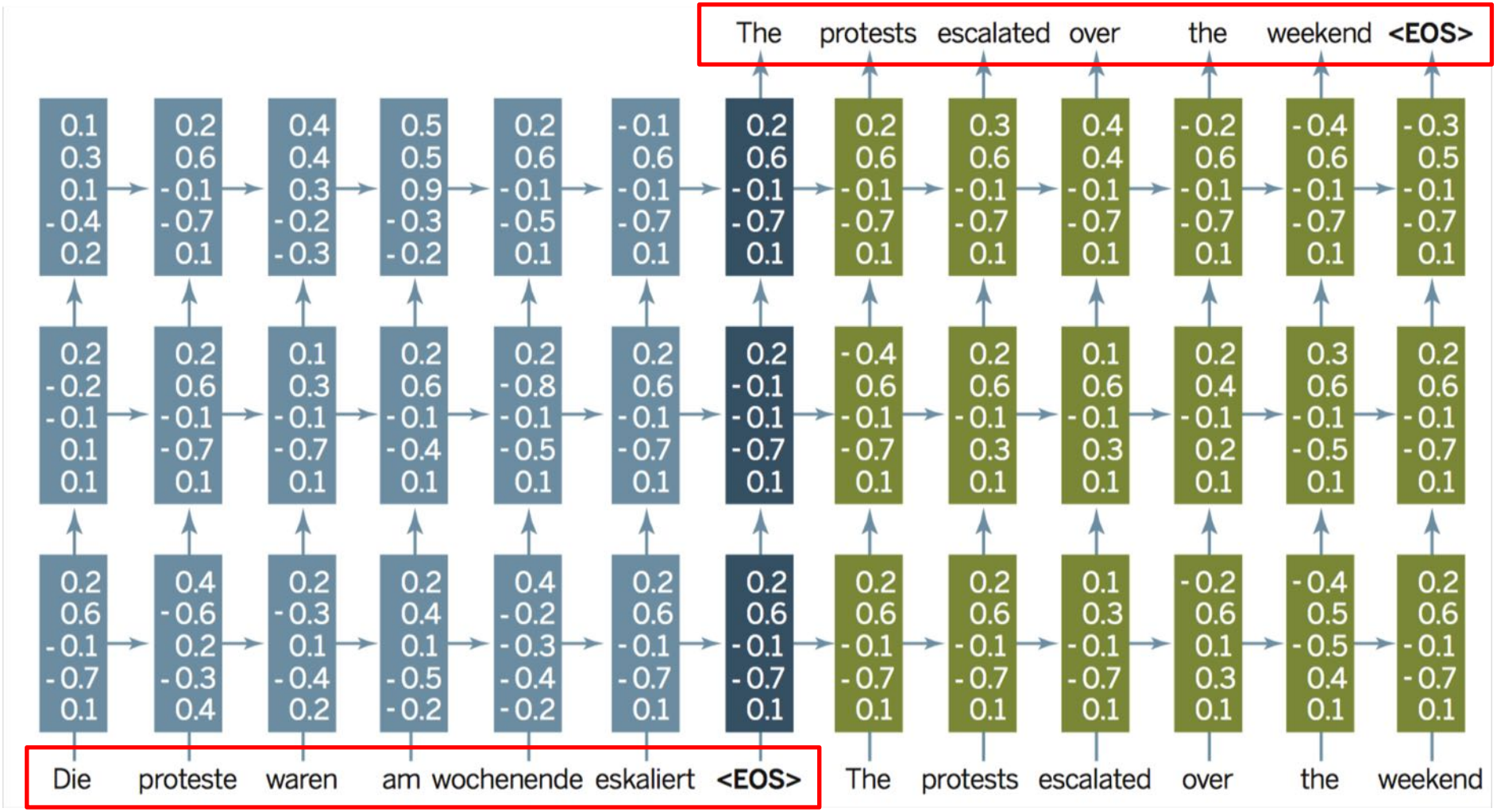
自然语言处理

- 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



Advances in Natural Language Processing. Science 2015.

数据驱动的自然语言处理：深度学习



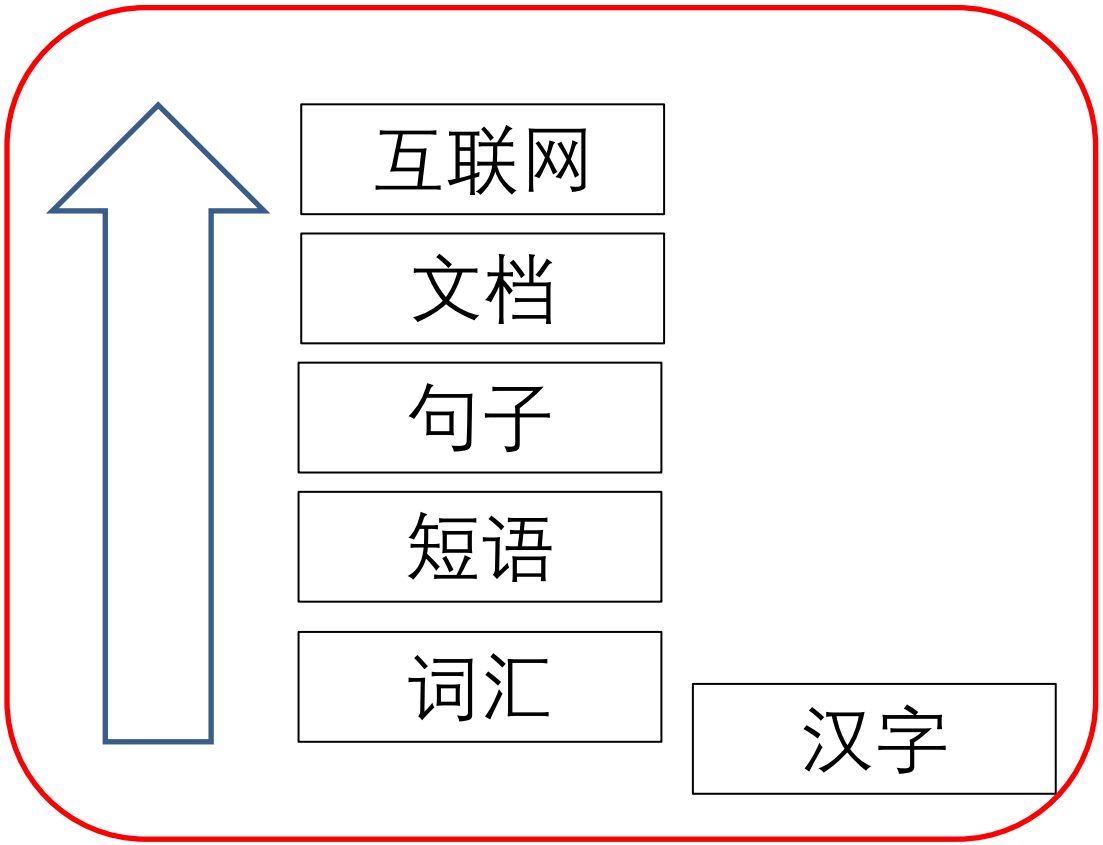
Advances in Natural Language Processing. Science 2015.

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

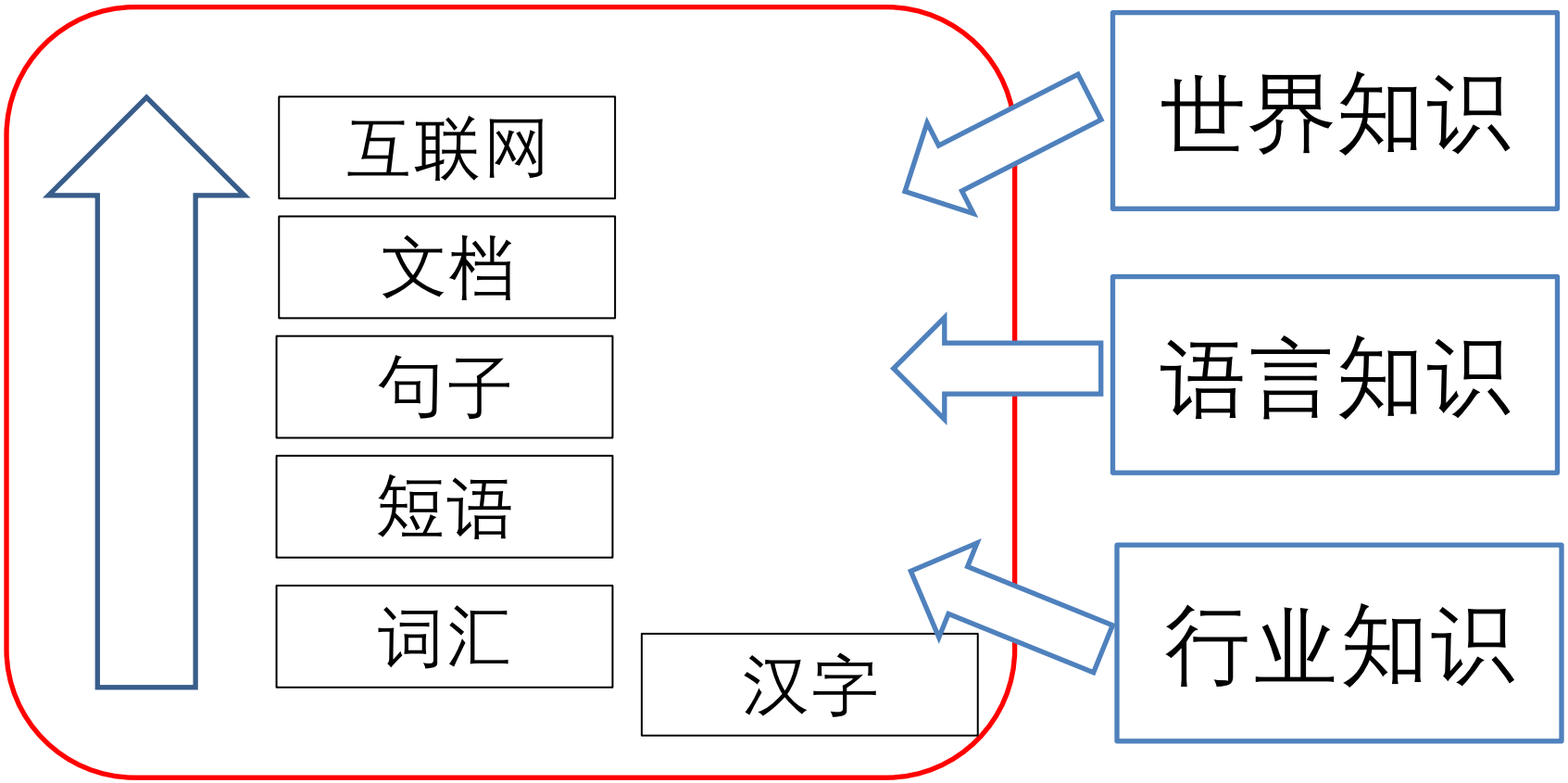
- 自然语言包含从汉字到文档的多粒度语言单位



数据驱动

自然语言特点

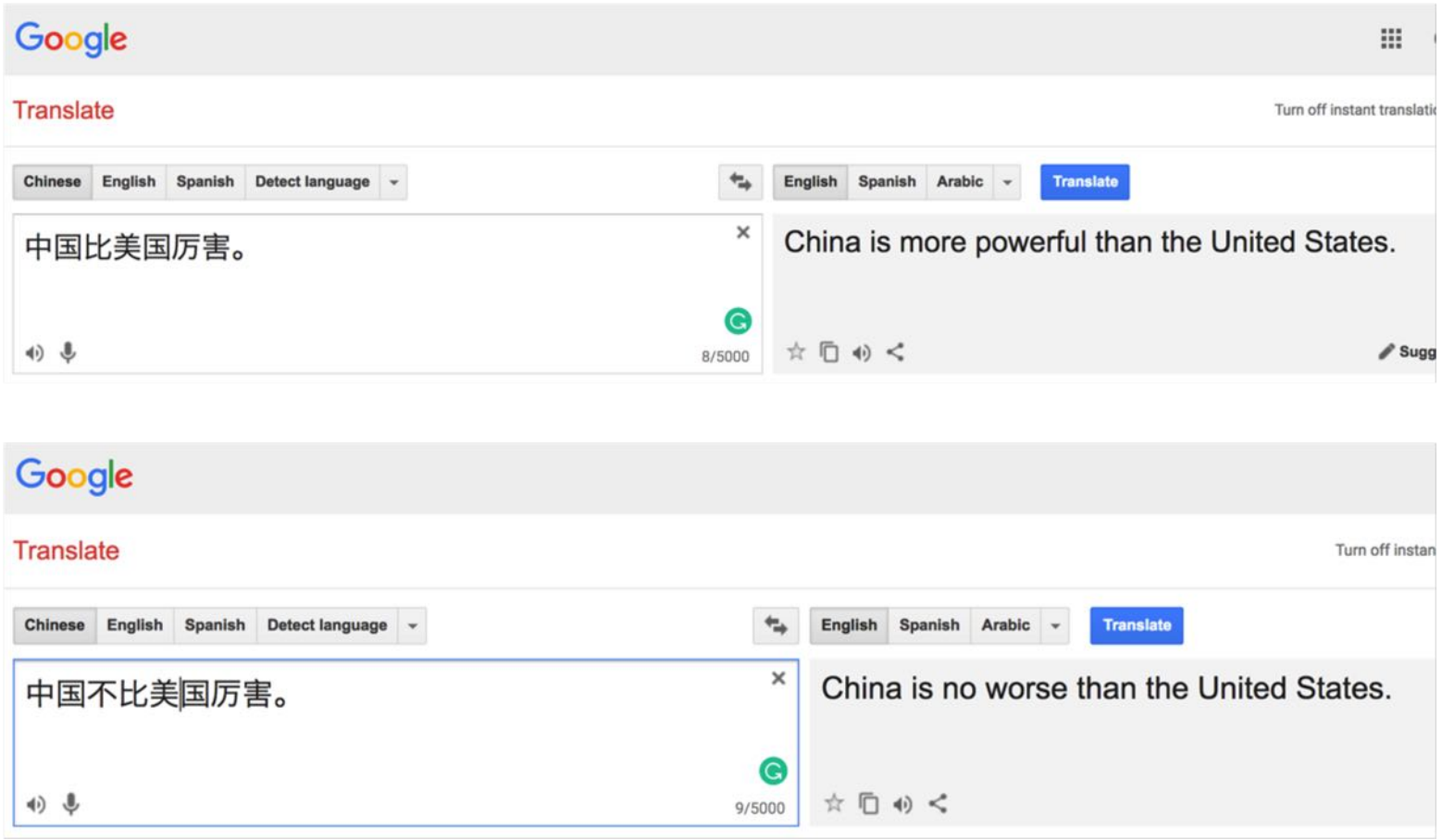
- 自然语言文本蕴含丰富的语言知识和世界知识



数据驱动 + 知识指导

数据驱动+知识指导：意义

- 缓解数据驱动方法的鲁棒性和可解释性问题

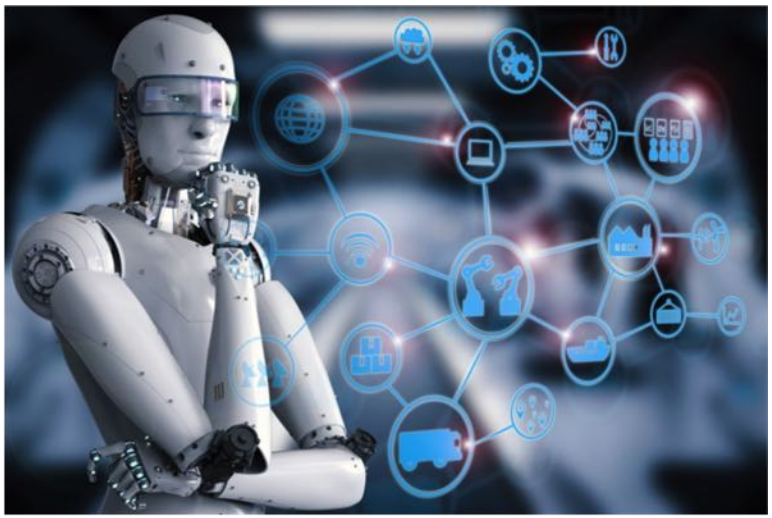


数据驱动+知识指导：意义

- 实现自然语言处理模型从感知到认知的飞跃

智能感知

智能认知

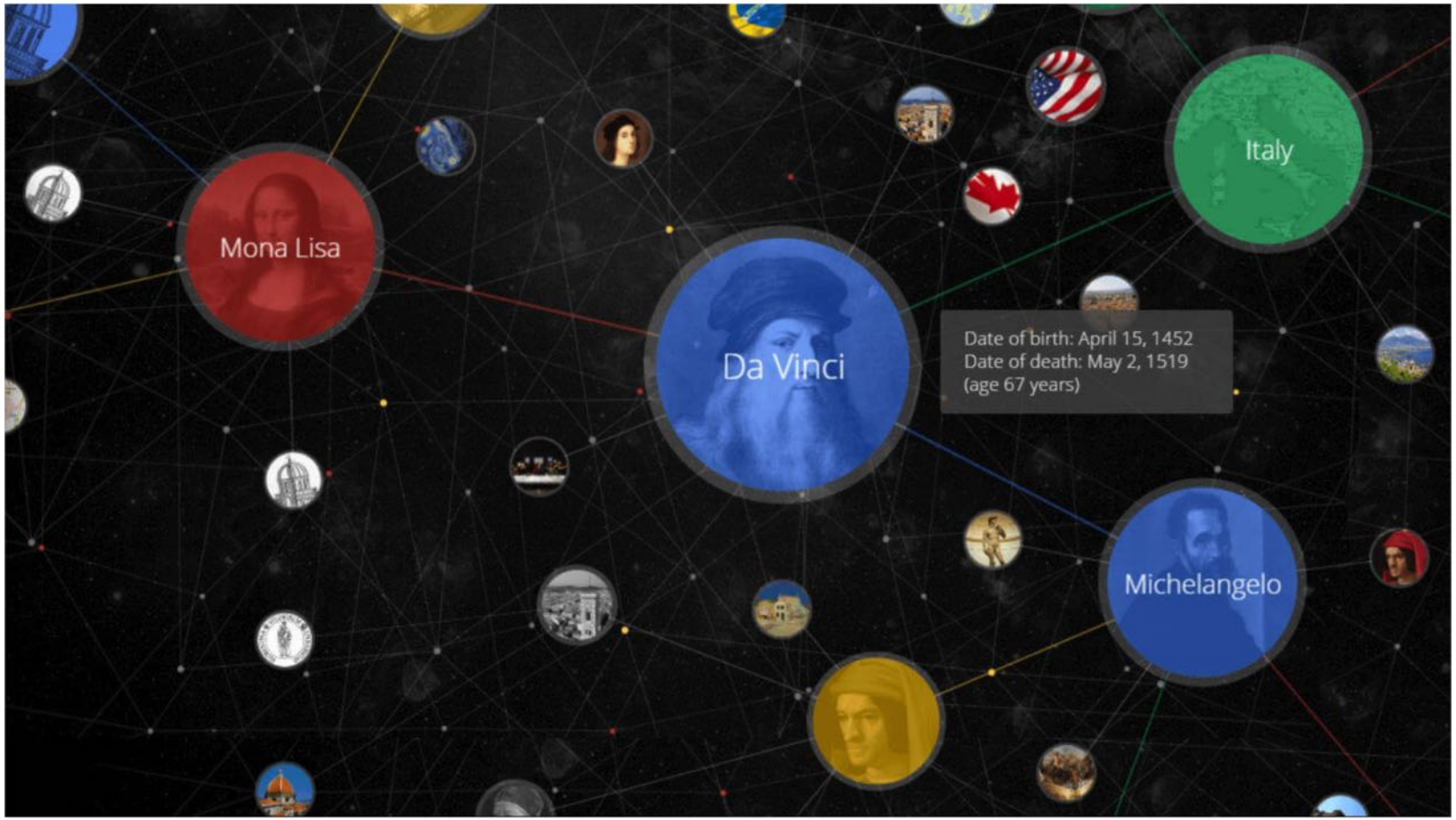


数据：耳聪目明

知识：深思熟虑

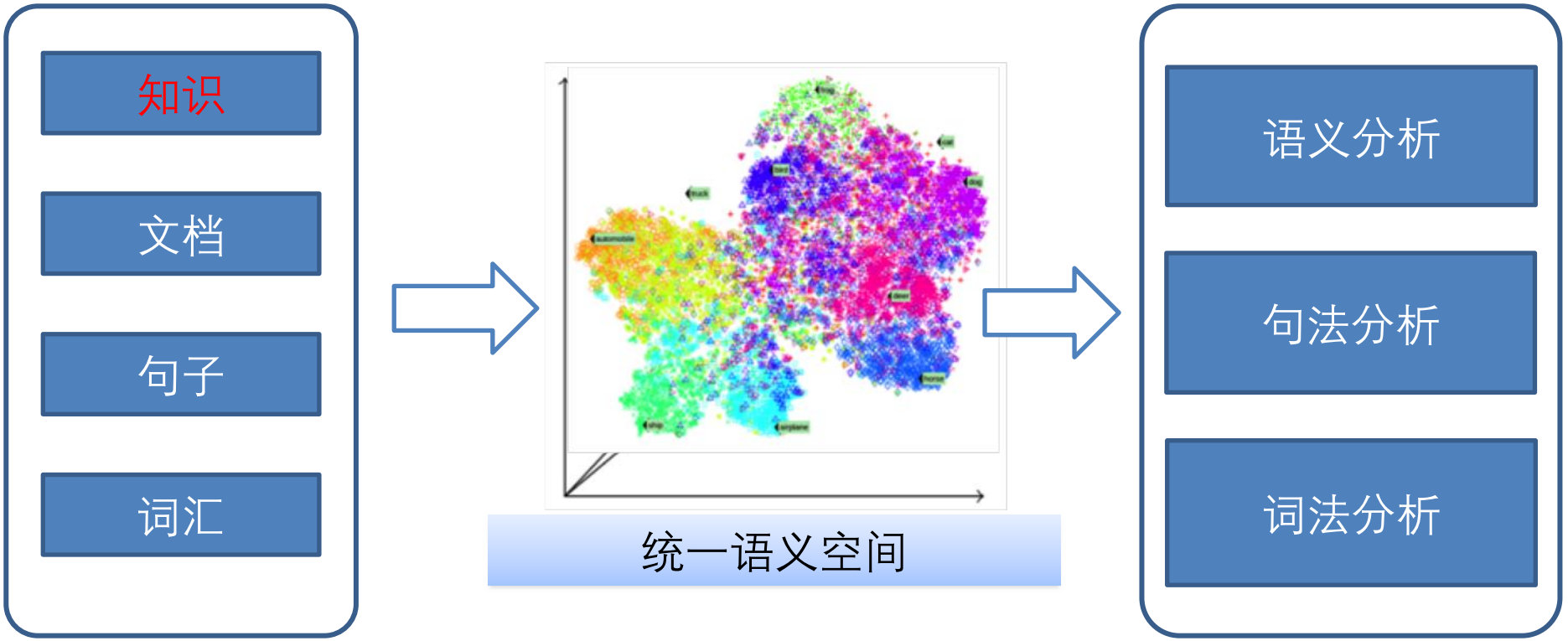
知识指导+数据驱动：难点

- 语言知识、世界知识均通过离散符号表示

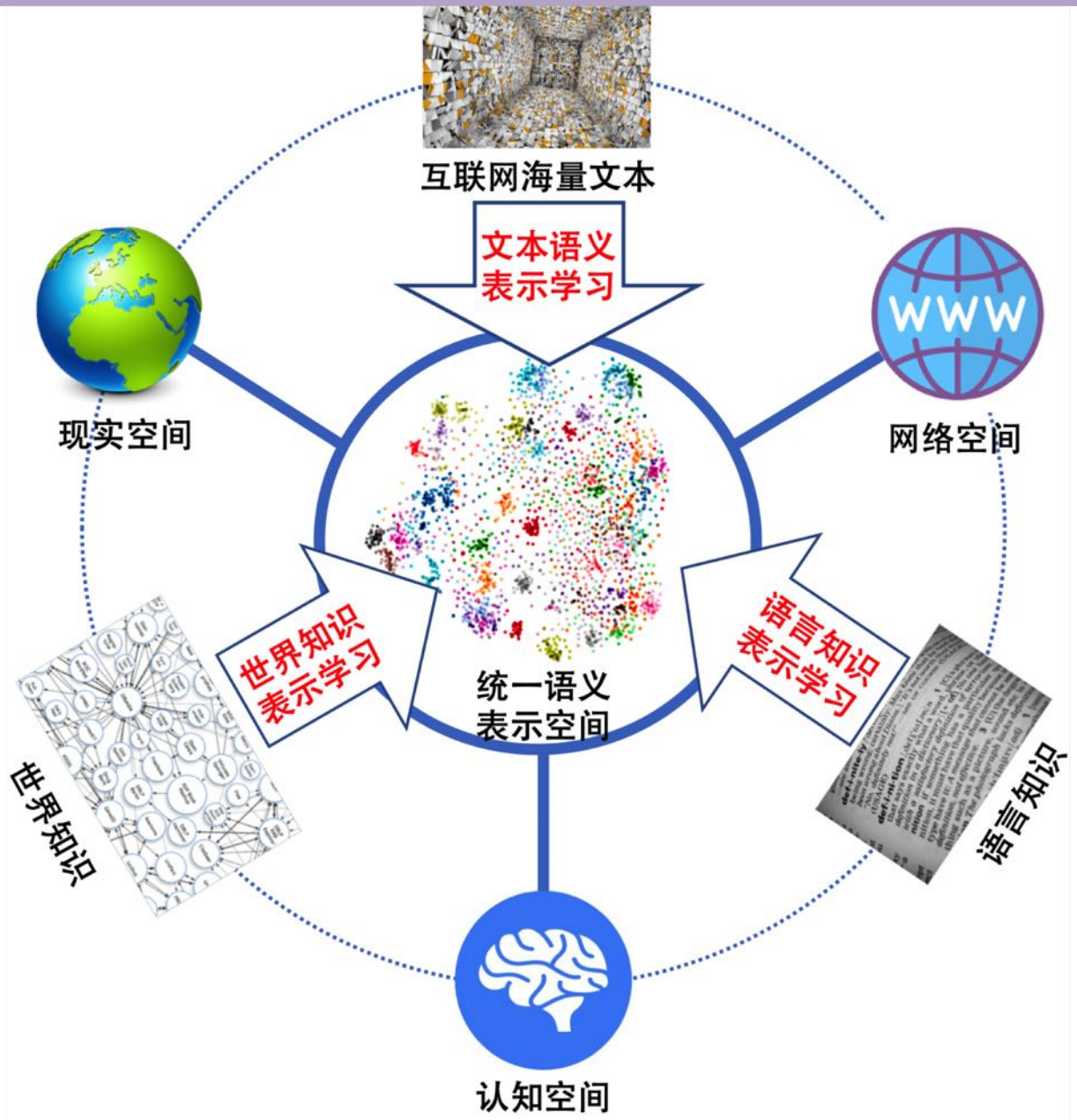


知识指导+数据驱动：实现

- 分布式表示：实现跨粒度、跨领域、富知识的语言理解



知识指导+数据驱动：实现


互联网海量文本
文本语义
表示学习
现实空间
网络空间
WWW
世界知识
表示学习
语言知识
表示学习
统一语义
表示空间
世界知识
语言知识
认知空间

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[1]

Word Similarity Computing Based on How-net

刘群[*] 、李素建[+]

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：

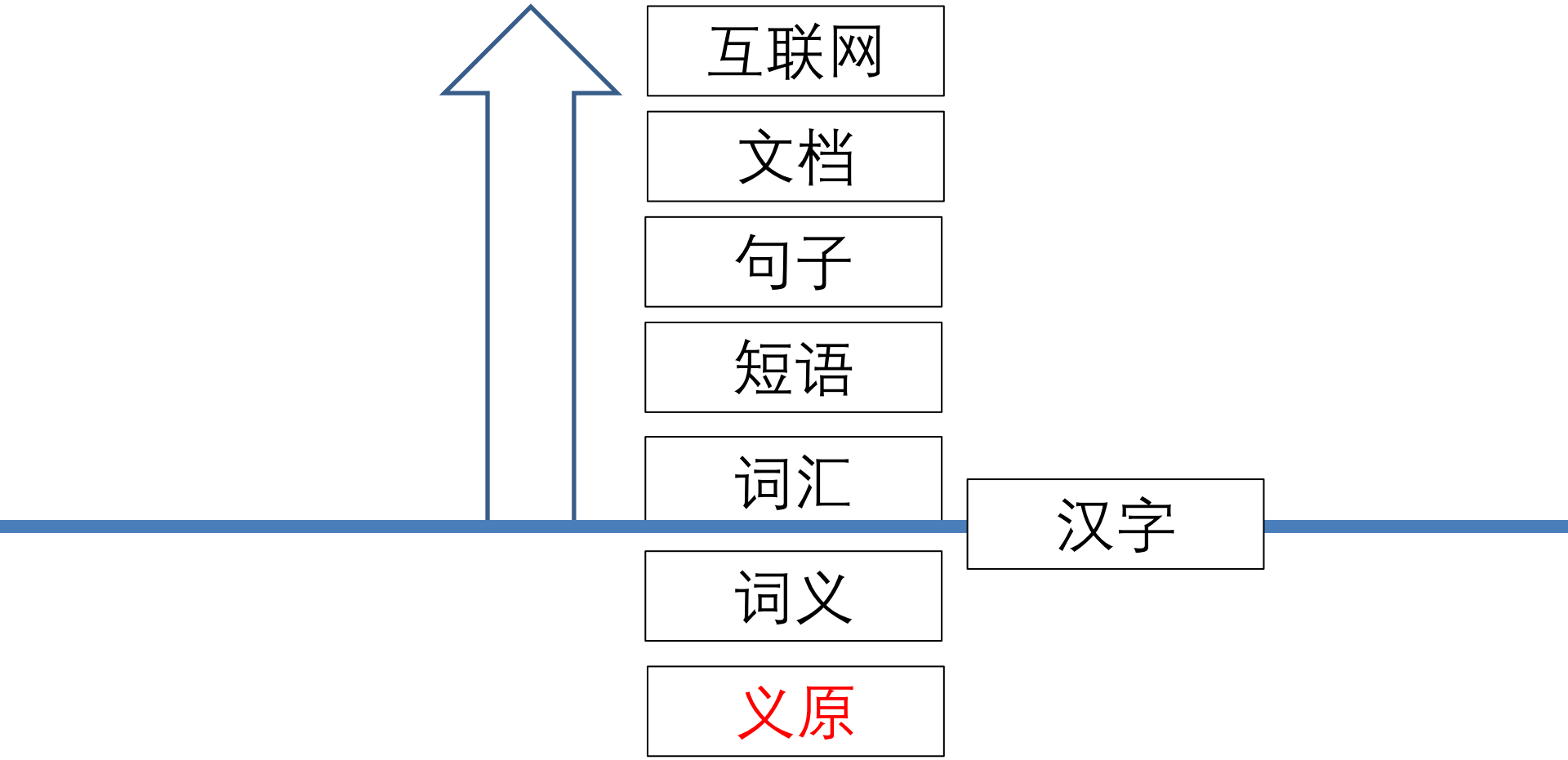
1. 每一个词的语义描述由多个义原组成；
2. 词语的语义描述中各个义原并不是平等的，它们之间有着复杂的关系，通过一种专门的知识描述语言来表示。

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

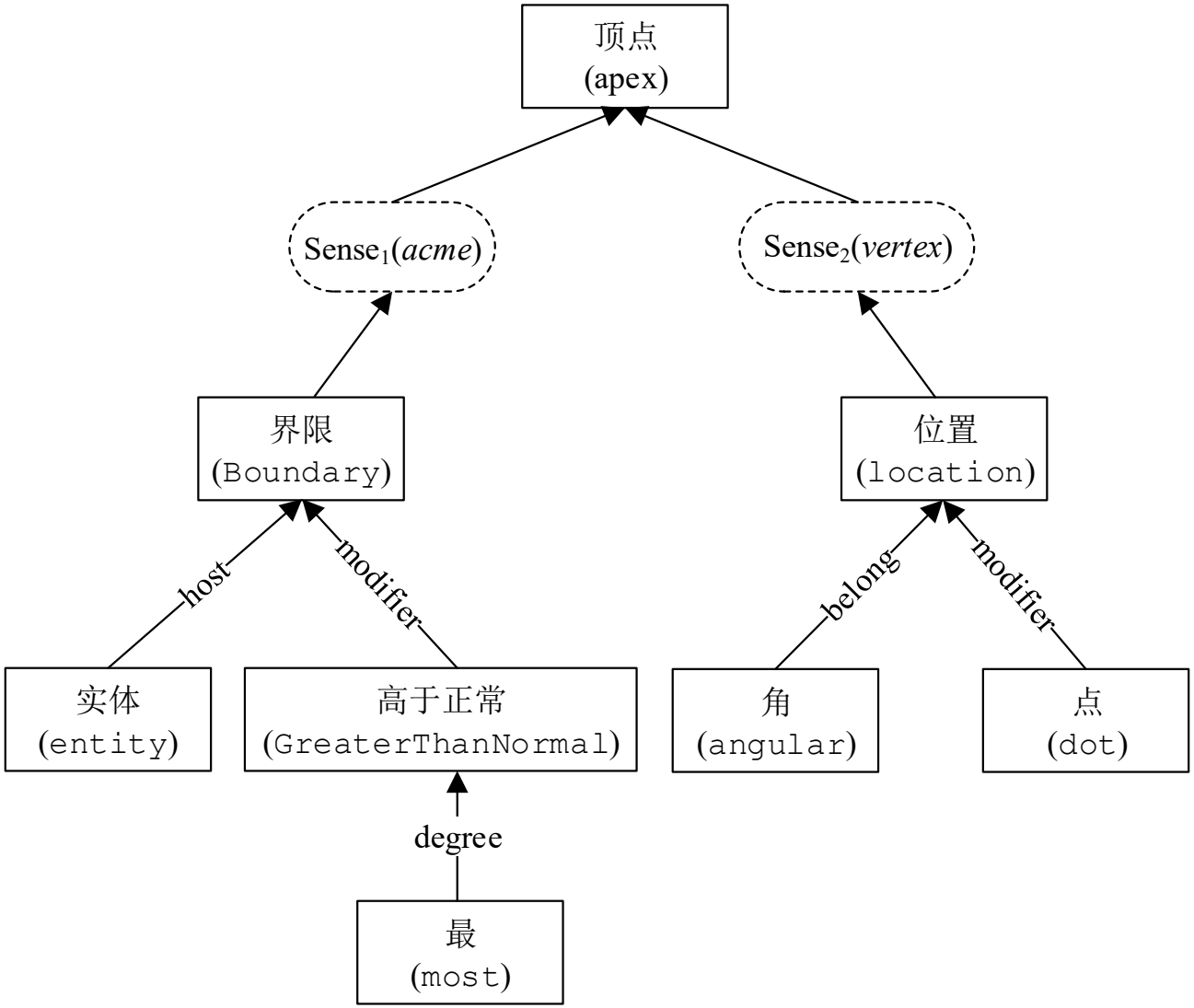
HowNet—瞥

- 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={GreaterThanNormal|高于正常:degree={most|最}}}
> 顶点#2
> DEF={location|位置:belong={angular|角},modifier={dot|点}}

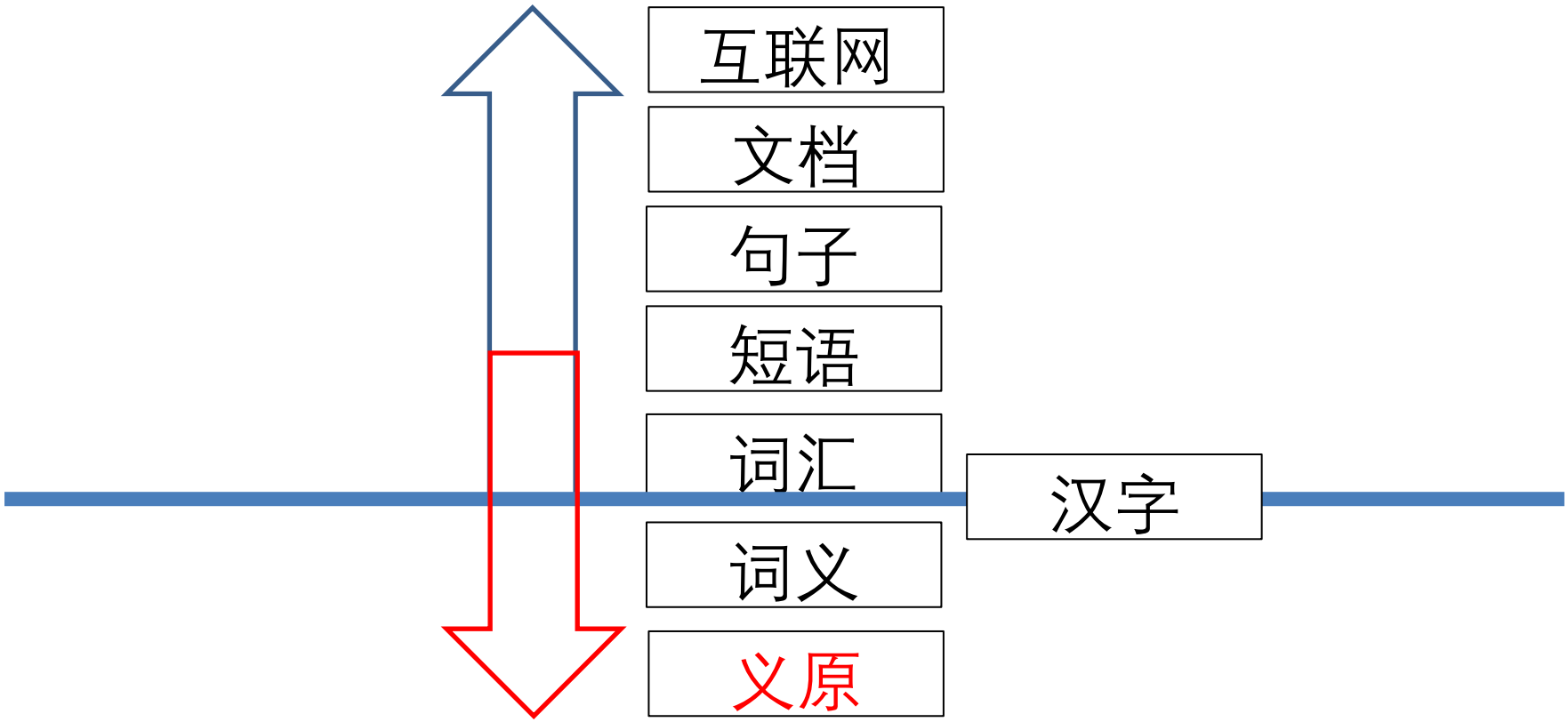
HowNet—瞥

- 义原知识带有层次结构



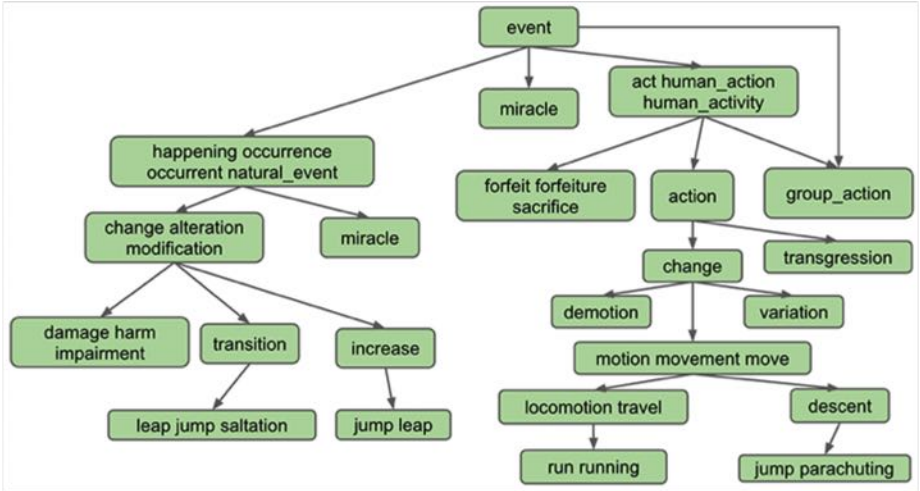
深度学习时代HowNet的意义

- **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

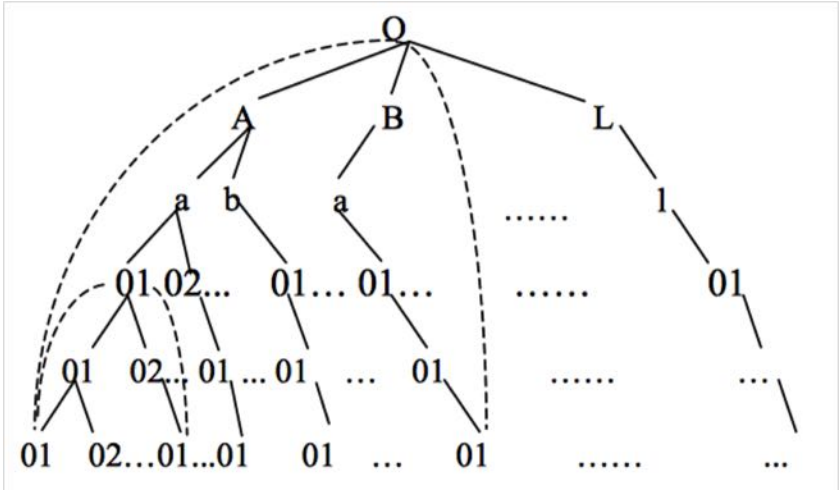


深度学习时代HowNet的意义

- **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

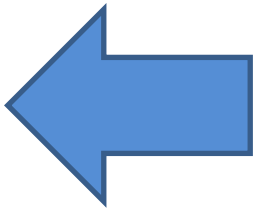


WordNet Synset体系



同义词词林层次类别体系

知识指导

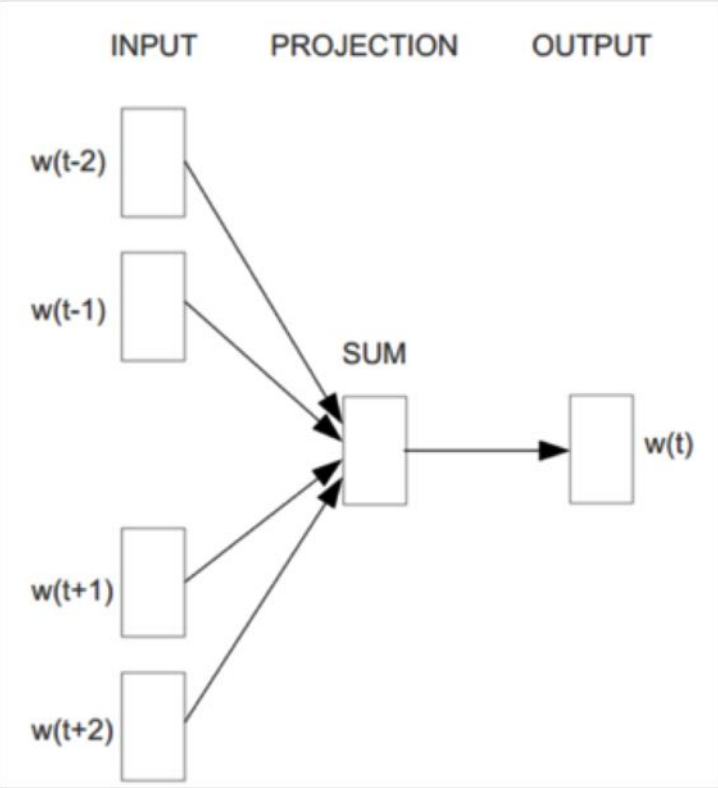


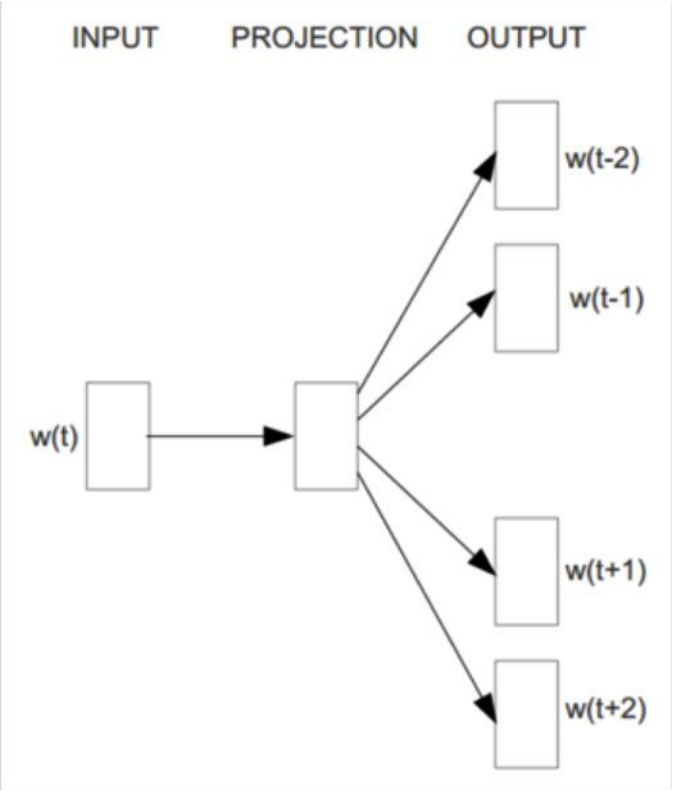
数据驱动的
深度学习

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



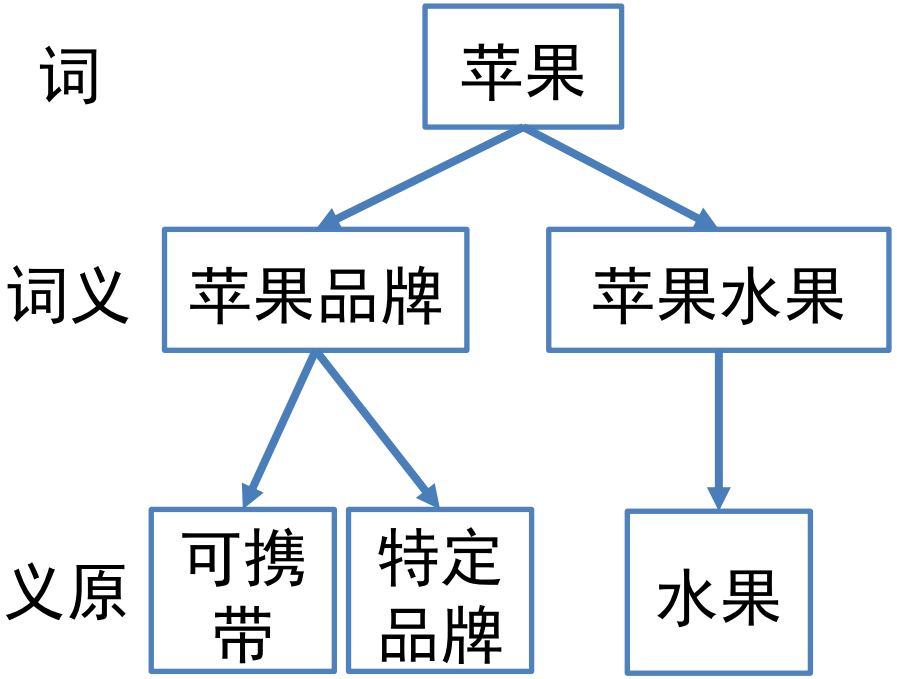


word2vec

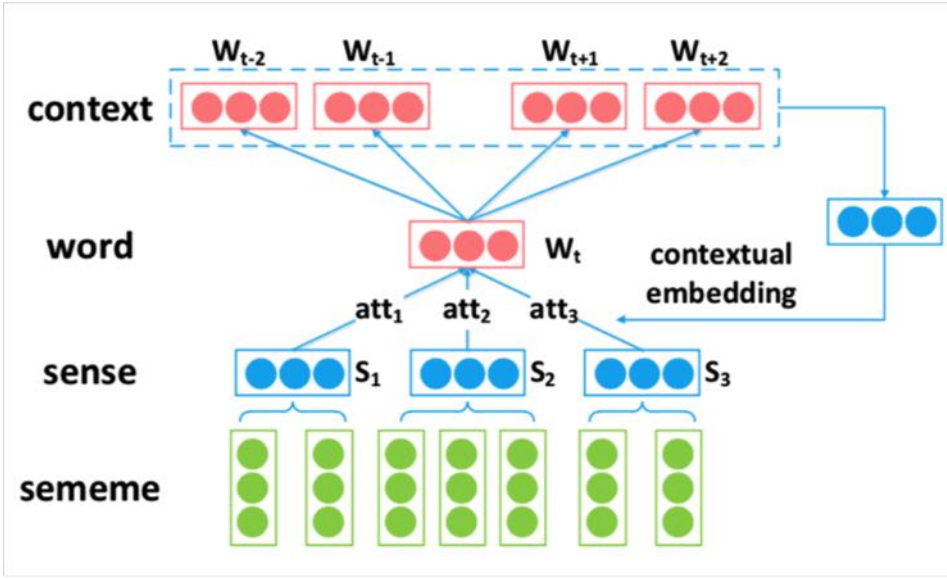
Tomas Mikolov et al. Distributed representations of words and phrases and their compositionality. NIPS 2013.

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

Yilin Niu, Ruobing Xie, Zhiyuan Liu, Maosong Sun. Improved Word Representation Learning with Sememes. ACL 2017.

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	**86.0**	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	**82.6**	**98.9**	80.1	**84.5**	**14.78**	**1.01**	**1.72**	**9.48**

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
苹果素有果中王美称	苹果品牌：0.28	苹果水果：0.72
苹果电脑无法正常启动	苹果品牌：0.87	苹果水果：0.13
八支**队伍**进入第二阶段团体赛	团体：0.90	部队：0.10
公安基层**队伍**组织建设	团体：0.15	部队：0.85

根据上下文消歧结果示例

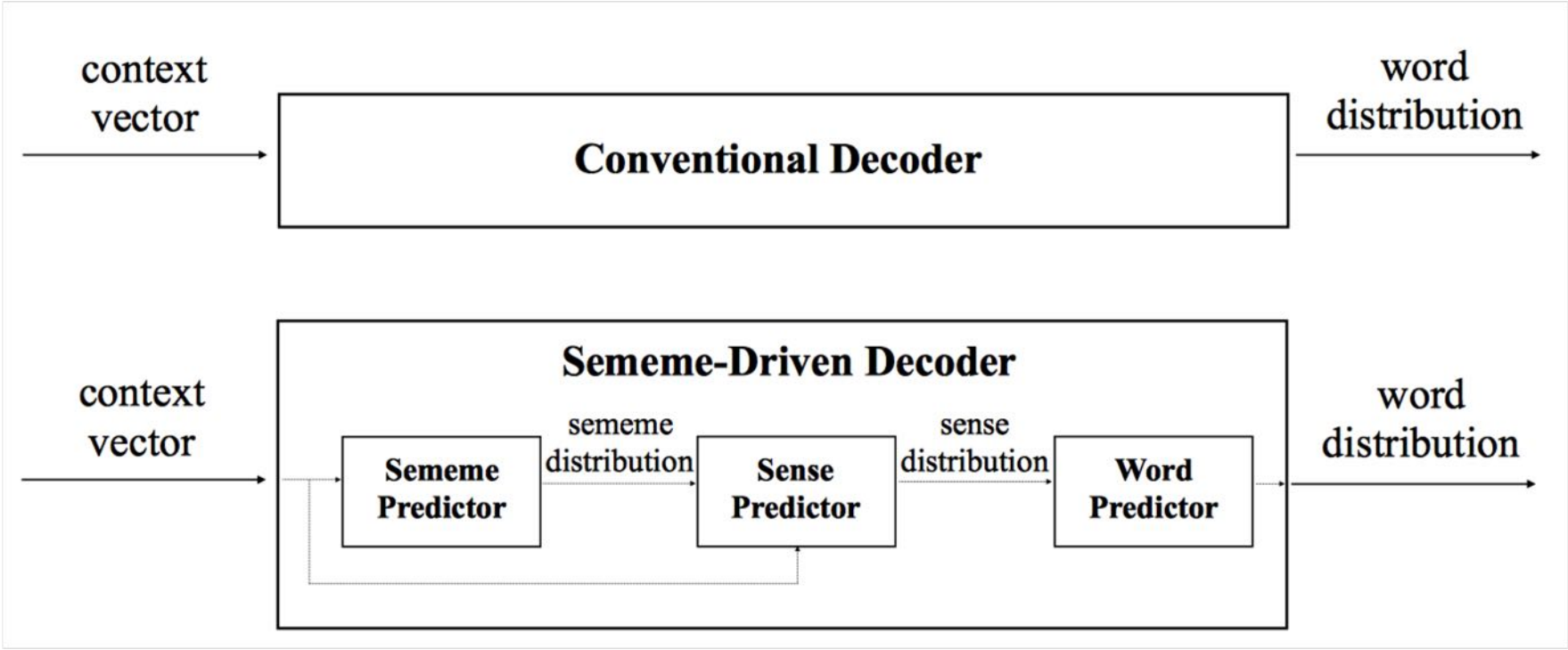
神经语言模型

- 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

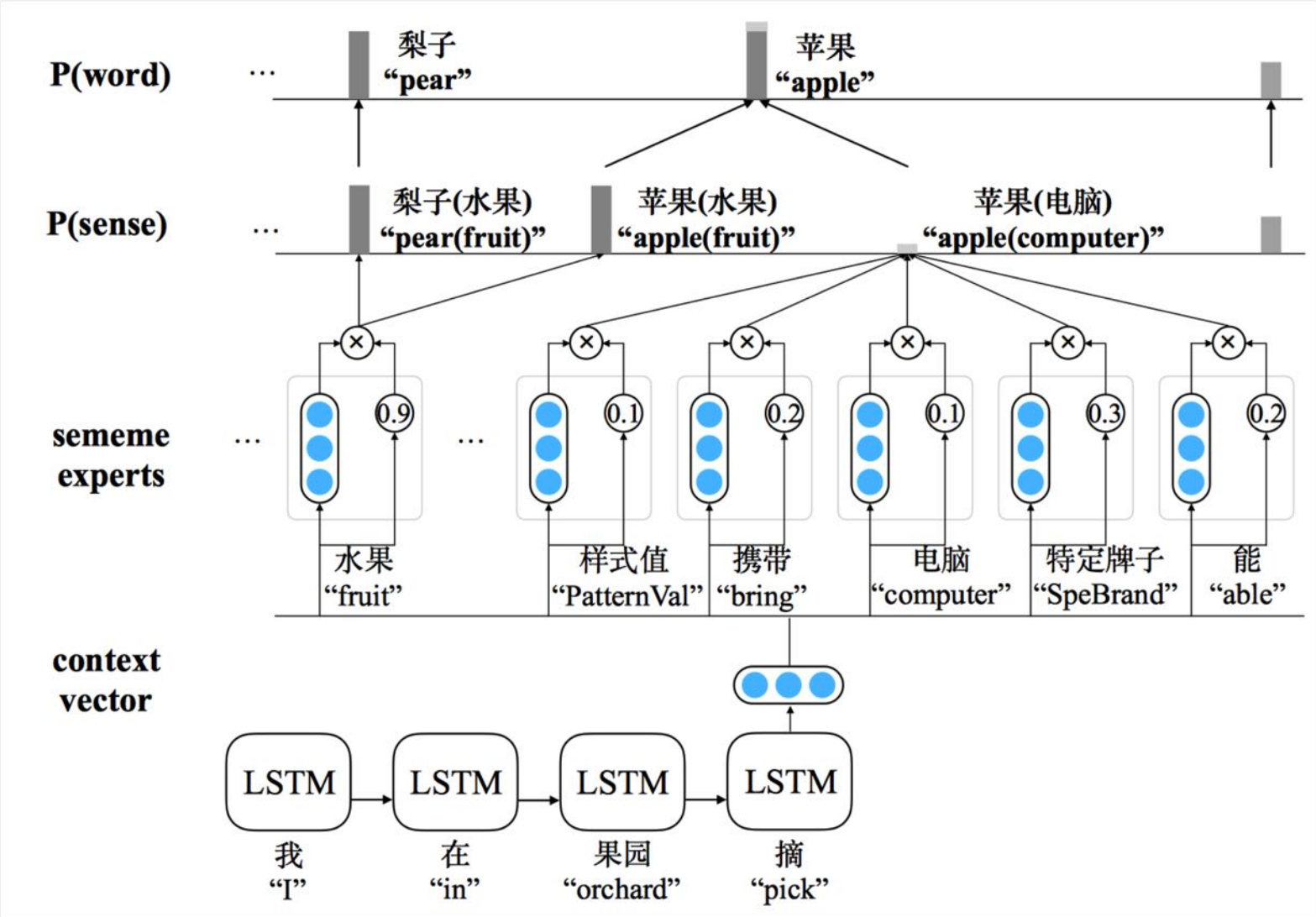
The U.S. trade deficit last year is initially estimated to be 40 billion _____ .

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



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.

实验结果

- 义原驱动的神经语言模型（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	**97.75**	**97.32**
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	**94.24**	**93.60**
AWD-LSTM[4]	26M	89.35	88.86
+ MoS	26M	92.98	92.76
+ SDLM	27M	**88.16**	**87.66**

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.

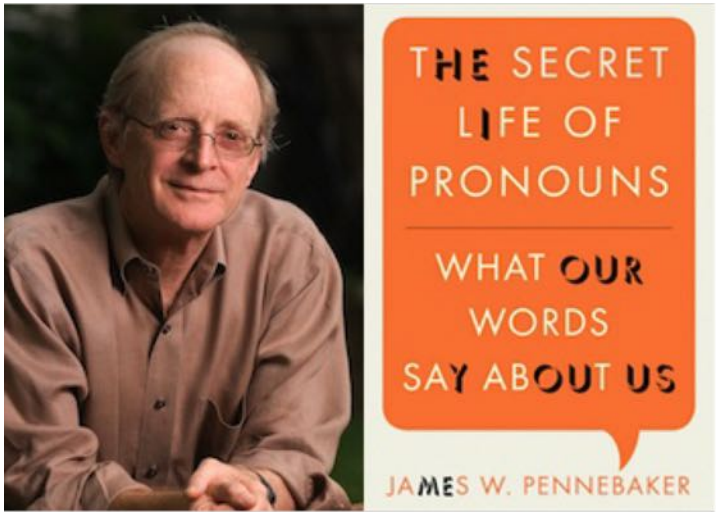
实验结果

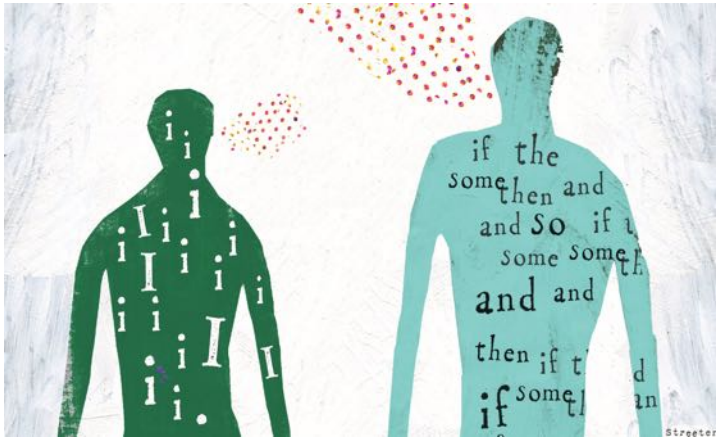
Example (1)		
去年 美国 贸易逆差 初步 估计 为 <N> _____ 。		
The U.S. trade deficit last year is initially estimated to be <N> _____ .		
Top 5 word prediction		
美元 “dollar”	，“,”	。“.”
日元 “yen”	和 “and”	
Top 5 sememe prediction		
商业 “commerce”	**金融 “finance”**	**单位 “unit”**
多少 “amount”	专 “proper name”	
Example (2)		
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”	

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.

融合HowNet义原标注的词典扩展

- 以中文LIWC (Linguistic Inquiry and Word Count)为例，是计算社会科学中的著名词典





LIWC Results

Details of Writer: 40 year old Female
Date/Time: 6 January 2014, 1:02 am

LIWC results from input text

LIWC categories

LIWC Dimension	Your Data	Personal Texts	Formal Texts
Self-references (I, me, my)	8.33	11.4	4.2
Social words	4.17	9.5	8.0
Positive emotions	2.08	2.7	2.6
Negative emotions	1.04	2.6	1.6
Overall cognitive words	3.12	7.8	5.4
Articles (a, an, the)	2.08	5.0	7.2
Big words (> 6 letters)	20.83	13.1	19.6

LIWC results from personal text and formal writing for comparison

The text you submitted was 96 words in length.

Your writing:

I'm newly diagnosed with type 2 diabetes. I also struggle with both calcium and uric acid kidney stones as well as the rare blood disorder LEIDEN FACTOR V. Is there anyone in this community who deals with Leiden as well as diabetes? If there is I would LOVE to be able to chat with you regarding diet and possible weight loss plans. I currently have no regular doctor and a insurance so my diabetes is uncontrolled at this time. I am working hard to educate myself AND make the necessary changes to improve my current health.

Input text: A post from a 40 year old female member in American Diabetes Association online community

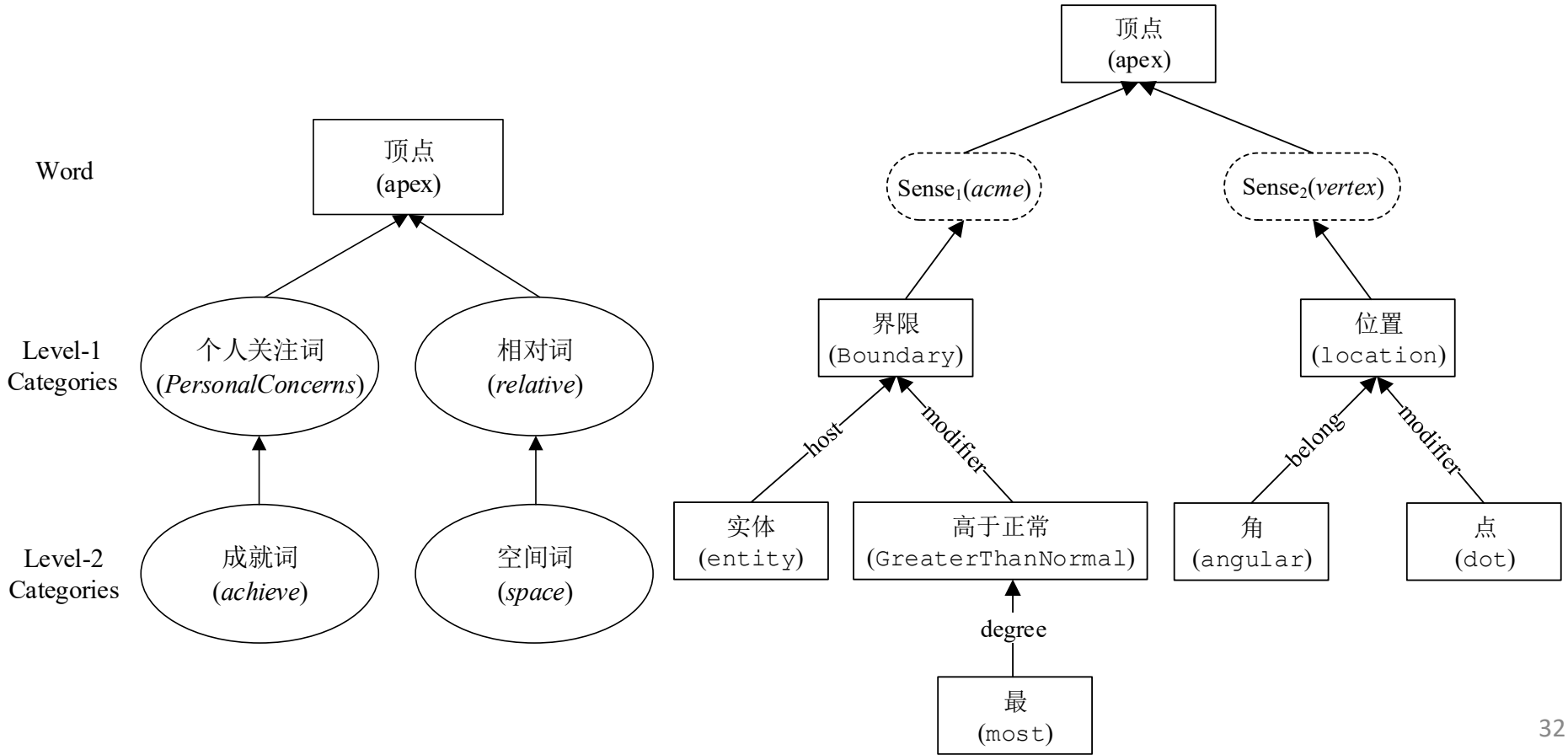
融合HowNet义原标注的词典扩展

- 以中文LIWC (Linguistic Inquiry and Word Count)为例，是计算社会科学中的著名词典
- LIWC中包含不到7000词，但中文中至少包括5万常用词

类别名称	英文简写	总词数	范例
认知历程词	cogmech	1255	理解、选择、质疑
洞察词	insight	328	了解、恍然大悟、体会
因果词	cause	128	引起、使得、变成
差距词	discrep	84	不足、纳闷、期待
暂订词	tentat	167	大约、未定、差不多
确切词	certain	145	不容置疑、必然、保证
限制词	inhib	292	废止、不准、规则
包含词	incl	82	包括、附近、添加
排除词	excl	39	取消、但是、除外

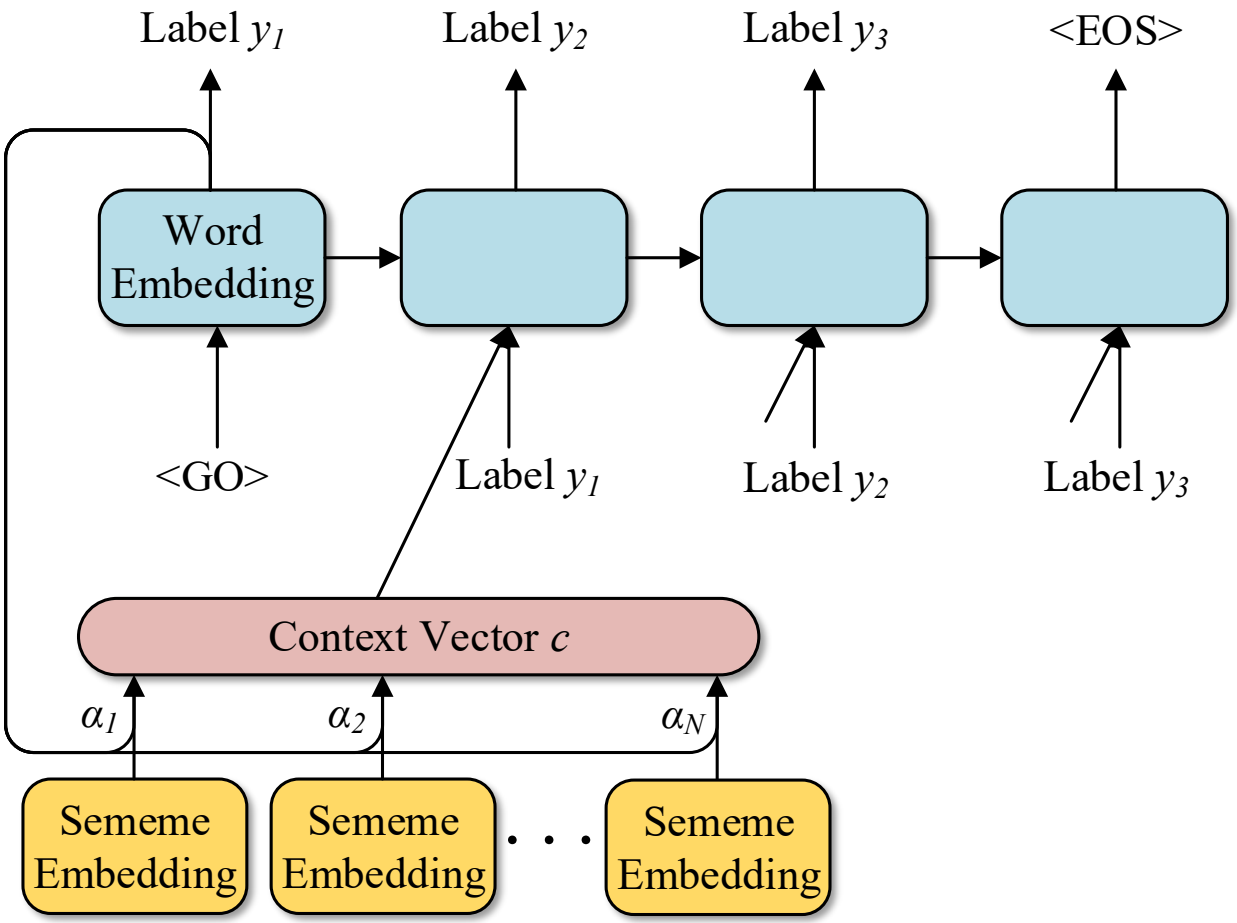
融合HowNet义原标注的词典扩展

- 以中文LIWC (Linguistic Inquiry and Word Count)为例，是计算社会科学中的著名词典
- 可以看做对词汇的层次分类



融合HowNet义原标注的词典扩展

- Hierarchical Decoder with Sememe Attention （AAAI 2018）



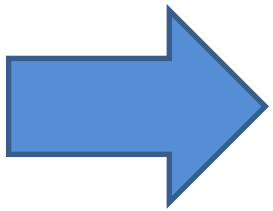
实验结果

- 我们提出的HDSA显著优于其他方法

	Overall		Level 1		Level 2		Level 3	
Model	Micro-F_1	W-M-F_1	Micro-F_1	W-M-F_1	Micro-F_1	W-M-F_1	Micro-F_1	W-M-F_1
TD k-NN	0.6198	0.6169	0.6756	0.6772	0.5716	0.5646	0.4884	0.4858
TD SVM	0.6283	0.6106	0.6858	0.6785	0.5766	0.5557	0.4503	0.4142
Structural SVM	0.6444	0.6448	0.7011	0.7010	0.5919	0.5919	0.5725	0.5718
CSSA	0.6511	0.6319	0.6880	0.6864	0.6172	0.5914	0.4729	0.4322
HD	0.7023	0.7000	0.7495	0.7476	0.6658	0.6614	0.6113	0.6064
HDSA	**0.7224**	**0.7204**	**0.7636**	**0.7616**	**0.6927**	**0.6874**	**0.6270**	**0.6234**

Word	Sememes	HD Prediction	HDSA Prediction	True Labels
恋人 (sweetheart)	交往 (associate), 人 (human), 爱恋 (love)	social←friend	social←friend, affect←posemo	social←friend, affect←posemo
今天 (today)	时间 (time), 现在 (present), 特定 (specific), 日 (day)	relativ←time	funct←TenseM←PresentM, relativ←time	funct←TenseM←PresentM, relativ←time
市镇 (town)	乡 (village), 市 (city), 地方 (place)	PersonalConcerns ←work	relativ←space	relativ←space
无望 (hopeless)	悲惨 (miserable)	cogmech←discrep	affect←negemo←sad	affect←negemo←sad
种种 (all kinds of)	多种 (various)	funct←negate	funct←quant	funct←quant
天空 (sky)	空域 (airspace)	relativ←time	relativ←space	relativ←space
联盟 (alliance)	结盟(ally), 团体(community)	PersonalConcerns ←work	social, PersonalConcerns←work	PersonalConcerns ←work
泪珠 (teardrop)	部件(part), 体液 (BodyFluid), 动物(AnimalHuman)	affect←negemo←sad	affect←negemo, bio←health	affect←negemo←sad

知识获取



数据驱动的
深度学习

符号表示的
义原知识

基于语义表示学习的义原推荐

- HowNet等知识库主要依赖人工标注，费时费力
- **义原自动推荐**：实现义原知识库与时俱进，提升标注一致性

基于词向量的近邻协同过滤方法（SPWE）

$$P(s_j, w) = \sum_{w_i \in W} \cos(\mathbf{w}, \mathbf{w_i}) \cdot \mathbf{M}_{ij} \cdot c^{r_i}$$

基于词义-义原矩阵分解的推荐方法（SPSE）

$$\mathcal{L} = \sum_{w_i \in W, s_j \in S} \left(\mathbf{w}_i \cdot (\mathbf{s}_j + \bar{\mathbf{s}}_j) + \mathbf{b}_i + \mathbf{b}'_j - \mathbf{M}_{ij}\right)^2 + \lambda \sum_{s_j, s_k \in S} \left(\mathbf{s}_j \cdot \bar{\mathbf{s}}_k - \mathbf{C}_{jk}\right)^2,$$

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

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
SPWE+SPASE	0.683
SPWE+SPSE	**0.713**

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

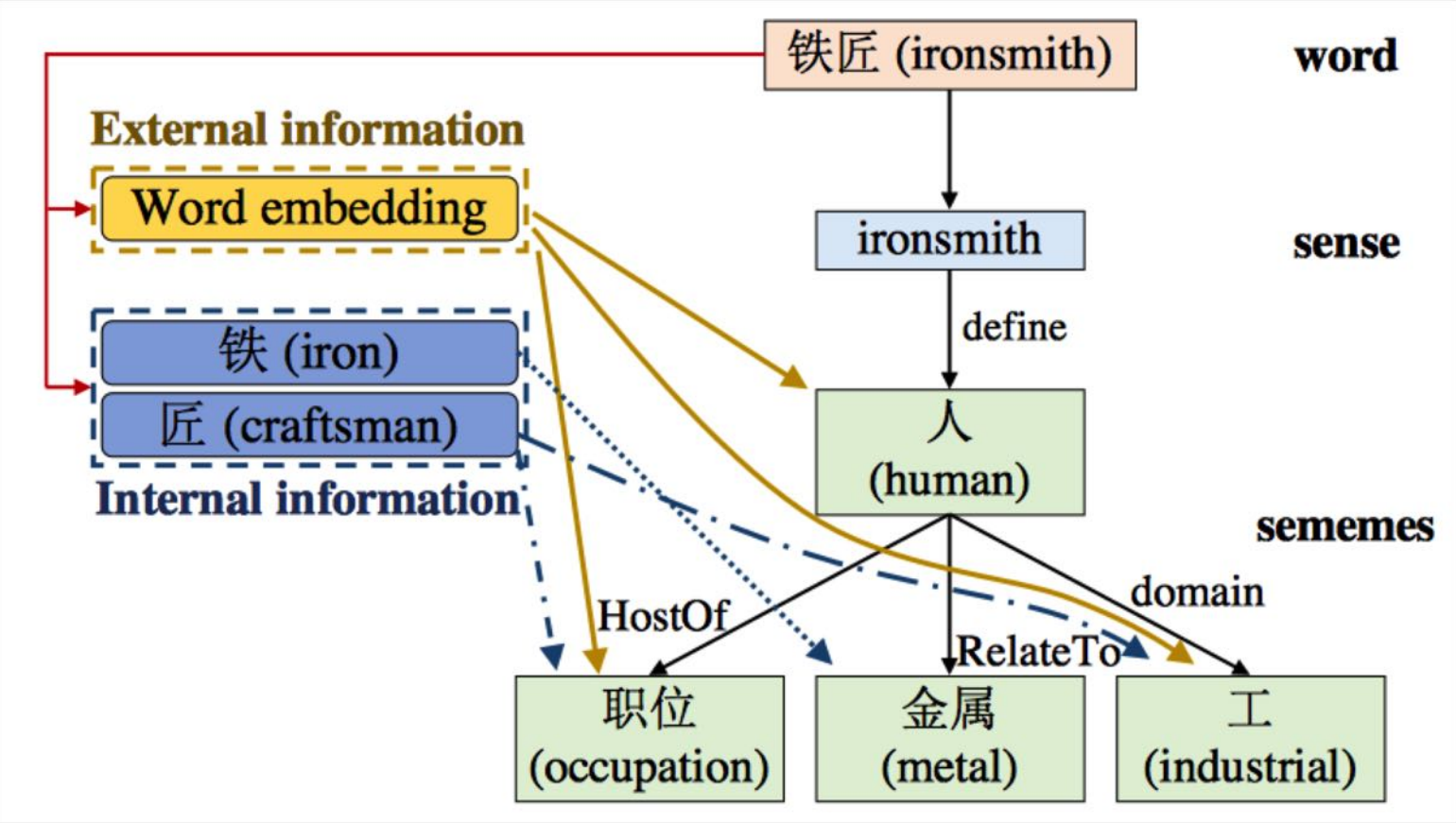
word frequency	number of words	MAP
<800	1,659	0.817
800 - 3,000	1,494	0.736
3,001 - 15,000	1,672	0.690
>15,000	1,311	0.596

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)

考虑内部汉字信息的义原推荐

- 词汇内部的汉字信息对语义理解具有重要意义，提出同时考虑内部汉字信息进行义原推荐



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

- 实验证明，考虑内部汉字信息，对于低频词的义原推荐提升尤为明显

word frequency occurrences	⩽ 50 8537	51– 100 4868	101 – 1,000 3236	1,001 – 5,000 2036	5,001 – 10,000 663	10,001 – 30,000 753	>30,000 686
SPWE	0.312	0.437	0.481	0.558	0.549	0.556	0.509
SPSE	0.187	0.273	0.339	0.409	0.407	0.424	0.386
SPWE + SPSE	0.284	0.414	0.478	0.556	0.548	0.554	0.511
SPWCF	0.456	0.414	0.400	0.443	0.462	0.463	0.479
SPCSE	0.309	0.291	0.286	0.312	0.339	0.353	0.342
SPWCF + SPCSE	0.467	0.437	0.418	0.456	0.477	0.477	0.494
SPWE + fastText	0.495	0.472	0.462	0.520	0.508	0.499	0.490
CSP	**0.527**	**0.555**	**0.555**	**0.626**	**0.632**	**0.641**	**0.624**

words	models	Top 5 sememes
钟表匠 (clockmaker)	internal	人**(human)**, 职位**(occupation)**, 部件(part), 时间**(time)**, 告诉**(tell)**
	external	人**(human)**, 专(ProperName), 地方(place), 欧洲(Europe), 政(politics)
	ensemble	人**(human)**, 职位**(occupation)**, 告诉**(tell)**, 时间**(time)**, 用具**(tool)**
奥斯卡 (Oscar)	internal	专**(ProperName)**, 地方(place), 市(city), 人(human), 国都(capital)
	external	奖励**(reward)**, 艺**(entertainment)**, 专**(ProperName)**, 用具(tool), 事情**(fact)**
	ensemble	专**(ProperName)**, 奖励**(reward)**, 艺**(entertainment)**, 著名(famous), 地方(place)

相关论文

- 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

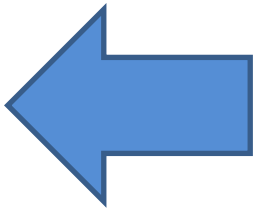
- 以Google Knowledge Graphs为代表的世界知识库，用三元组形式记录知识

莎士比亚

写作

罗密欧与朱丽叶

知识指导

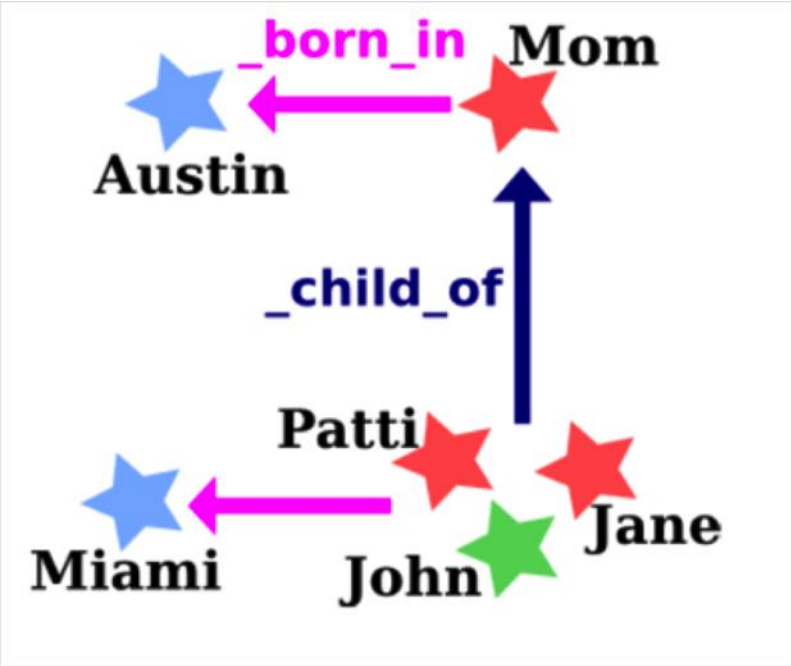


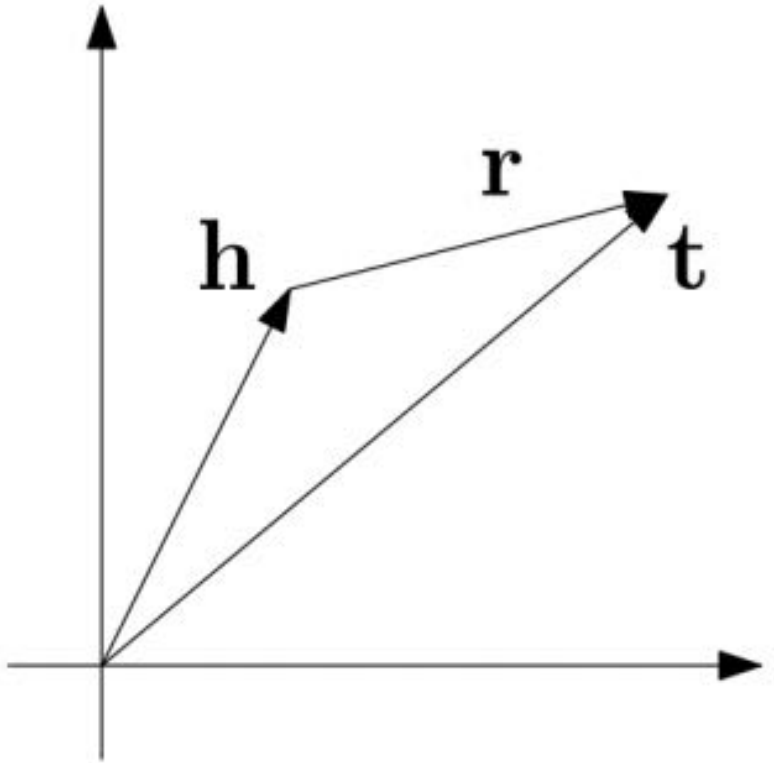
数据驱动的
深度学习

符号表示的
世界知识

世界知识的分布式表示学习

- TransE对每个事实 (head, relation, tail)，将其中的relation作为从head到tail的平移操作





优化目标: **h + r = t**

评价任务：链接预测

WALL-E _has_genre ?

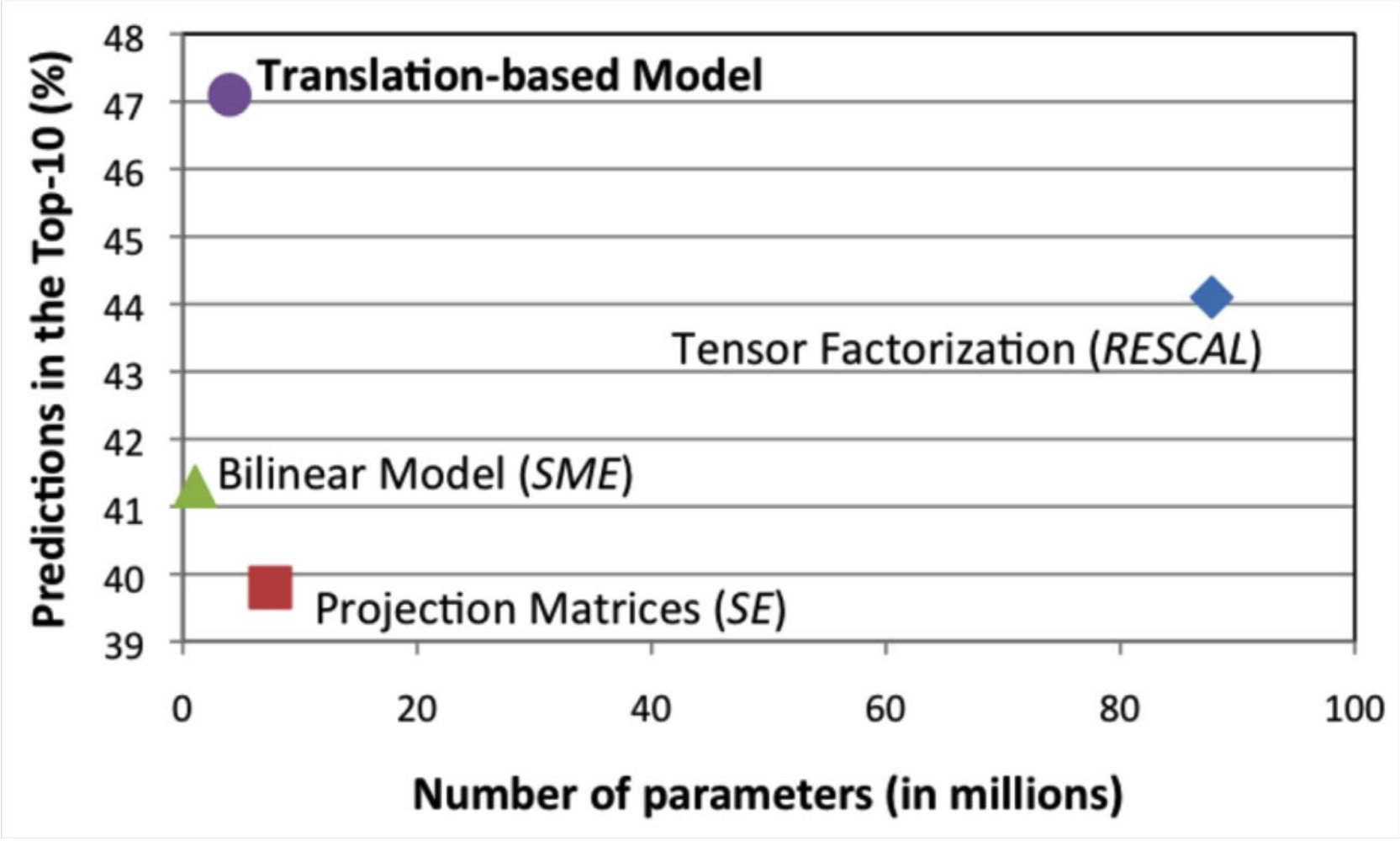
评价任务：链接预测

WALL-E _has_genre

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

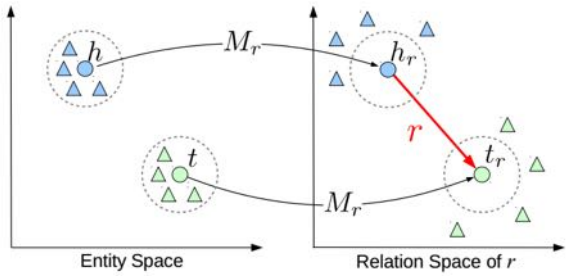
链接预测性能比较

Freebase15K

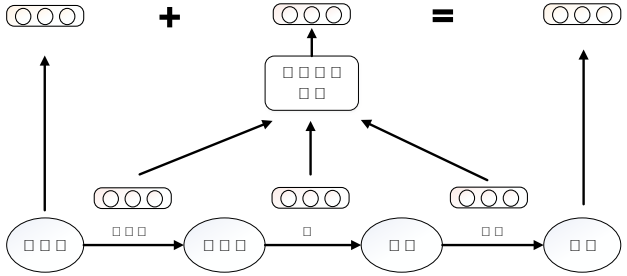


世界知识的分布式表示学习

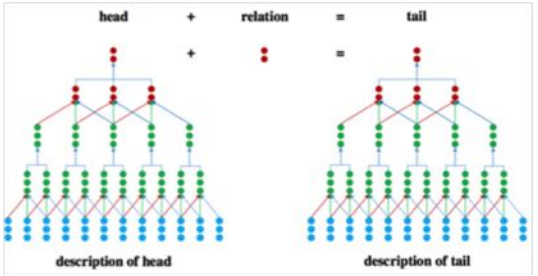
- 利用知识图谱和实体描述、类别和图像等外部信息，实现高效知识表示学习



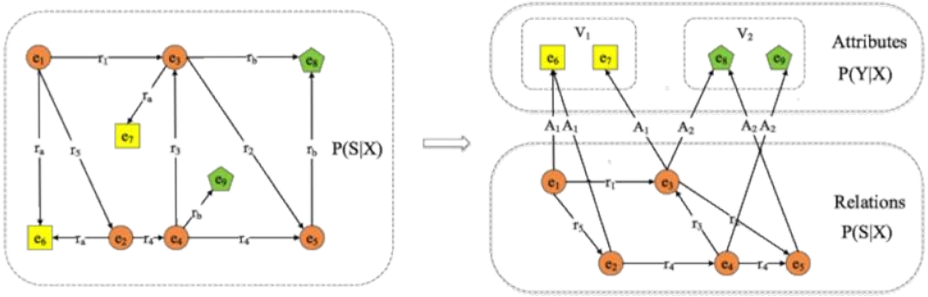
考虑复杂关系类型的知识表示
TransR (AAAI 2015)



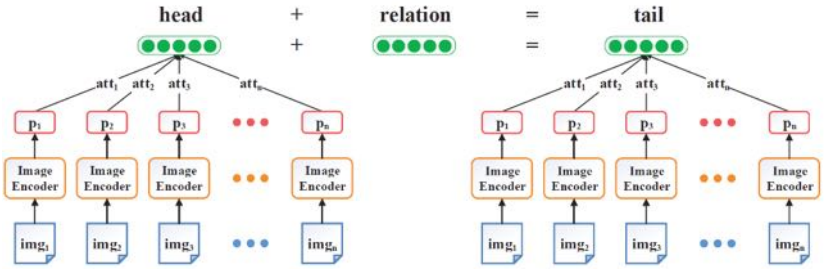
考虑关系路径的知识表示
PTransE (EMNLP 2015)



考虑实体描述信息的知识表示
DKRL (AAAI 2016)



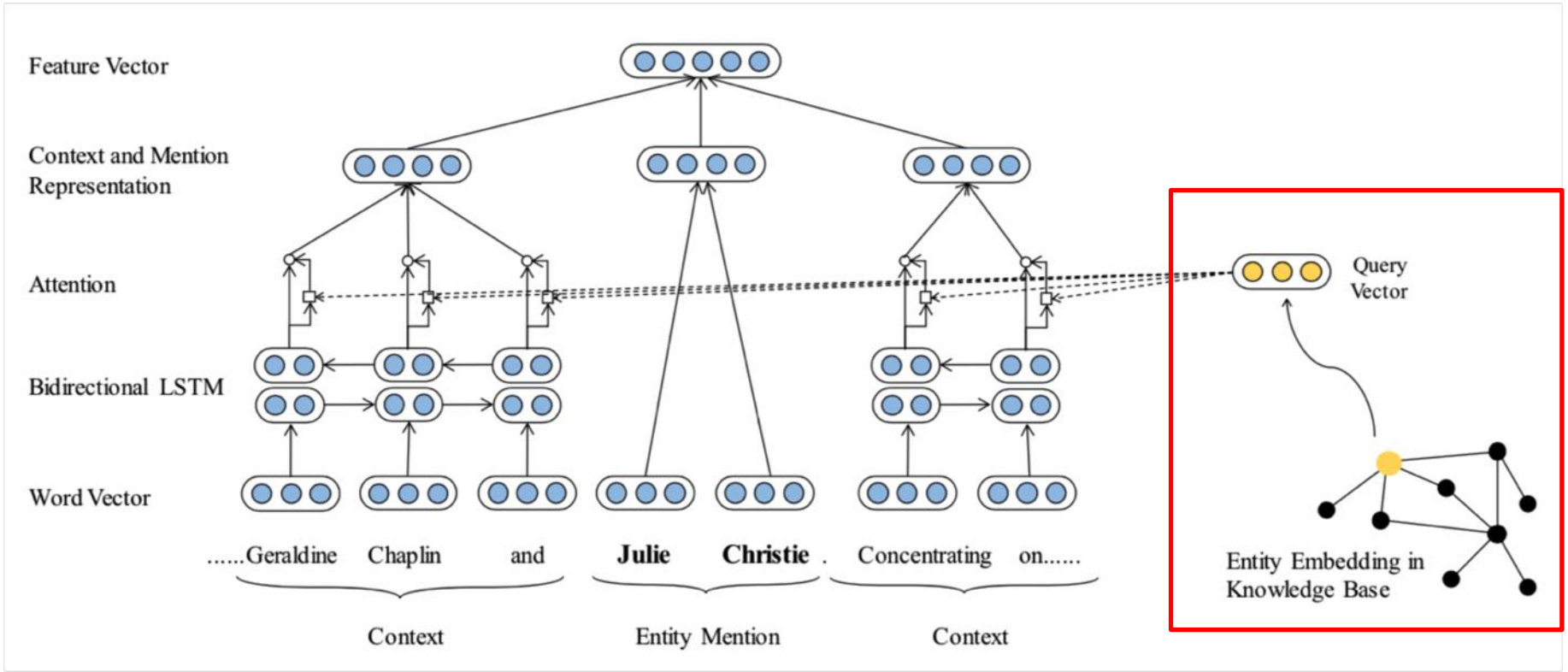
综合考虑实体、属性与关系的知识表示
KR-EAR (IJCAI 2016)



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



Ji Xin, Yankai Lin, Zhiyuan Liu, Maosong Sun. Improving Neural Fine-Grained Entity Typing with Knowledge Attention. The 32th AAAI Conference on Artificial Intelligence (AAAI 2018).

实验结果

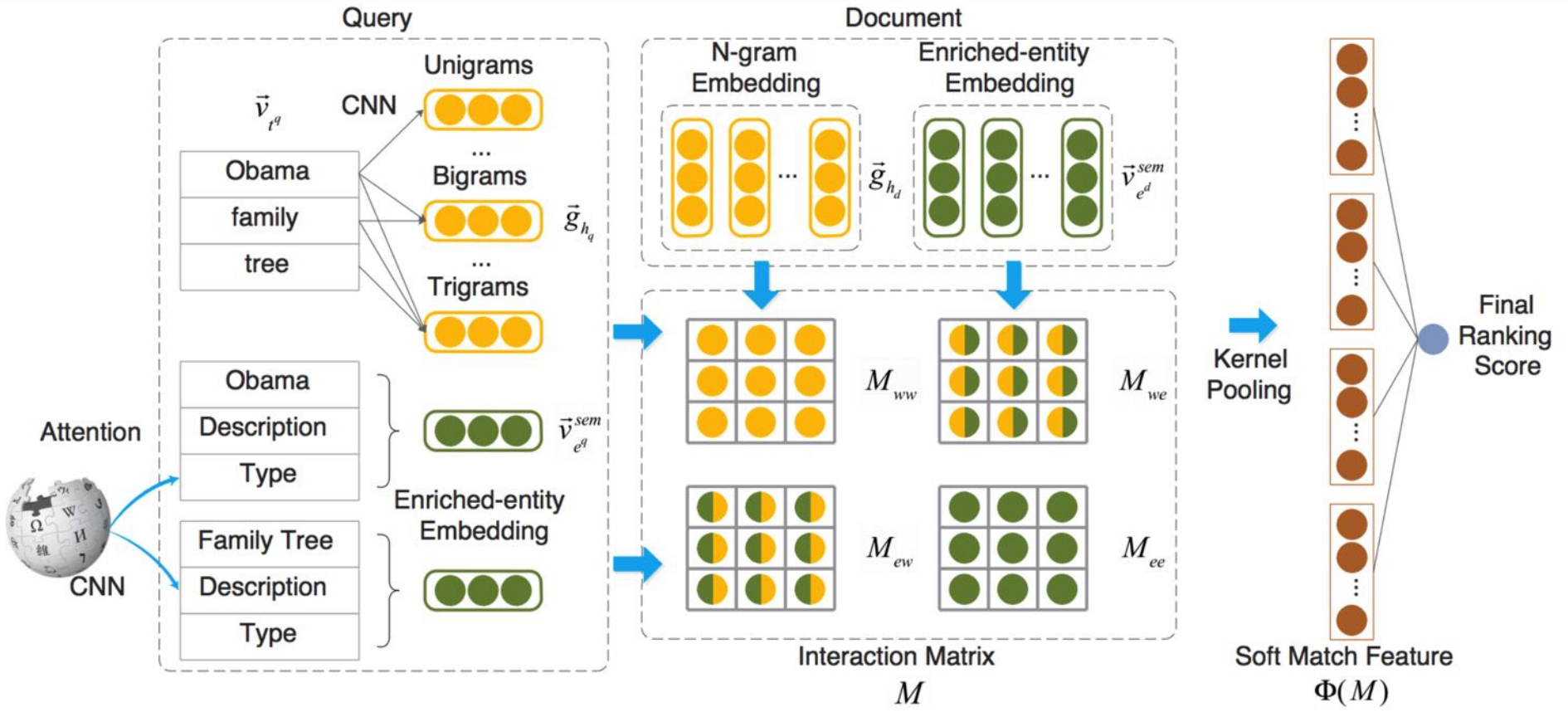
- 对文本实体进行细粒度分类，助力深度分析
- 充分利用KG实体表示，提出知识注意力机制，建立对上下文的高效建模
- 显著提升实体分类性能

Dateset	WIKI-AUTO							WIKI-MAN						
Metrics	Strict	Macro			Micro			Strict	Macro			Micro		
	Acc	Pre	Rec	F1	Pre	Rec	F1	Acc	Pre	Rec	F1	Pre	Rec	F1
AFET	20.32	67.00	45.82	54.75	69.29	42.40	52.61	18.00	64.50	50.00	56.33	64.29	50.43	56.52
KB-ONLY	35.12	69.65	71.35	70.49	54.85	74.99	63.36	17.00	55.50	72.83	63.00	27.81	74.57	40.52
HNM	34.88	68.09	61.03	64.37	72.80	64,48	68.39	15.00	61.80	68.00	64.75	62.35	68.53	65.30
SA	42.77	75.33	69.69	72.40	77.35	72.63	74.91	18.00	66.67	73.67	69.44	65.54	75.43	70.14
MA	41.58	73.64	71.71	72.66	75.94	75.52	75.72	26.00	65.13	78.50	71.19	64.09	82.33	72.08
KA	45.49	74.82	72.46	73.62	76.96	75.49	76.22	23.00	64.69	78.92	71.10	63.25	82.68	71.67
KA+D	**47.20**	**75.72**	**74.03**	**74.87**	**77.96**	**77.87**	**77.92**	**34.00**	**68.41**	**82.83**	**74.94**	**66.12**	**87.50**	**75.32**

Ji Xin, Yankai Lin, Zhiyuan Liu, Maosong Sun. Improving Neural Fine-Grained Entity Typing with Knowledge Attention. The 32th AAAI Conference on Artificial Intelligence (AAAI 2018).

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



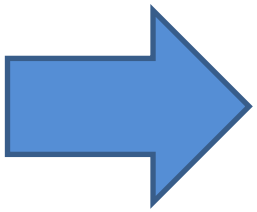
Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. Entity-Duet Neural Ranking: Understanding the Role of Knowledge Graph Semantics in Neural Information Retrieval. ACL 2018.

实验结果

- 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
- 显著提升文档匹配效果

	Testing-SAME				Testing-DIFF				Testing-RAW	
Method	**NDCG@1**		**NDCG@10**		**NDCG@1**		**NDCG@10**		**MRR**	
BM25	0.1422	−46.24%	0.2868	−31.67%	0.1631	−45.63%	0.3254	−23.04%	0.2280	−33.86%
RankSVM	0.1457	−44.91%	0.3087	−26.45%	0.1700	−43.33%	0.3519	−16.77%	0.2241	−34.99%
Coor-Ascent	0.1594	−39.74%	0.3547	−15.49%	0.2089	−30.37%	0.3775	−10.71%	0.2415	−29.94%
DRMM	0.1367	−48.34%	0.3134	−25.34%	0.2126‡	−29.14%	0.3592§	−15.05%	0.2335	−32.26%
CDSSM	0.1441	−45.53%	0.3329	−20.69%	0.1834	−38.86%	0.3534	−16.41%	0.2310	−33.00%
MP	0.2184†‡	−17.44%	0.3792†‡	−9.67%	0.1969	−34.37%	0.3450	−18.40%	0.2404	−30.27%
K-NRM	0.2645	–	0.4197	–	0.3000	–	0.4228	–	0.3447	–
Conv-KNRM	0.3357†‡§¶	+26.90%	0.4810†‡§¶	+14.59%	0.3384†‡§¶	+12.81%	0.4318†‡§	+2.14%	0.3582†‡§	+3.91%
EDRM-KNRM	0.3096†‡§¶	+17.04%	0.4547†‡§¶	+8.32%	0.3327†‡§¶	+10.92%	0.4341†‡§¶	+2.68%	0.3616†‡§¶	+4.90%
EDRM-CKNRM	**0.3397**†‡§¶	+28.42%	**0.4821**†‡§¶	+14.86%	**0.3708**†‡§¶*	+23.60%	**0.4513**†‡§¶*	+6.74%	**0.3892**†‡§¶*	+12.90%

知识获取

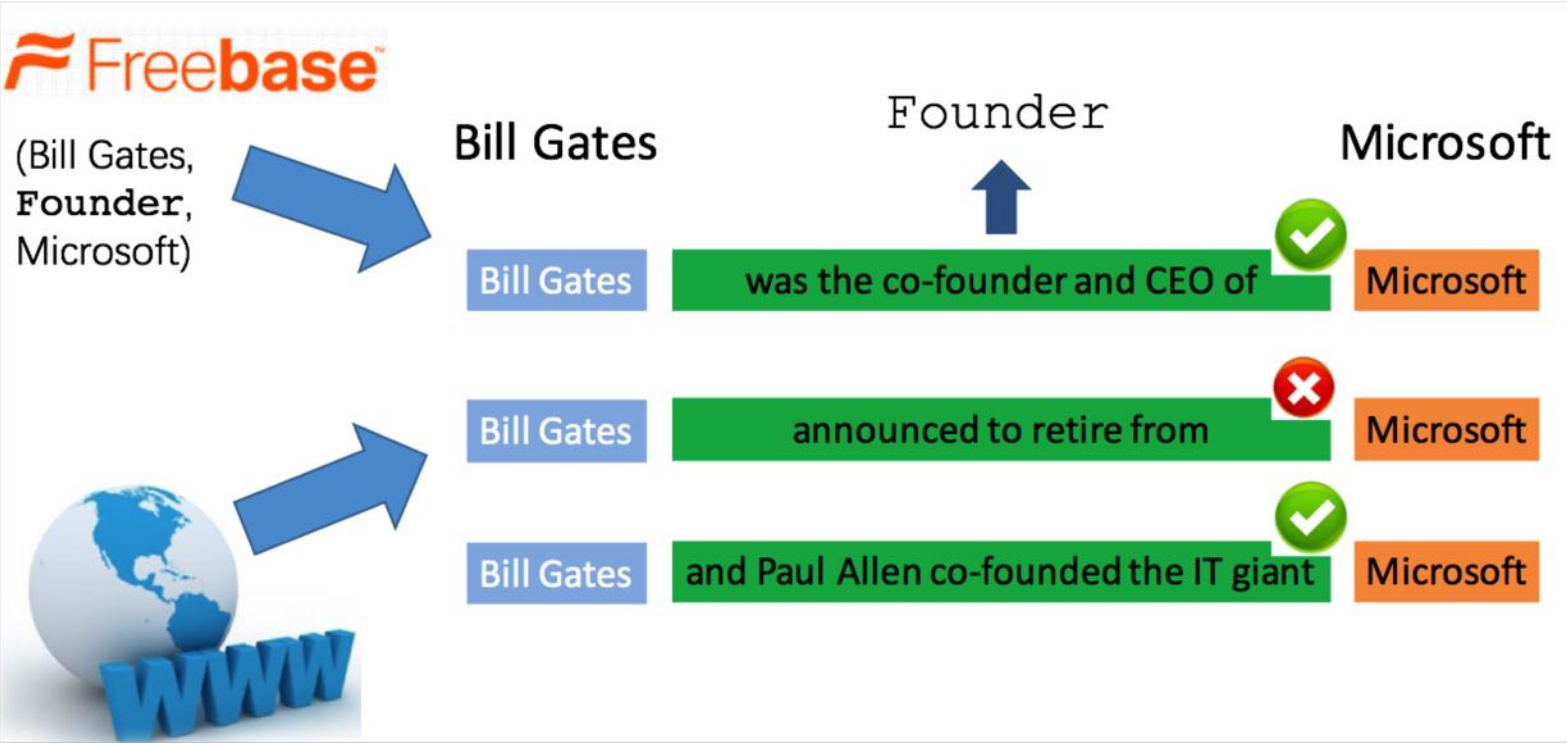


数据驱动的
深度学习

符号表示的
世界知识

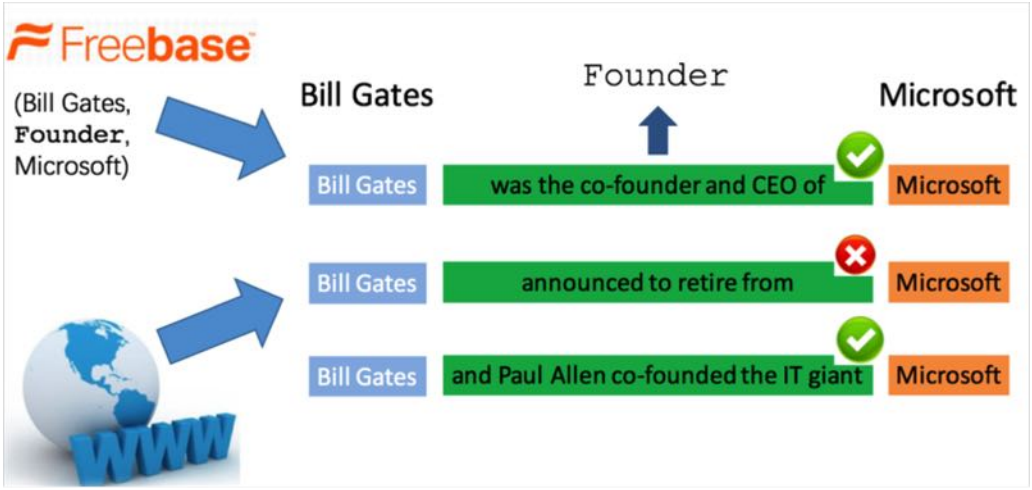
知识获取

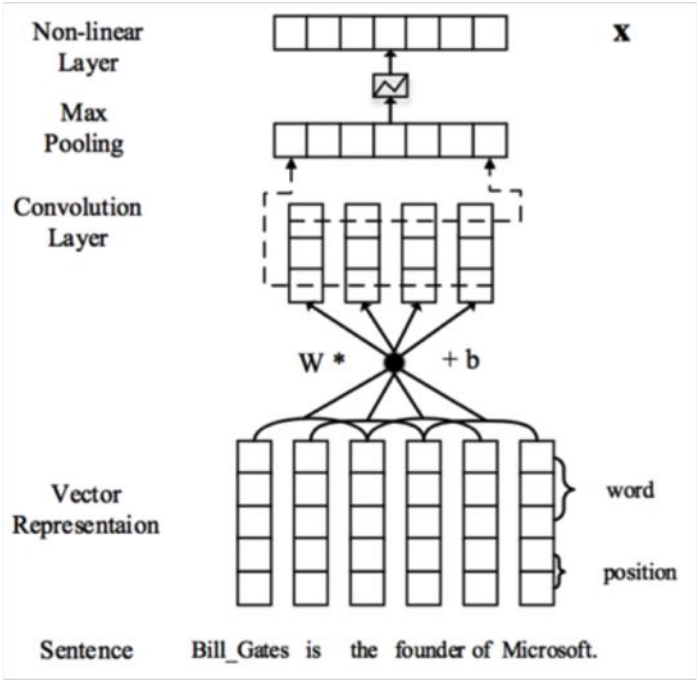
- 基于已有知识和海量文本信息获取结构化知识
- 解决标注数据噪音，融合多源信息



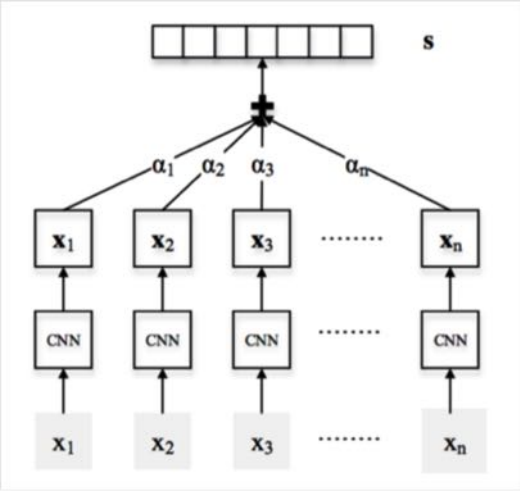
神经网络知识获取技术

- 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

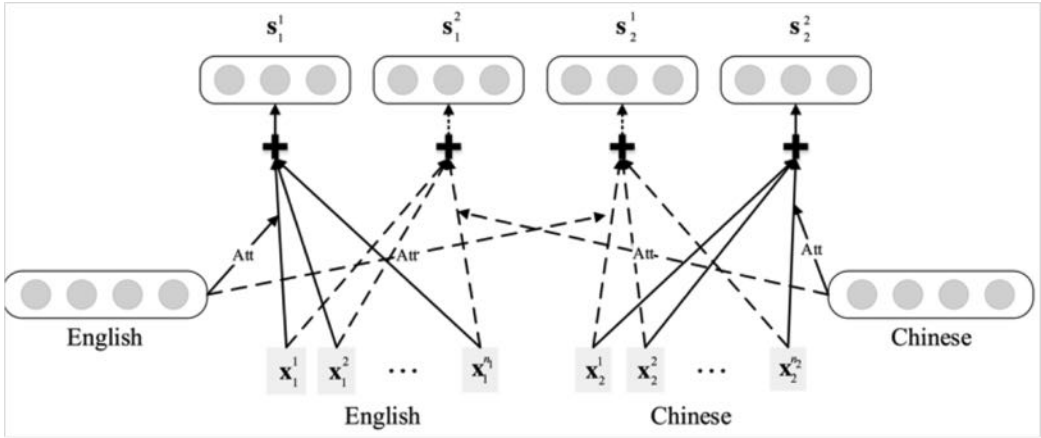




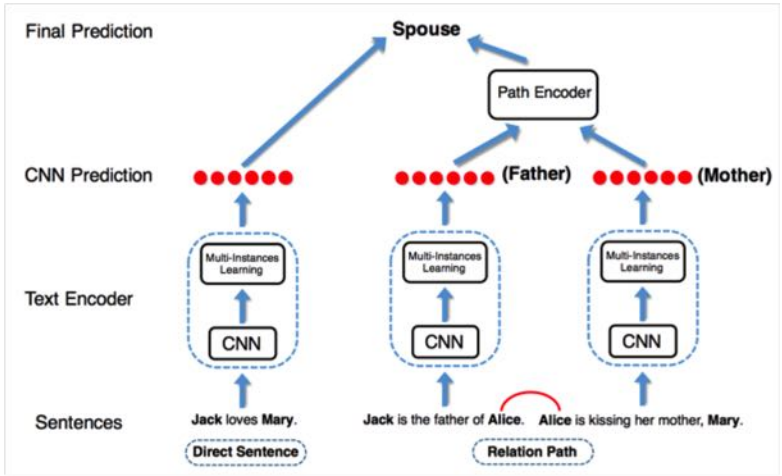
高效鲁棒的知识获取技术



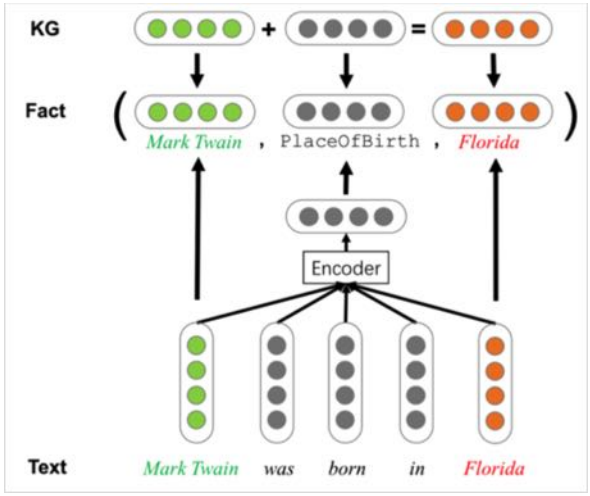
基于句级注意力机制的远程监督
神经网络关系抽取模型(ACL 2016)



基于跨语言注意力机制的
神经网络关系抽取模型(ACL 2017)



考虑关系路径的
神经网络关系抽取模型(EMNLP 2017)



综合知识图谱与文本的
神经网络关系抽取模型(AAAI 2018)

相关论文

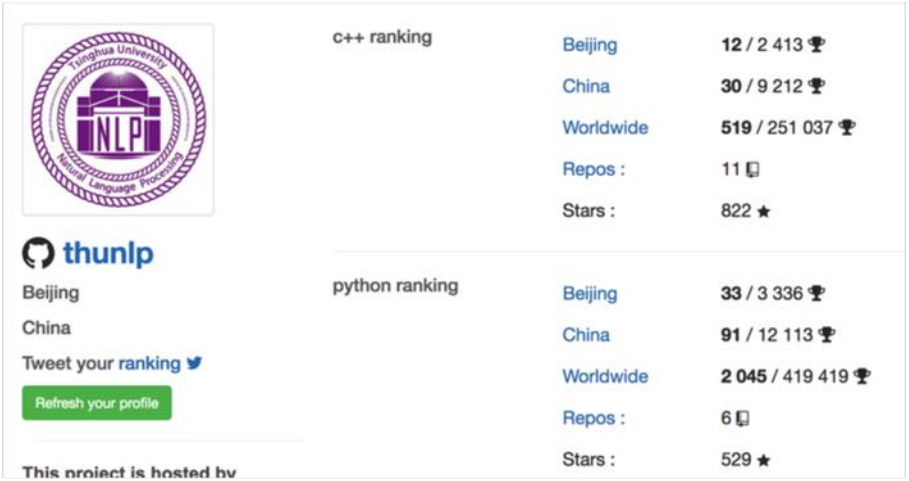
- Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, Peng Li. **Hierarchical Relation Extraction with Coarse-to-Fine Grained Attention**. EMNLP 2018.
- Zhenghao Liu, Chenyan Xiong, Maosong Sun, Zhiyuan Liu. **Entity-Duet Neural Ranking: Understanding the Role of Knowledge Graph Semantics in Neural Information Retrieval**. ACL 2018.
- Xu Han, Zhiyuan Liu, Maosong Sun. **Neural Knowledge Acquisition via Mutual Attention between Knowledge Graph and Text**. 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. **Neural Relation Extraction with Multi-lingual Attention**. ACL 2017.
- Wenyuan Zeng, Yankai Lin, Zhiyuan Liu, Maosong Sun. **Incorporating Relation Paths in Neural Relation Extraction**. EMNLP 2017.
- Yankai Lin, Zhiyuan Liu, Maosong Sun. **Knowledge Representation Learning with Entities, Attributes and Relations**. IJCAI 2016.

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10000+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

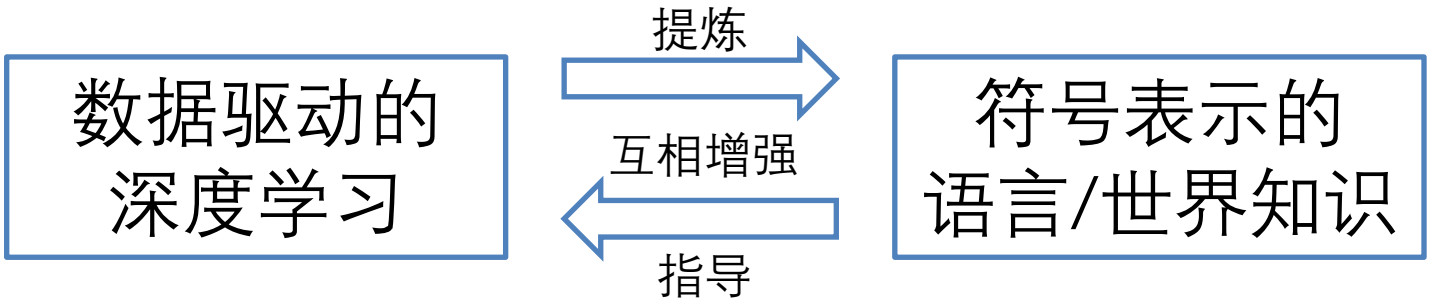


总结展望

- 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

AI = 数据驱动 + 知识指导

- 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



感谢各位老师同学

http://nlp.csai.tsinghua.edu.cn/~lzy/