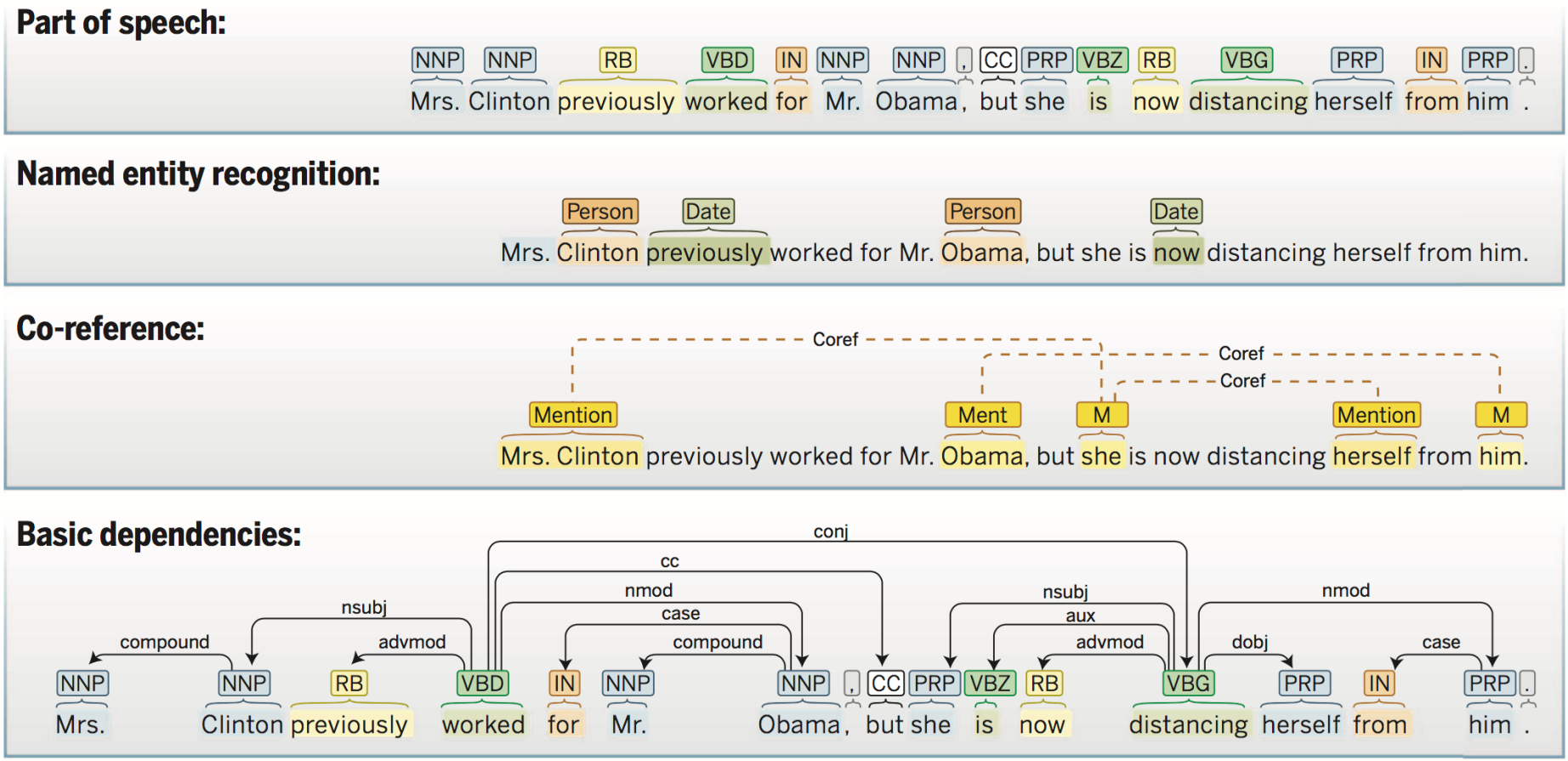
在深度学习时代用HowNet搞事情

清华大学自然语言处理实验室

刘知远

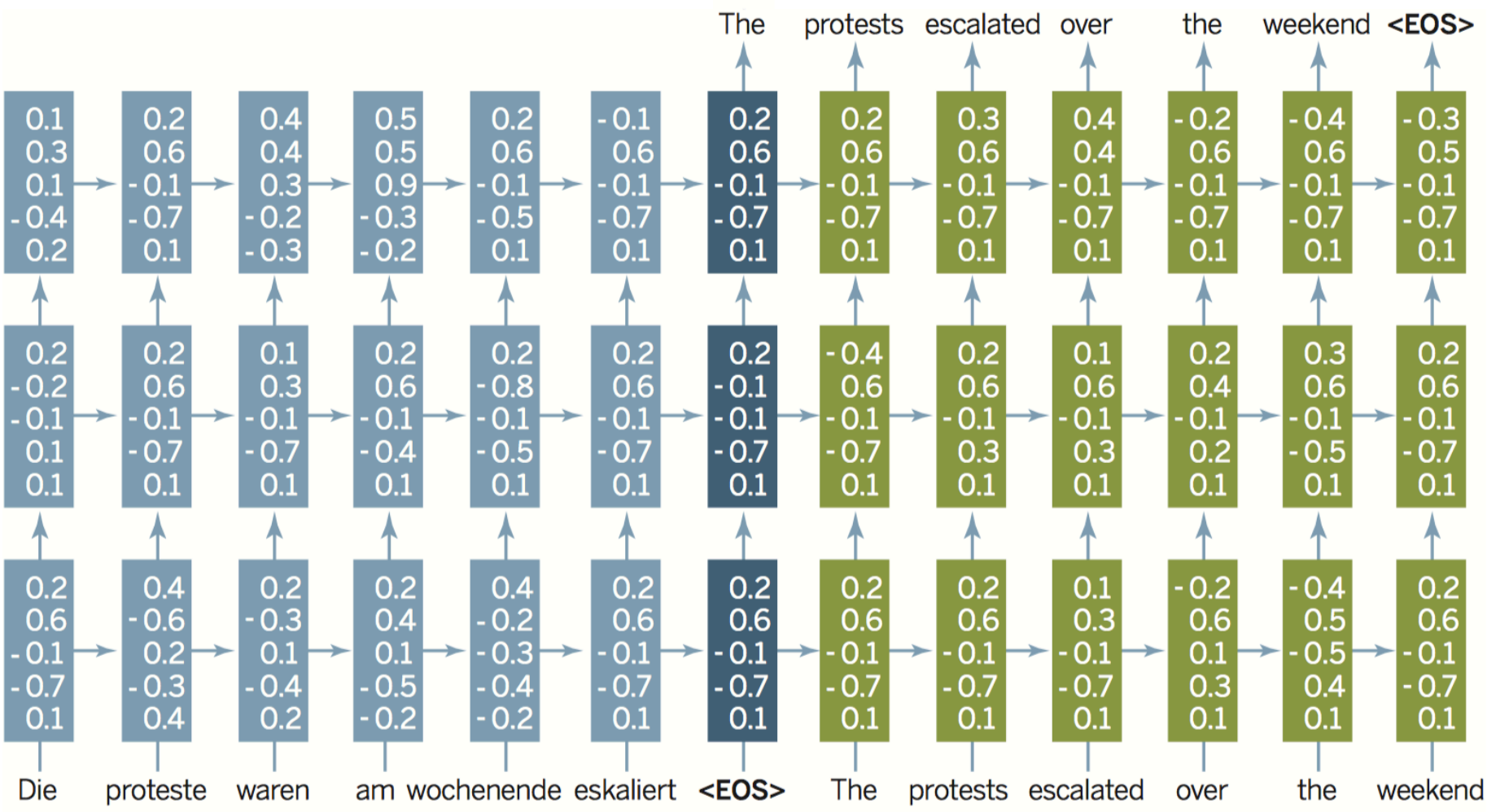
自然语言处理

- 自然语言旨在理解与表示人类语言的语义信息



Advances in Natural Language Processing. Science 2015.

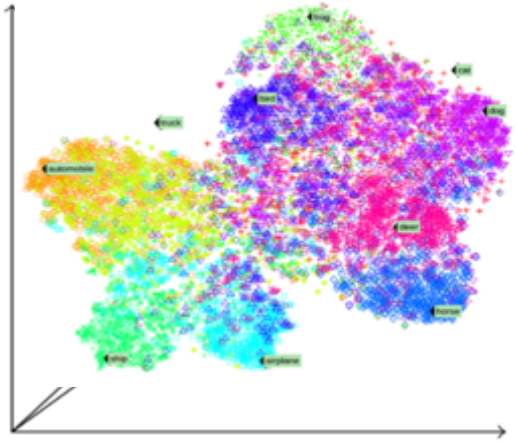
自然语言处理与深度学习



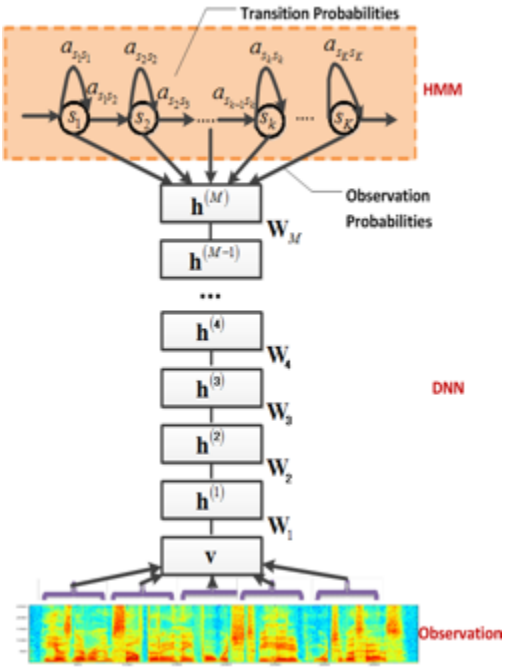
Advances in Natural Language Processing. Science 2015.

深度学习特点

- 分布式（Distributed）表示
  - 嵌入（Embeddings）
  - 稠密、实值、低维向量



- 层次（Hierarchical）结构
  - 对应层次的真实世界
  - 具有抽象和泛化能力



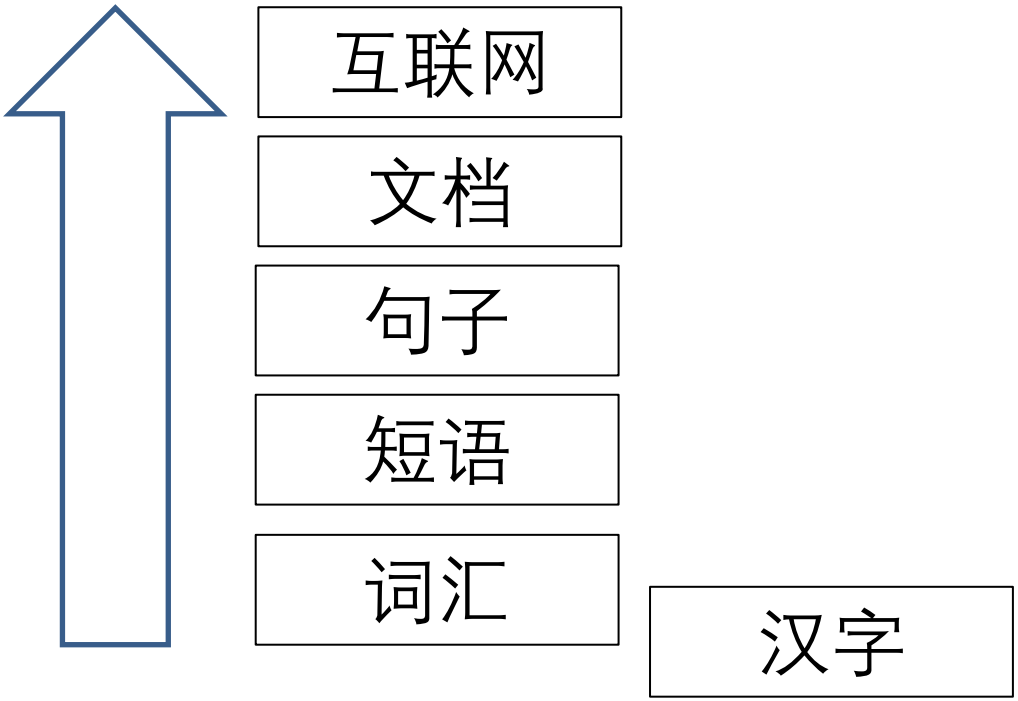
自然语言处理与深度学习

In the short term, 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



语义符号表示

- 又名one-hot表示，词袋模型的基础

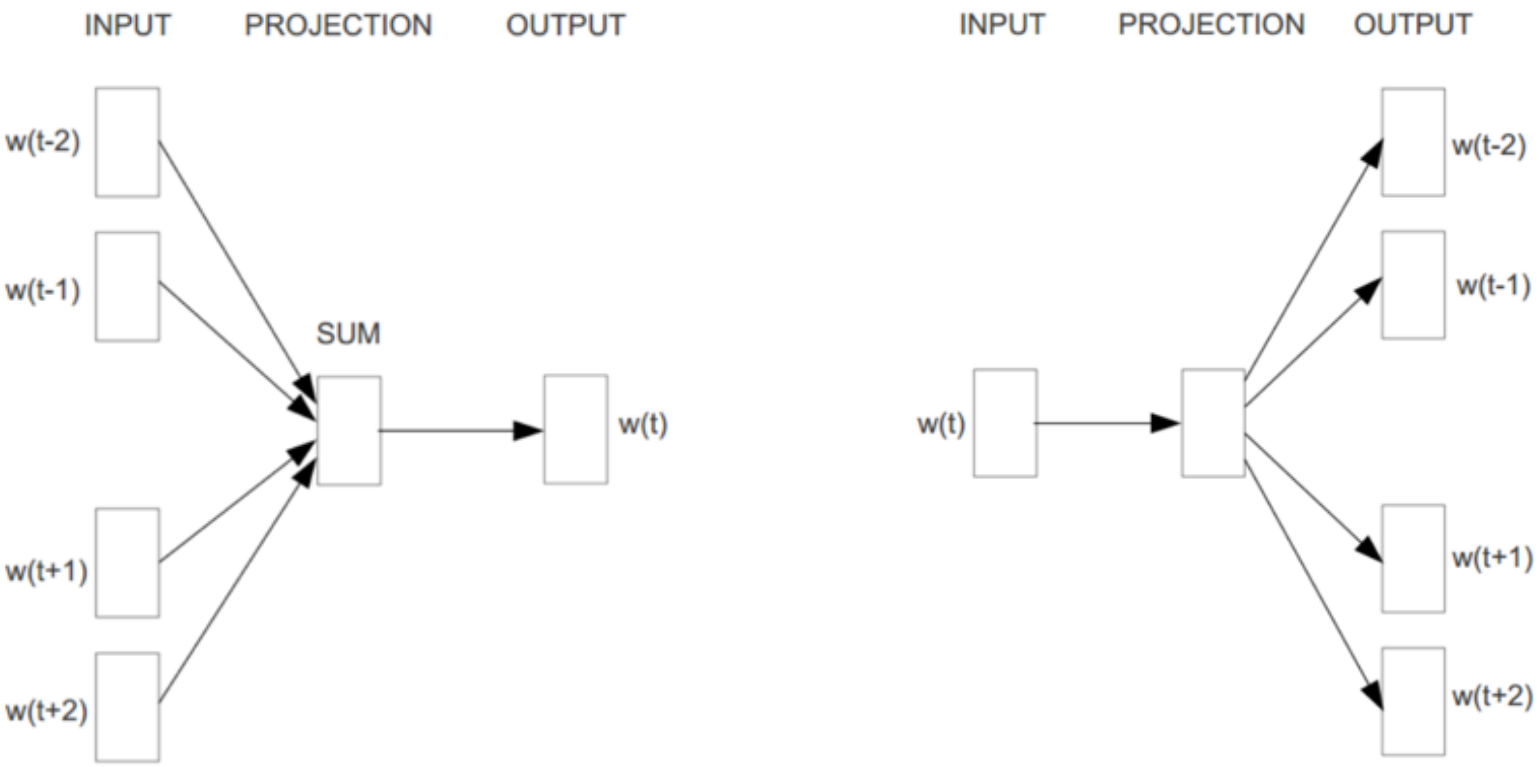
star [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, …]

sun [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …]

Sim (star, sun) = 0

语义分布式表示

- 深度学习利用纯数据驱动方法学习语义表示

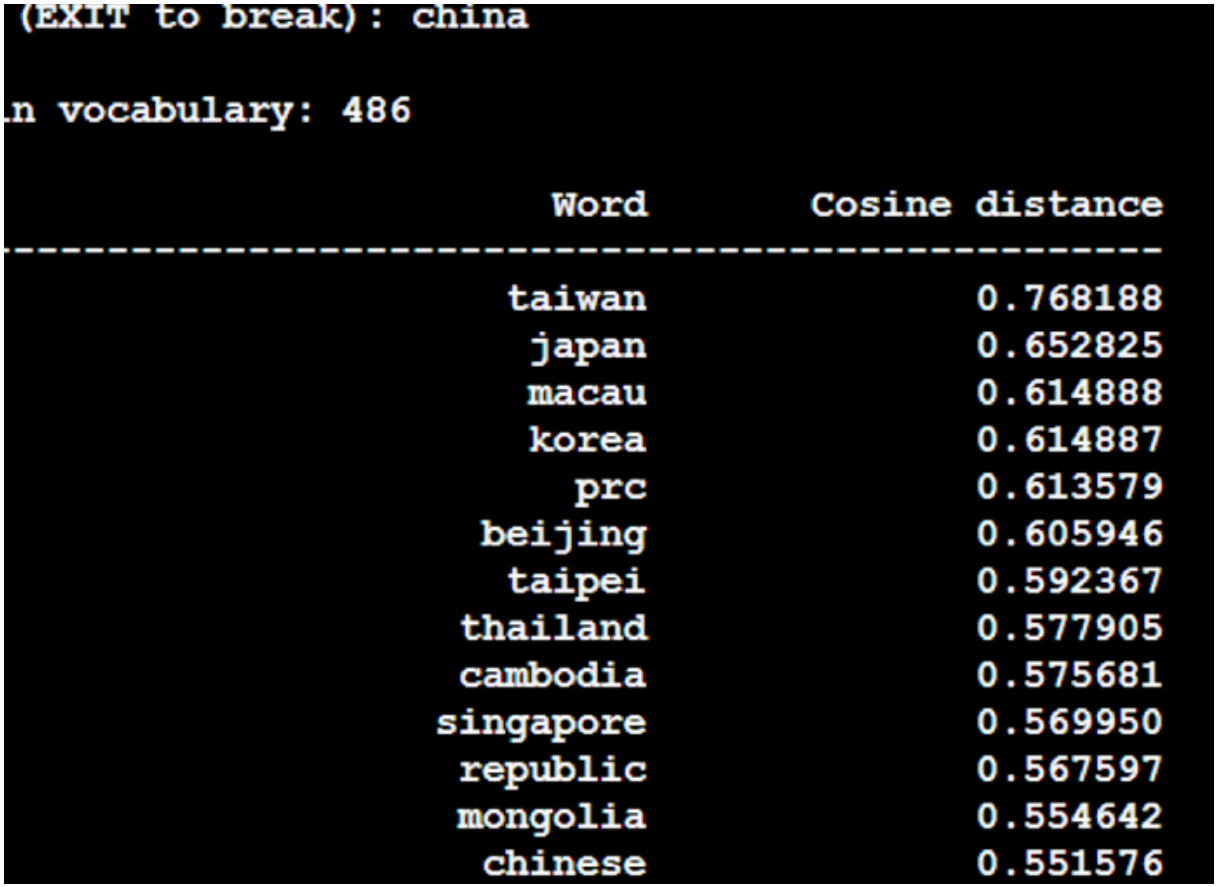


word2vec

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

语义分布式表示

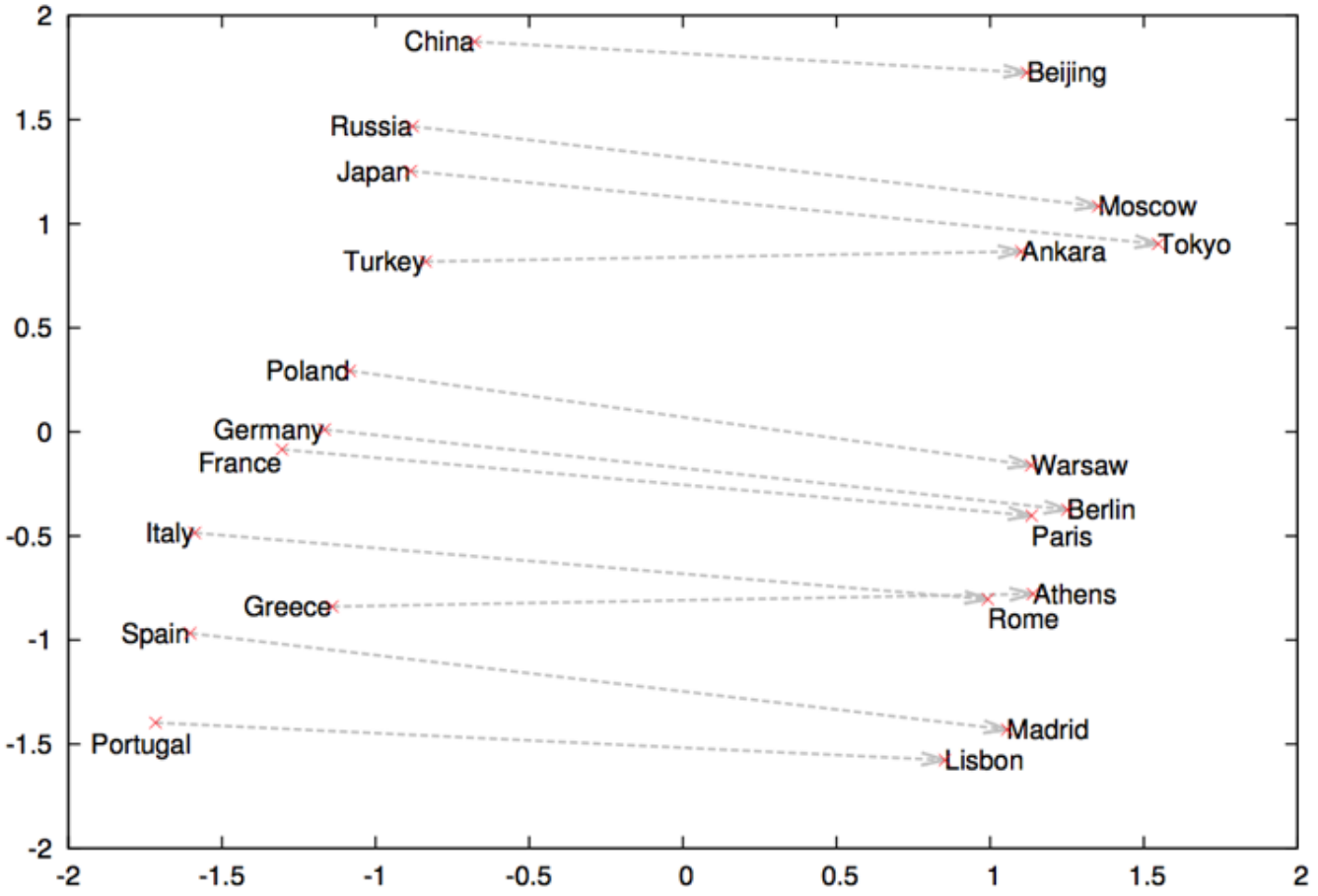
- 深度学习利用纯数据驱动方法学习语义表示



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

语义分布式表示

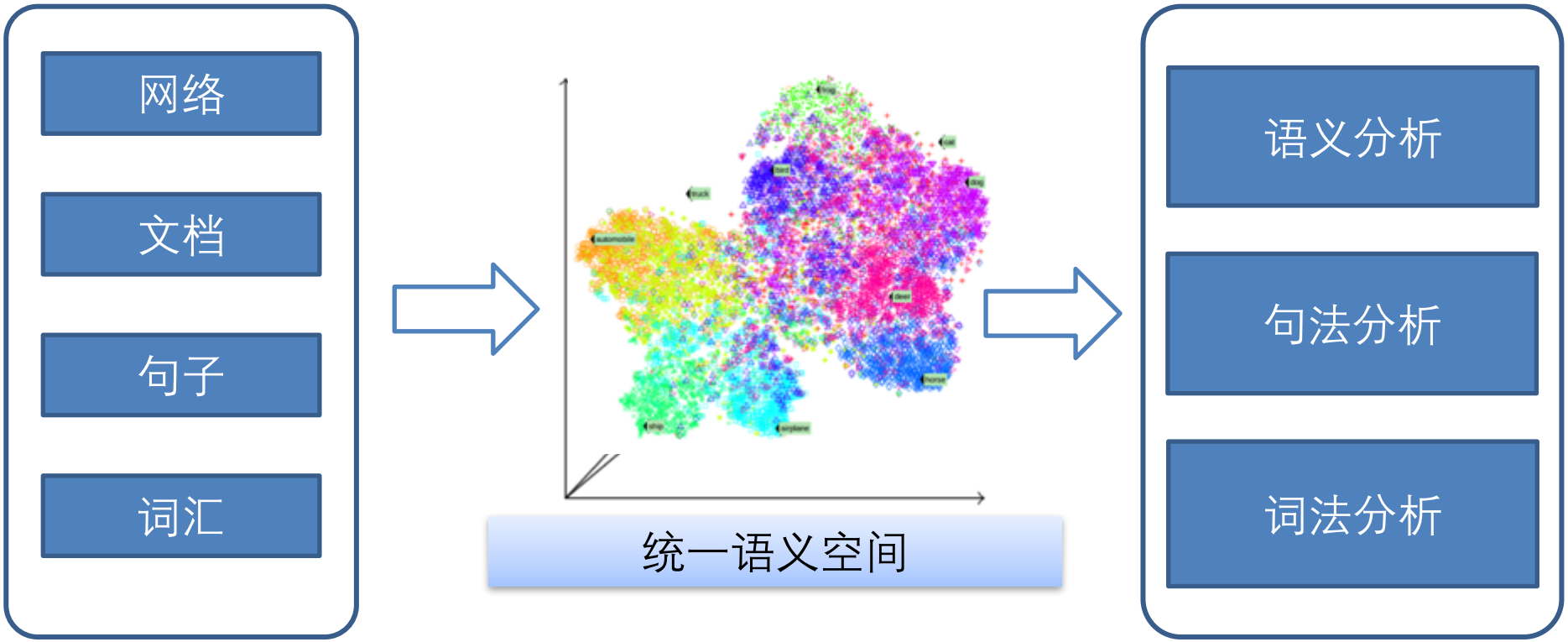
- 深度学习利用纯数据驱动方法学习语义表示



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

分布式表示优势

- 解决大数据NLP的数据稀疏问题
- 实现跨领域、跨对象的知识迁移
- 提供多任务学习的统一底层表示



语言知识库

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：

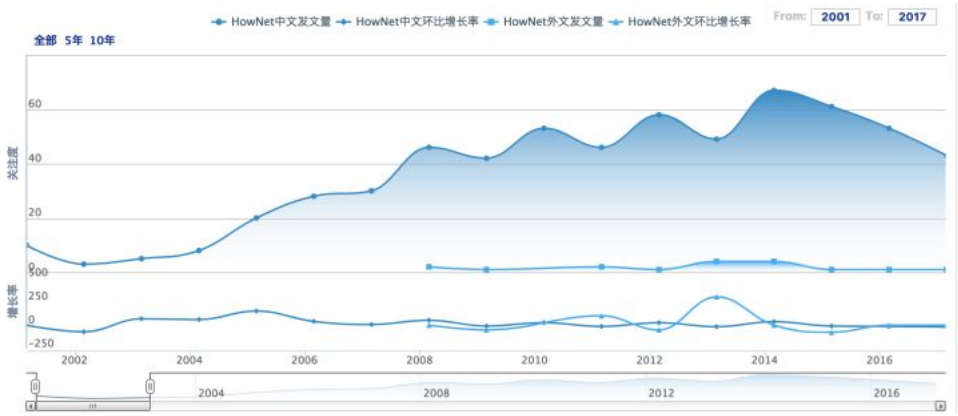
1. 每一个词的语义描述由多个义原组成；
2. 词语的语义描述中各个义原并不是平等的，它们之间有着复杂的关系，通过一种专门的知识描述语言来表示。

我们的工作主要包括：

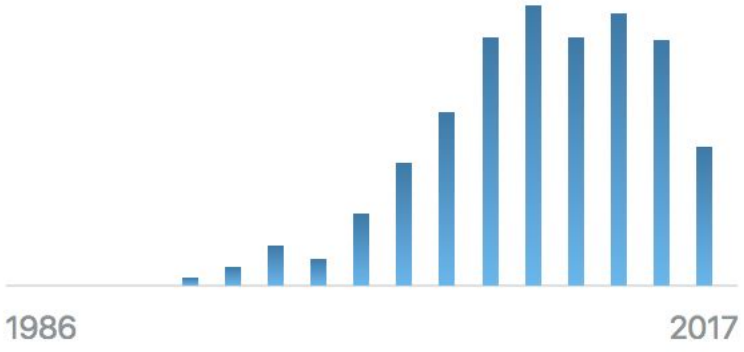
1. 研究《知网》中知识描述语言的语法，了解其描述一个词义所用的多个义原之间的关系，区分其在词语相似度计算中所起的作用；我们采用一种更

深度学习时代的语言知识库

- 由于大规模文本数据日益增长，以及深度学习的数据驱动特性，语言知识库关注度逐年下降



中国期刊网（CNKI)统计HowNet学术关注度变化趋势

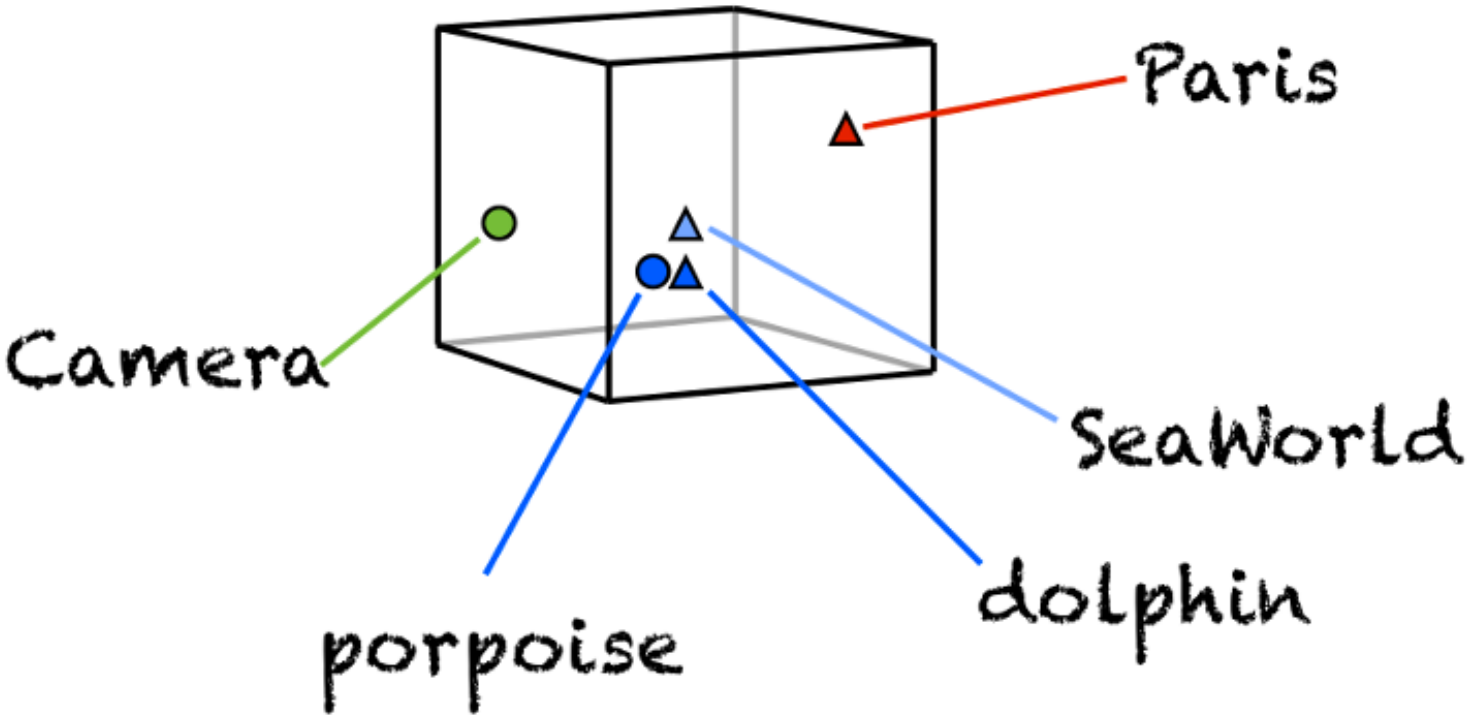


Semantic Scholar统计WordNet相关论文变化趋势

进入大数据和深度学习时代，HowNet等语言知识库还有什么用？

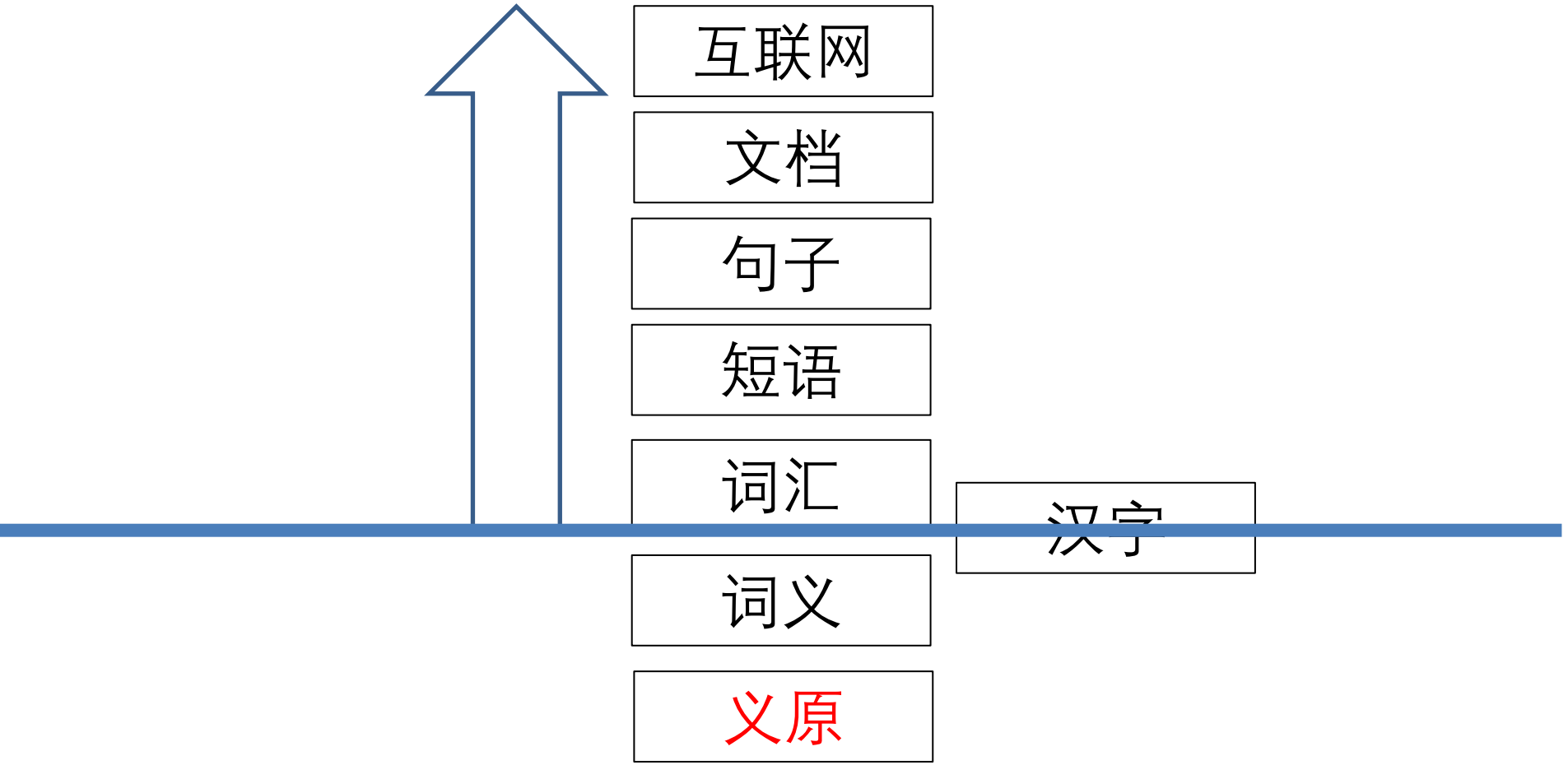
分布式表示缺点

- 可解释性和鲁棒性差



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

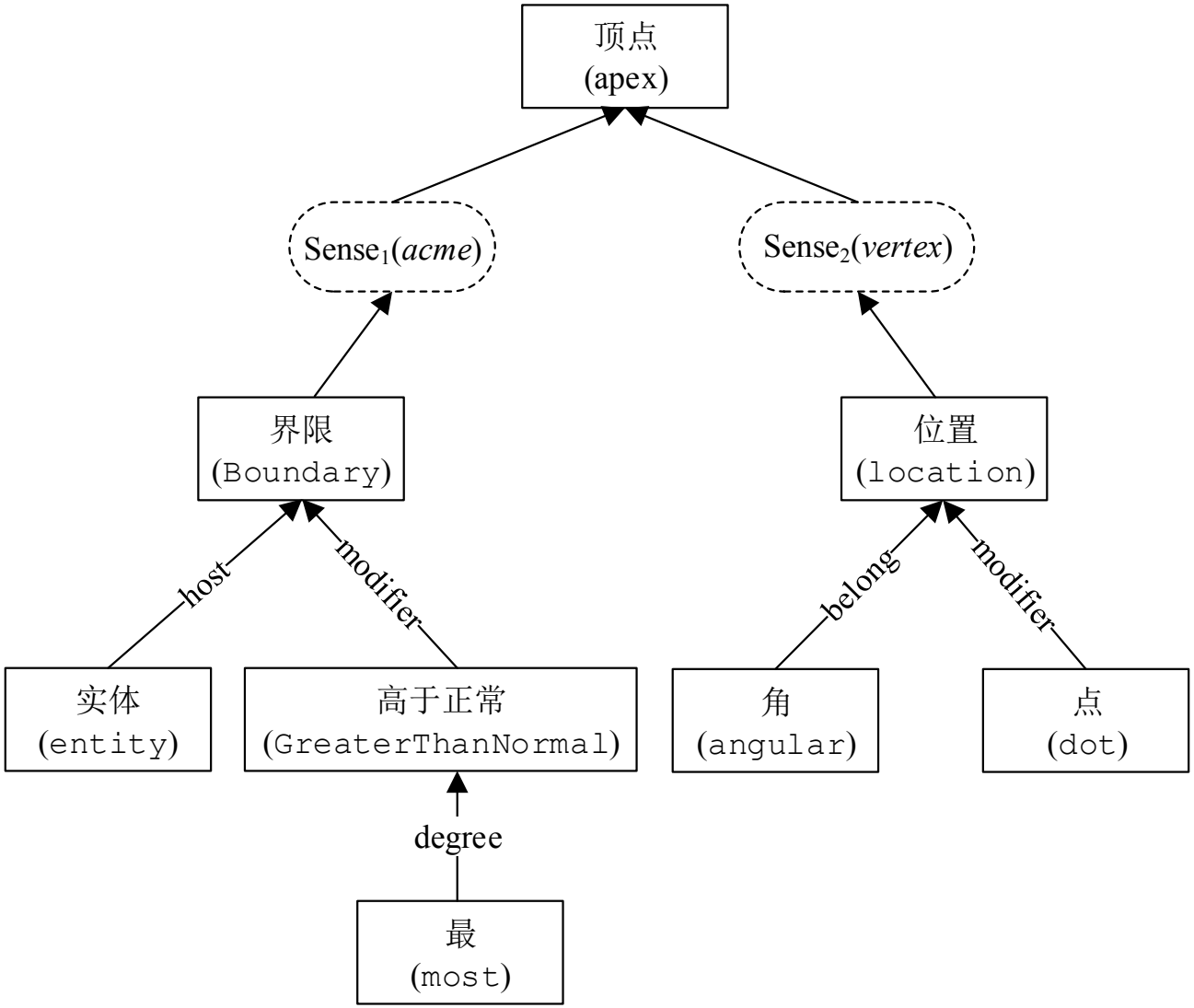
HowNet—瞥

- 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}}

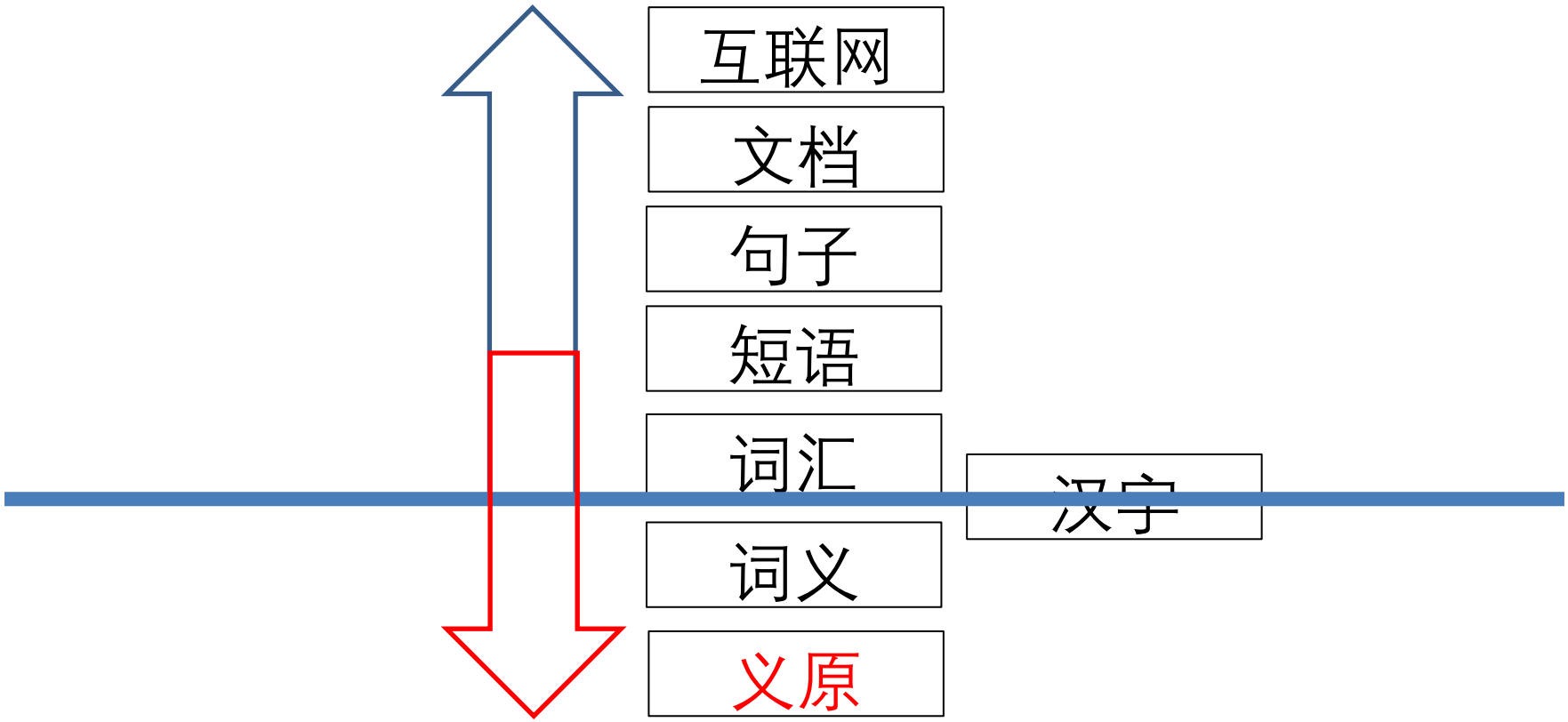
HowNet—瞥

- 义原知识带有层次结构



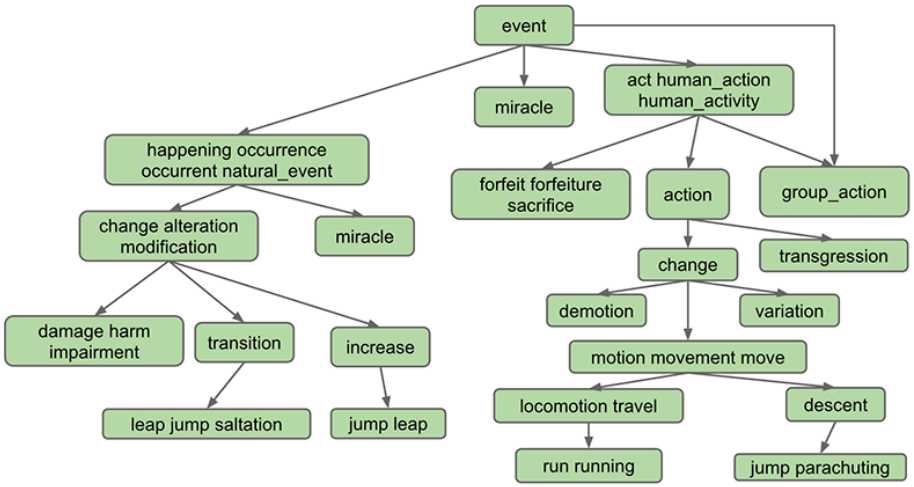
HowNet特点

- **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

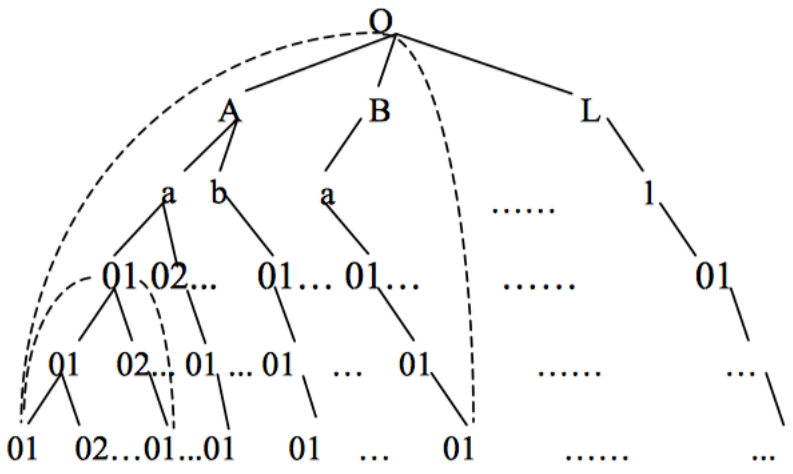


HowNet特点

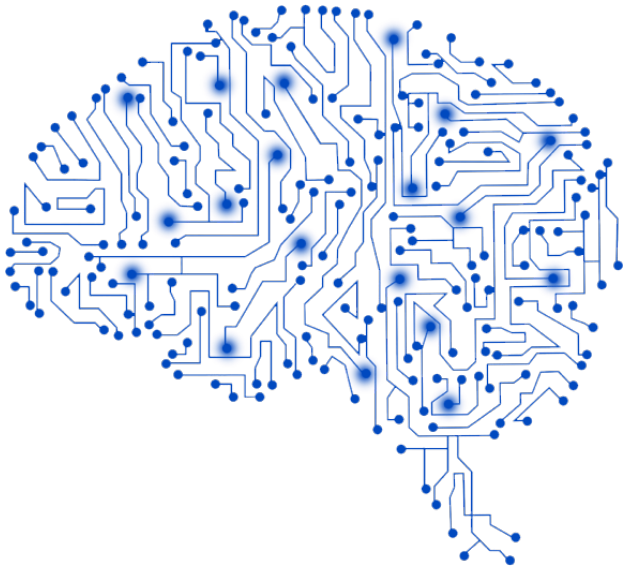
- **在融入深度学习方面，具有无可比拟优势**
  - 与WordNet、同义词词林等知识库组织模式不同
  - HowNet通过统一义原标注体系直接精准刻画语义信息。每个义原含义明确固定，可被直接作为语义标签融入机器学习模型



WordNet Synset体系



同义词词林层次类别体系

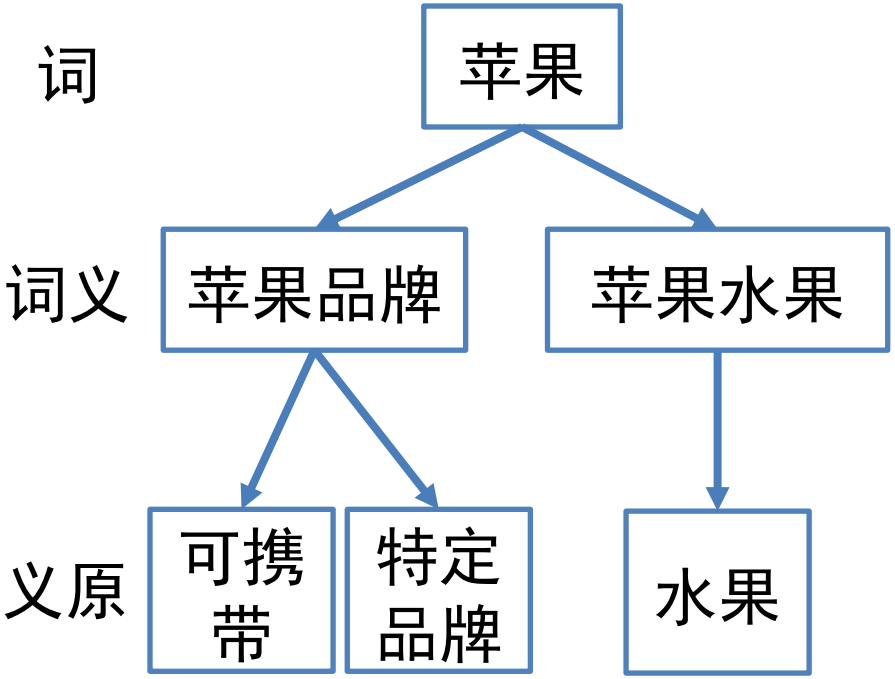


深度学习

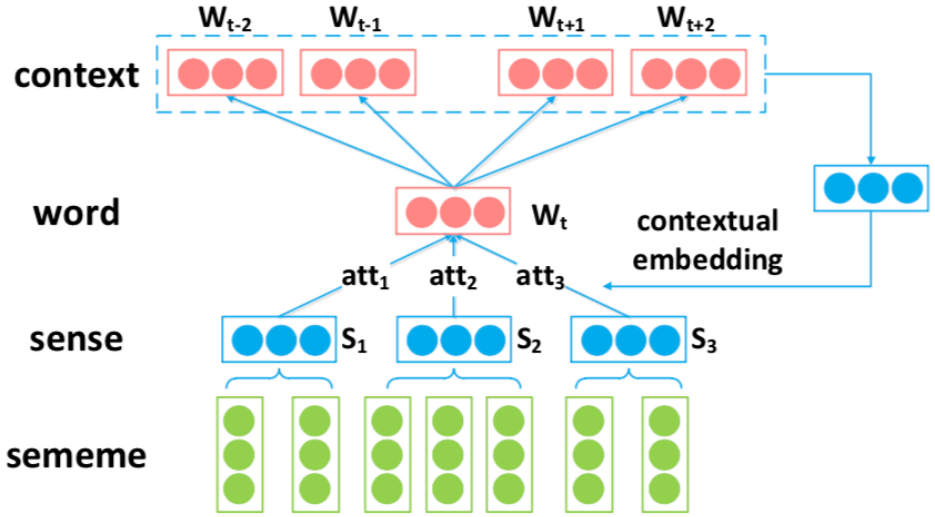
义原知识

融合义原知识的词义表示

- 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（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

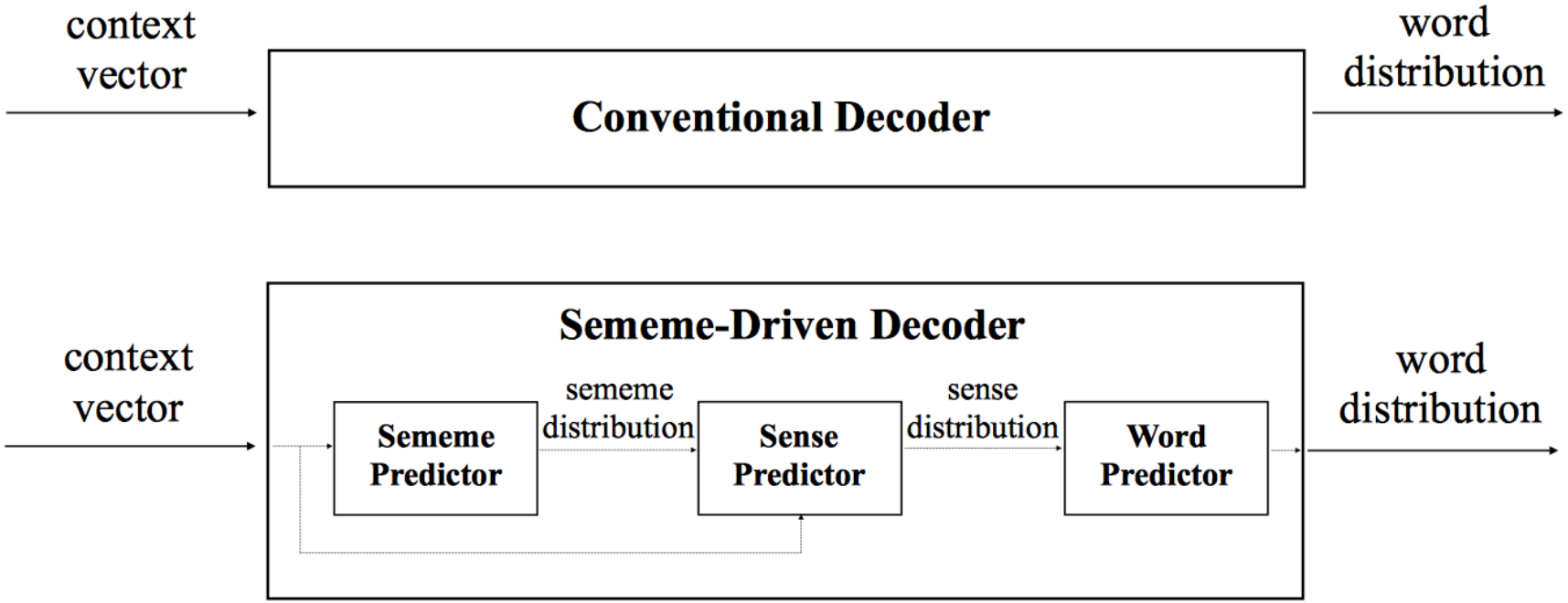
融合义原知识的神经语言模型

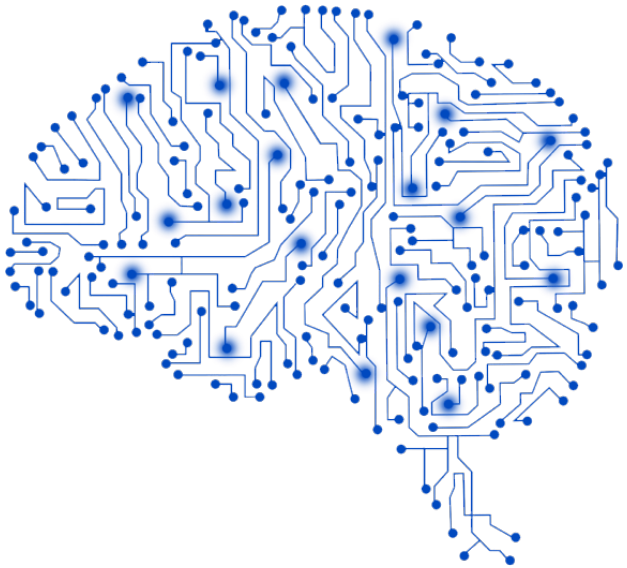
- 语言模型是自然语言处理的核心任务
- NGram是前深度学习时代的代表语言模型，深度学习框架CNN、RNN即用来学习语言模型
- 马尔科夫性：当前词出现的概率，依赖于上下文出现的词

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





深度学习

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）

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

(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

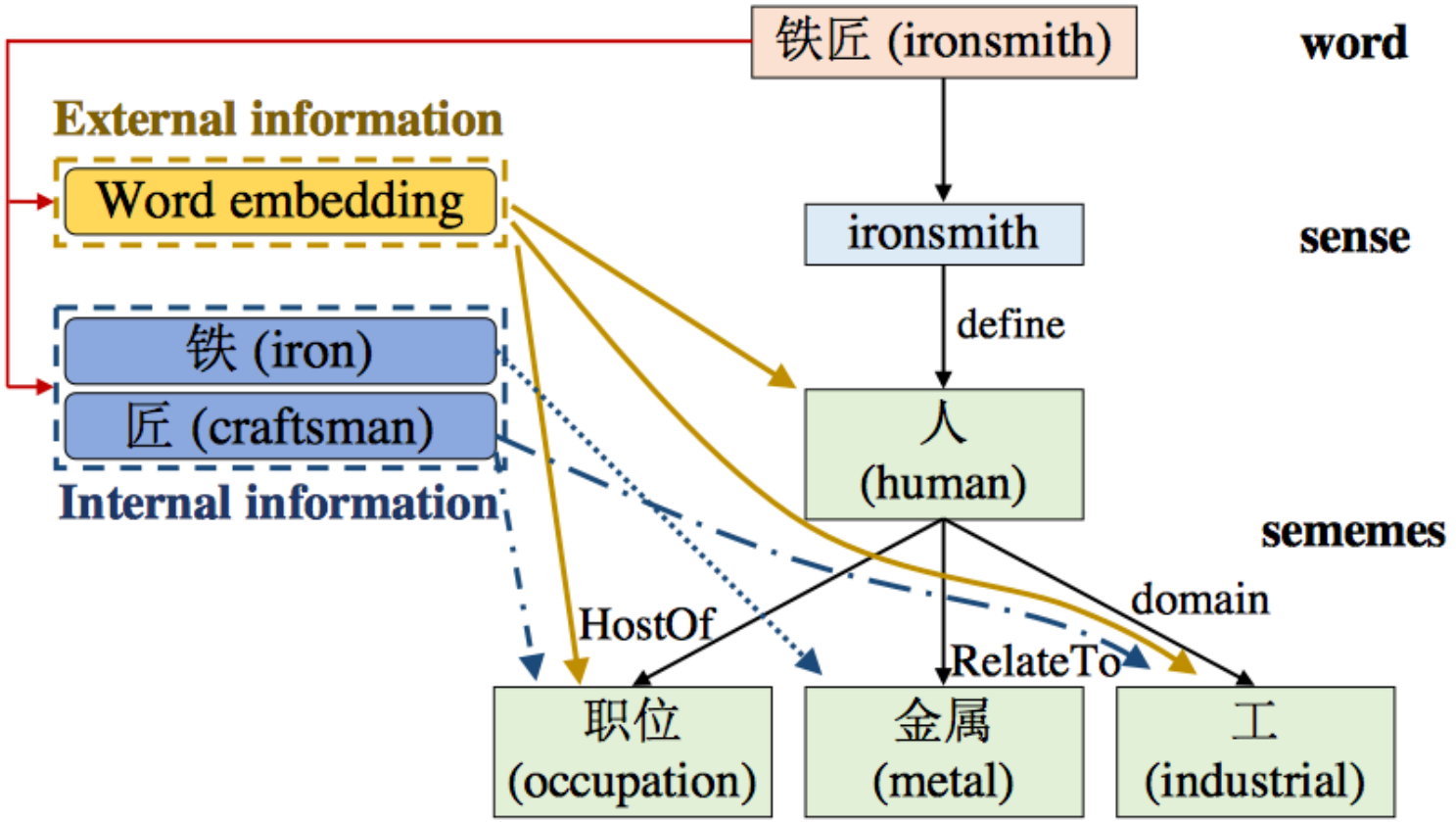
不同词频的词汇义原推荐效果

实验结果

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



(ACL 2018)

实验结果

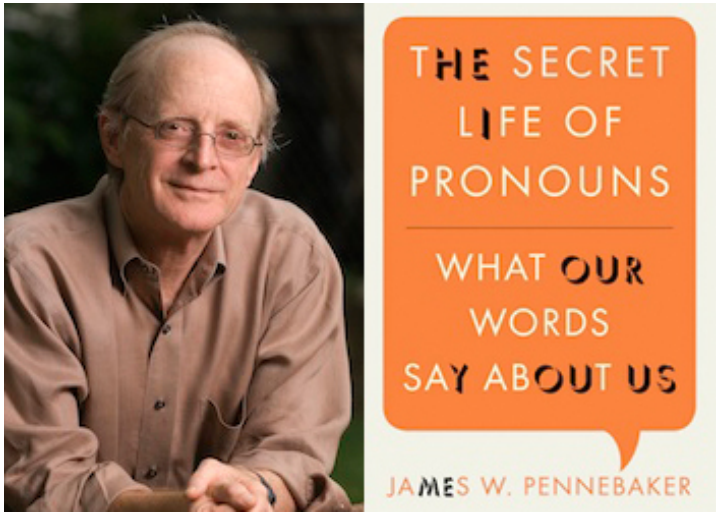
- 实验证明，考虑内部汉字信息，对于低频词的义原推荐提升尤为明显

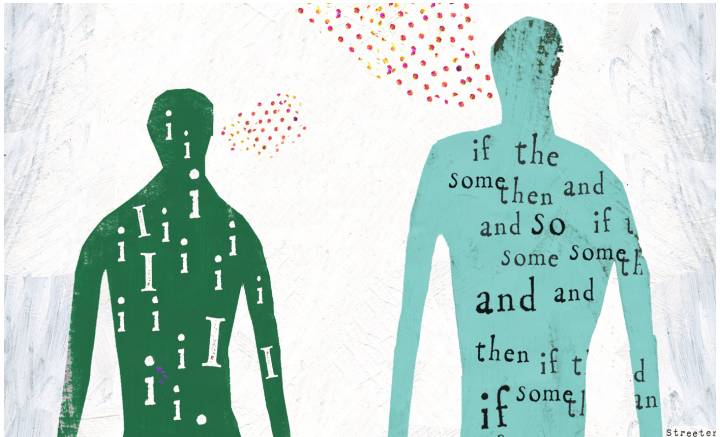
word frequency occurrences	⩽ 50 8537	51– 100 4868	101 – 1,000 3236	1,001 – 5,000 2036	5,001 – 10,000 663	10,001 – 30,000 753	>30,000 686
SPWE	0.312	0.437	0.481	0.558	0.549	0.556	0.509
SPSE	0.187	0.273	0.339	0.409	0.407	0.424	0.386
SPWE + SPSE	0.284	0.414	0.478	0.556	0.548	0.554	0.511
SPWCF	0.456	0.414	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**

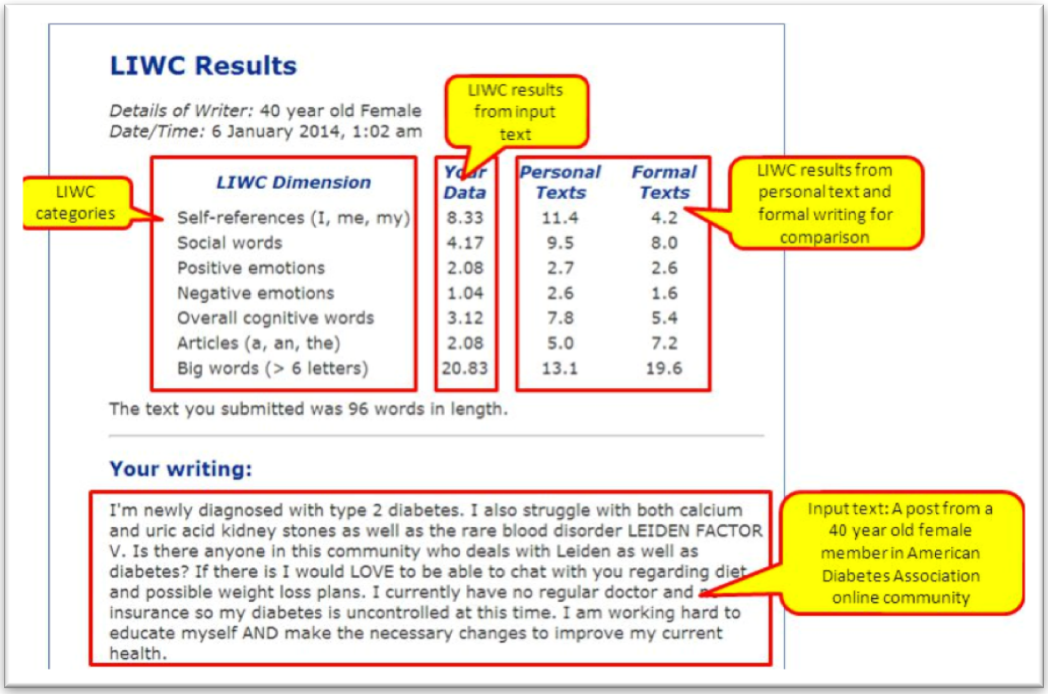
words	models	Top 5 sememes
钟表匠 (clockmaker)	internal	人**(human)**, 职位**(occupation)**, 部件(part), 时间**(time)**, 告诉**(tell)**
	external	人**(human)**, 专(ProperName), 地方(place), 欧洲(Europe), 政(politics)
	ensemble	人**(human)**, 职位**(occupation)**, 告诉**(tell)**, 时间**(time)**, 用具**(tool)**
奥斯卡 (Oscar)	internal	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)

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







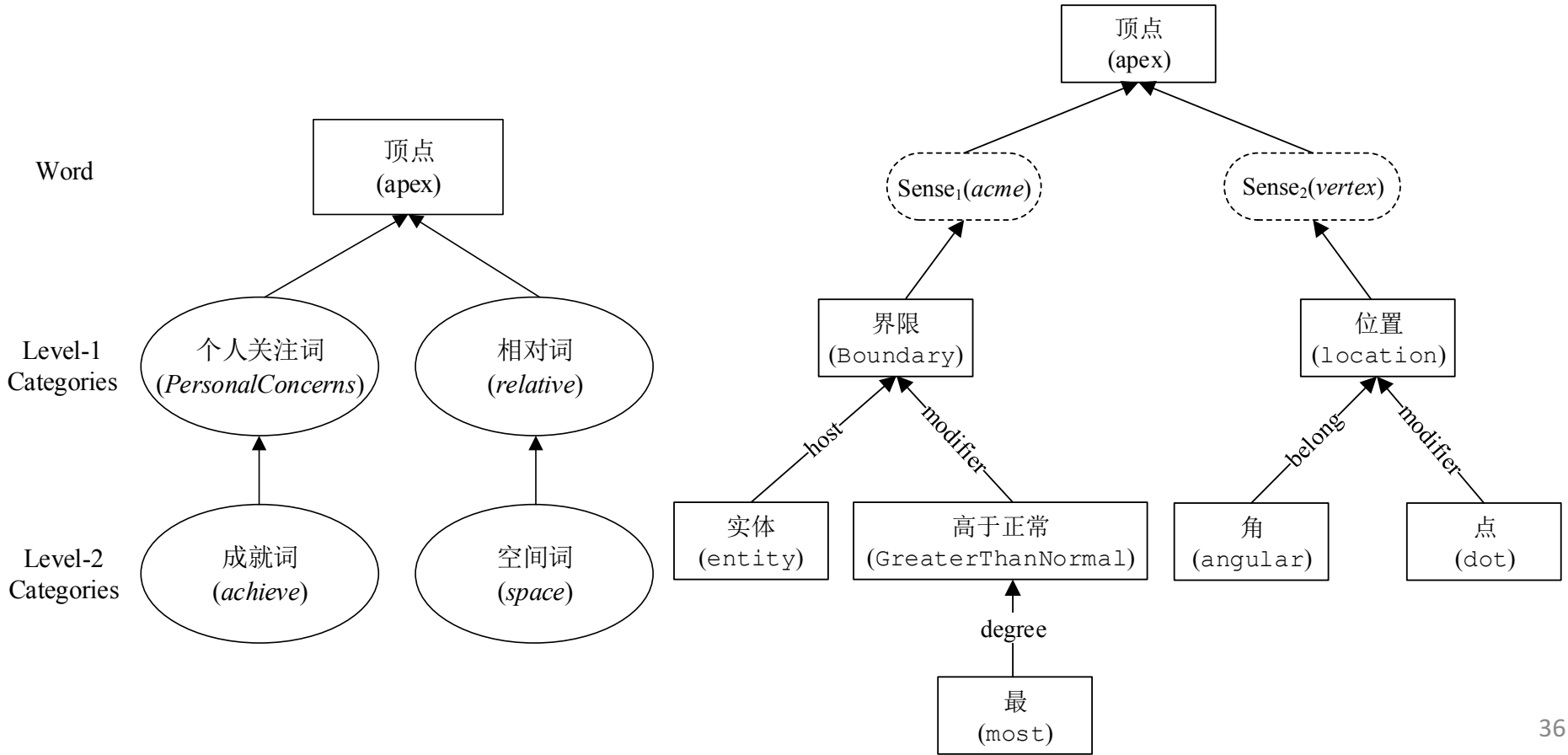
融合HowNet义原标注的词典扩展

- 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

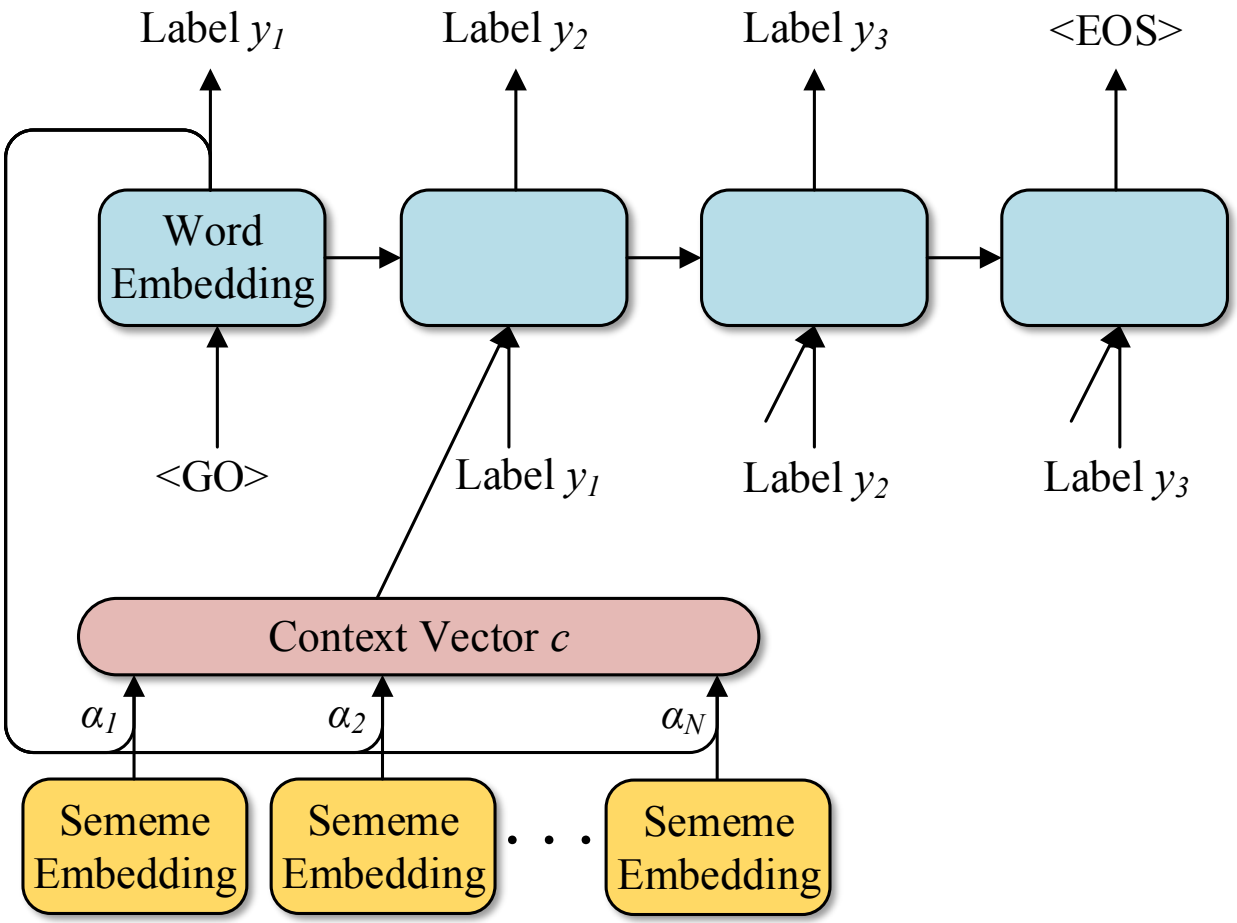
融合HowNet义原标注的词典扩展

- 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

- 在CLIWC词汇层次分类任务上，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**

实验结果

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

相关文献

- 下载地址：http://nlp.csai.tsinghua.edu.cn/~lzy/publication.html
- Huiming Jin, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, Leyu Lin. **Incorporating Chinese Characters of Words for Lexical Sememe Prediction**. The 56th Annual Meeting of the Association for Computational Linguistics (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**. The 32nd AAAI Conference on Artificial Intelligence (AAAI 2018).
- Ruobing Xie, Xingchi Yuan, Zhiyuan Liu, Maosong Sun. **Lexical Sememe Prediction via Word Embeddings and Matrix Factorization**. International Joint Conference on Artificial Intelligence (IJCAI 2017).
- Yilin Niu, Ruobing Xie, Zhiyuan Liu, Maosong Sun. **Improved Word Representation Learning with Sememes**. The 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017).

开源代码

- 义原语义计算、义原自动推荐等相关算法代码均在github.com开源

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
OpenNE：	网络表示学习
SE-WRL：	义原增强的词表示学习
Auto_CLIWC：	利用义原知识扩展LIWC词典
sememe_prediction：	义原自动推荐
Character-enhanced-Sememe-Prediction：	汉字增强的义原自动推荐

总结与展望

- 义原知识与HowNet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
- 义原知识亦可通过深度学习自然语言处理技术提升标注效率与一致性
- 随着义原知识的进一步扩充，有望显著改善深度学习可解释性与鲁棒性问题
- 在语言模型、义原推荐、词典扩充等具体任务上，仍有众多问题待深度探索
  - 义原结构利用、语言模型应用（翻译、对话）

感谢各位老师同学

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