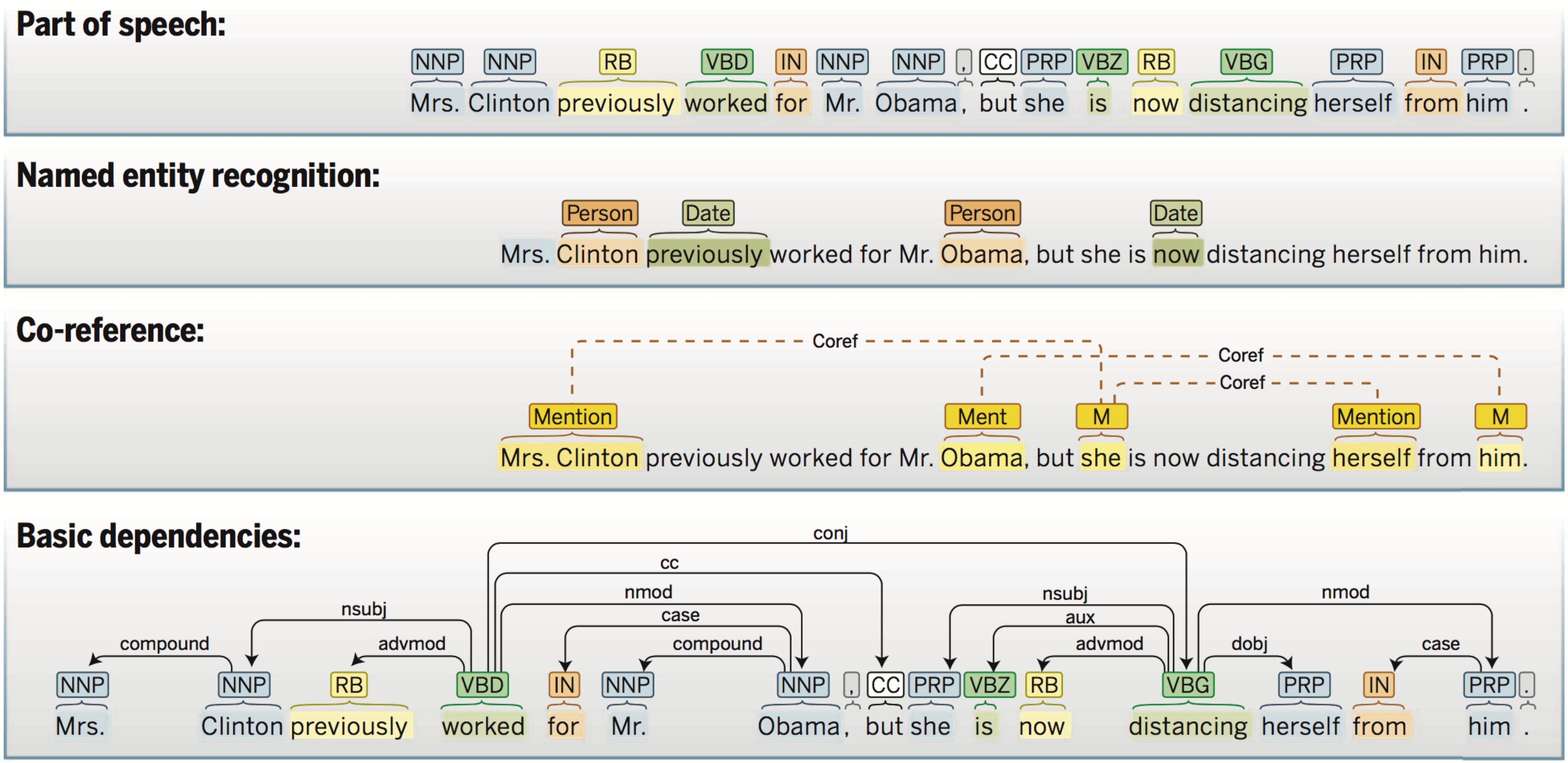
Knowledge-Guided NLP

Zhiyuan Liu
Tsinghua University

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

Natural Language Processing

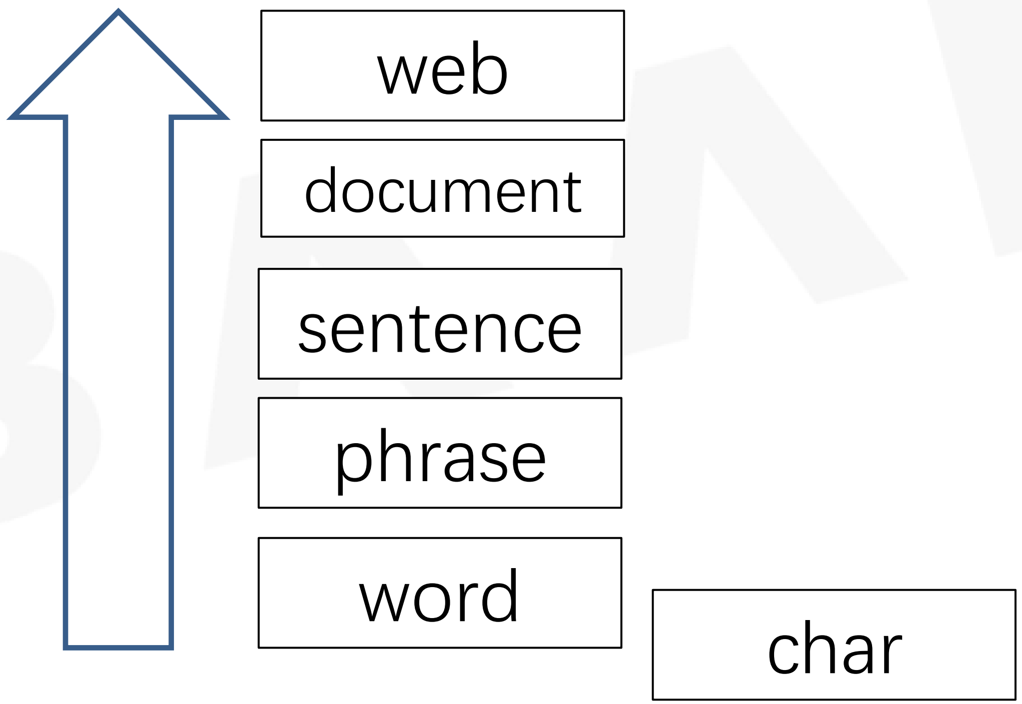
- NLP aims to make computers understand languages
- The nature of NLP is structure prediction



Advances in Natural Language Processing. Science 2015.

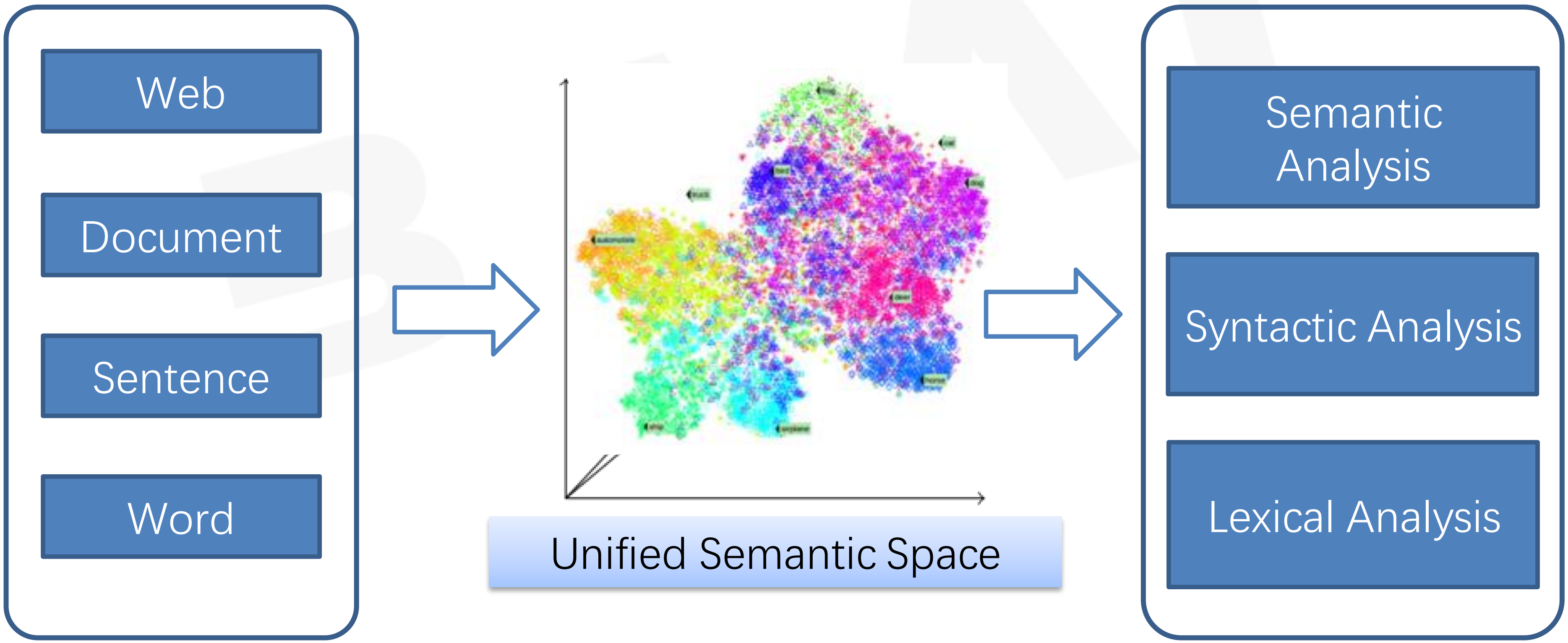
Characteristics of Natural Language - 1

- There are multiple-grained units in languages



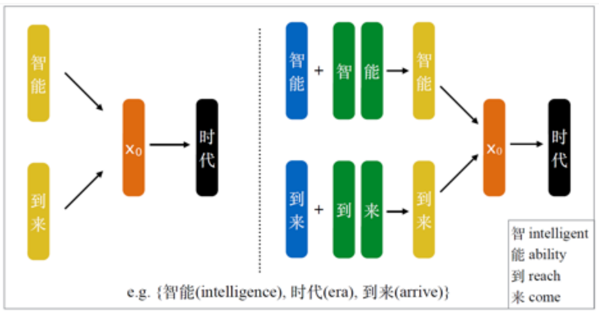
Distributed Representation

- Bridge the multiple-grained units in languages
- Alleviate the issue of data sparsity

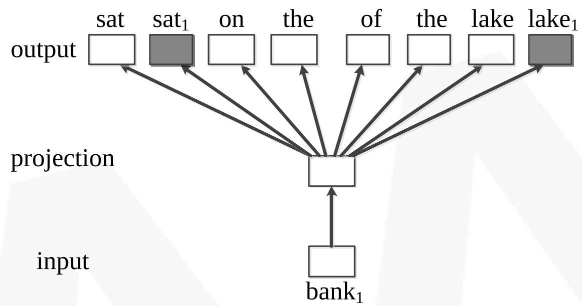


Language Representation Learning

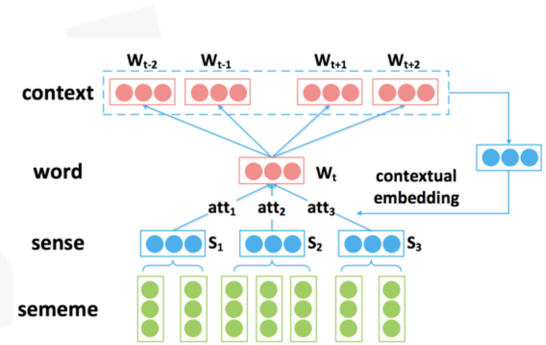
- Learn semantic representations of multi-grained language units



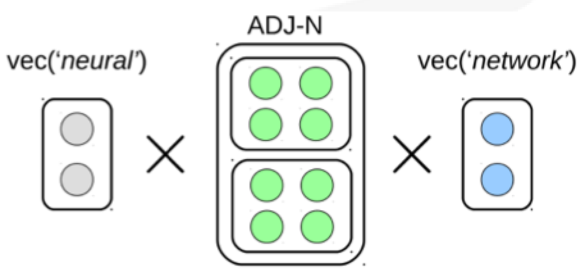
Character and Word Embedding (IJCAI 2015)



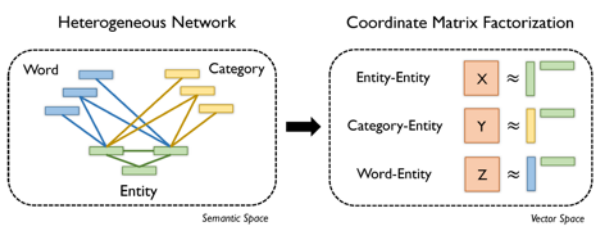
English Sense Embedding (EMNLP 2014)



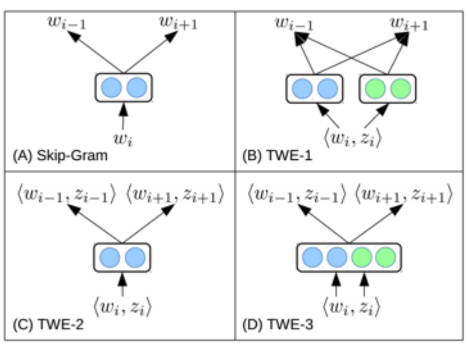
Chinese Sense Embedding (ACL 2017)



Phrase Embedding (AAAI 2015)



Entity Embedding (IJCAI 2015)



Document Embedding (IJCAI 2015)

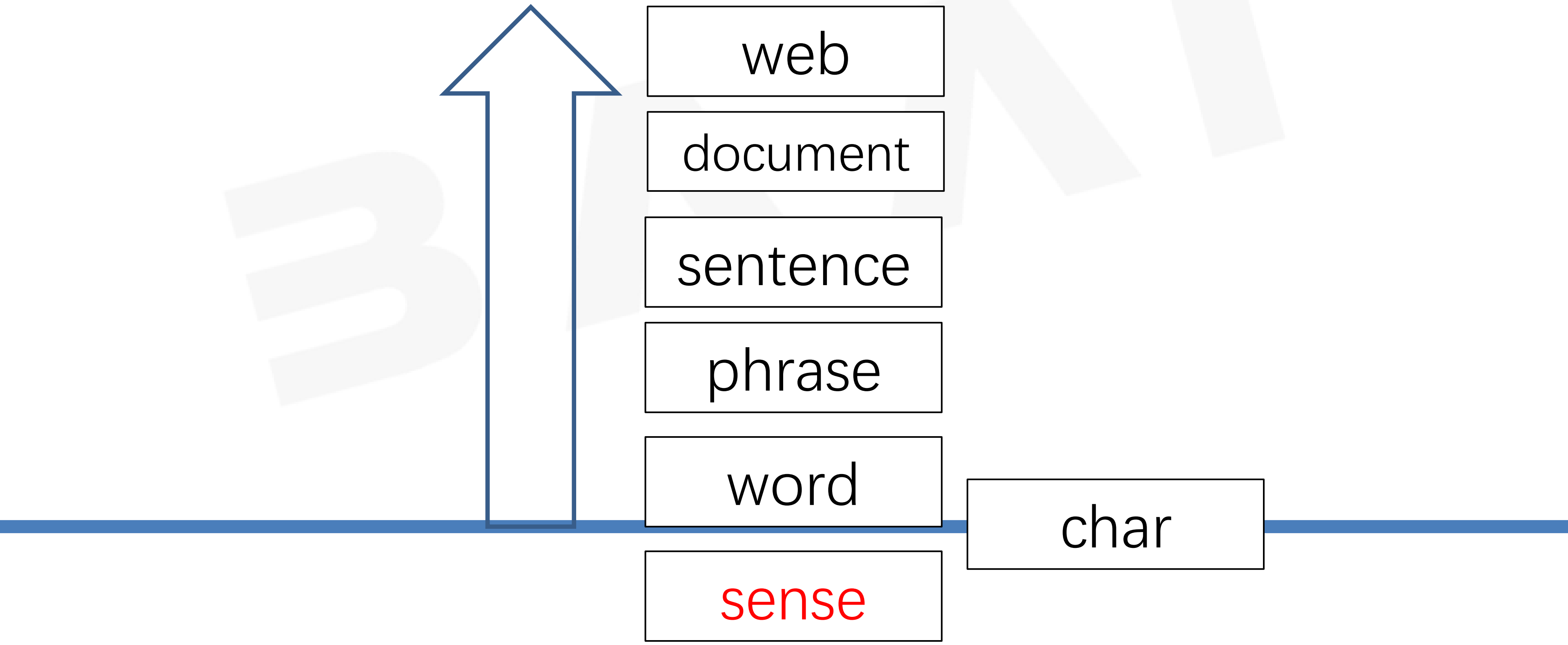
Challenges of DL for NLU & NLP

… 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 <u>difficult problems of semantics, context, and knowledge</u> will probably require new discoveries in linguistics and inference.

Advances in Natural Language Processing. Science 2015.

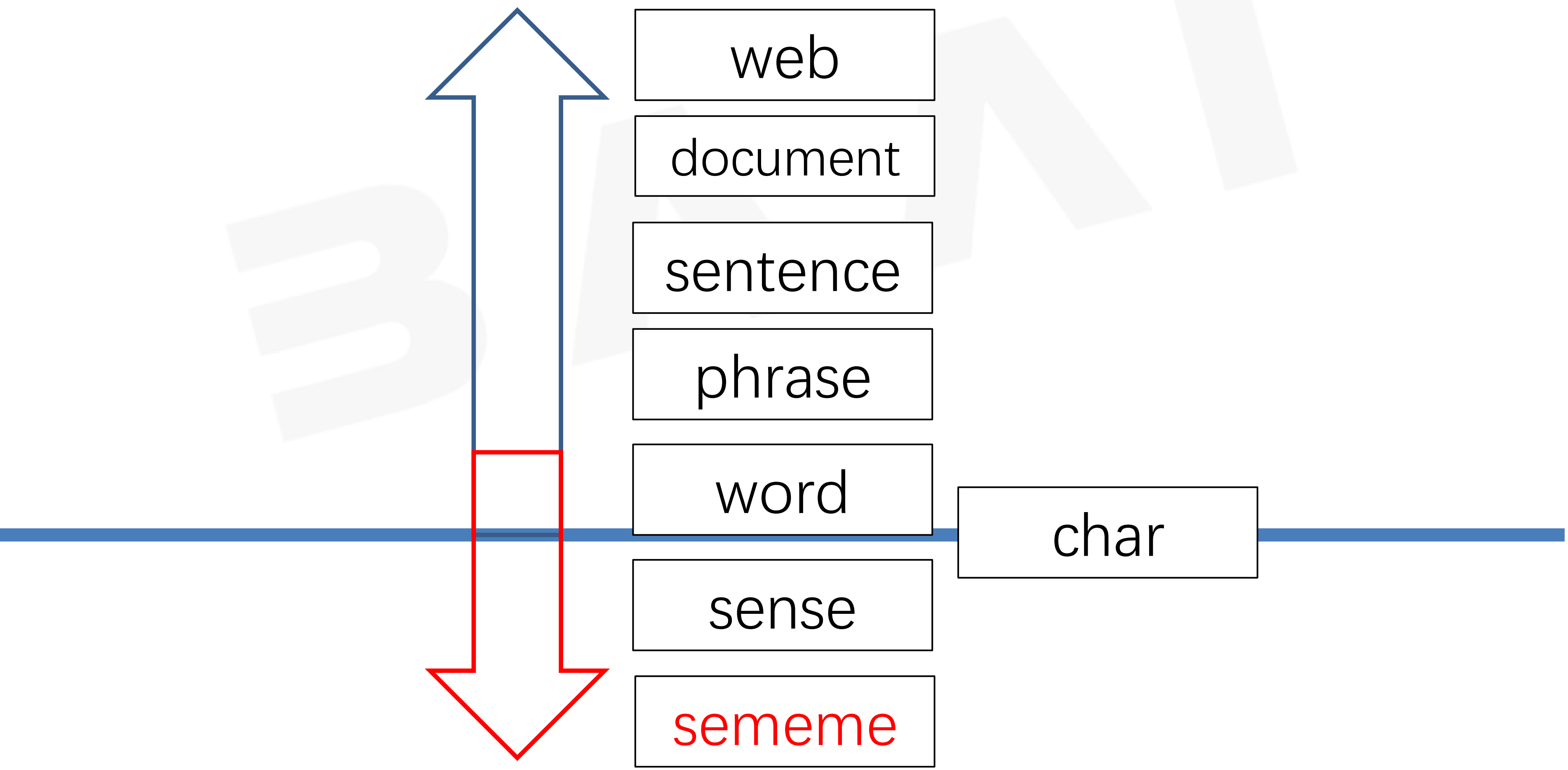
Characteristics of Natural Language - 2

- Words/Chinese characters are minimal units of usages, but not minimal units of semantics



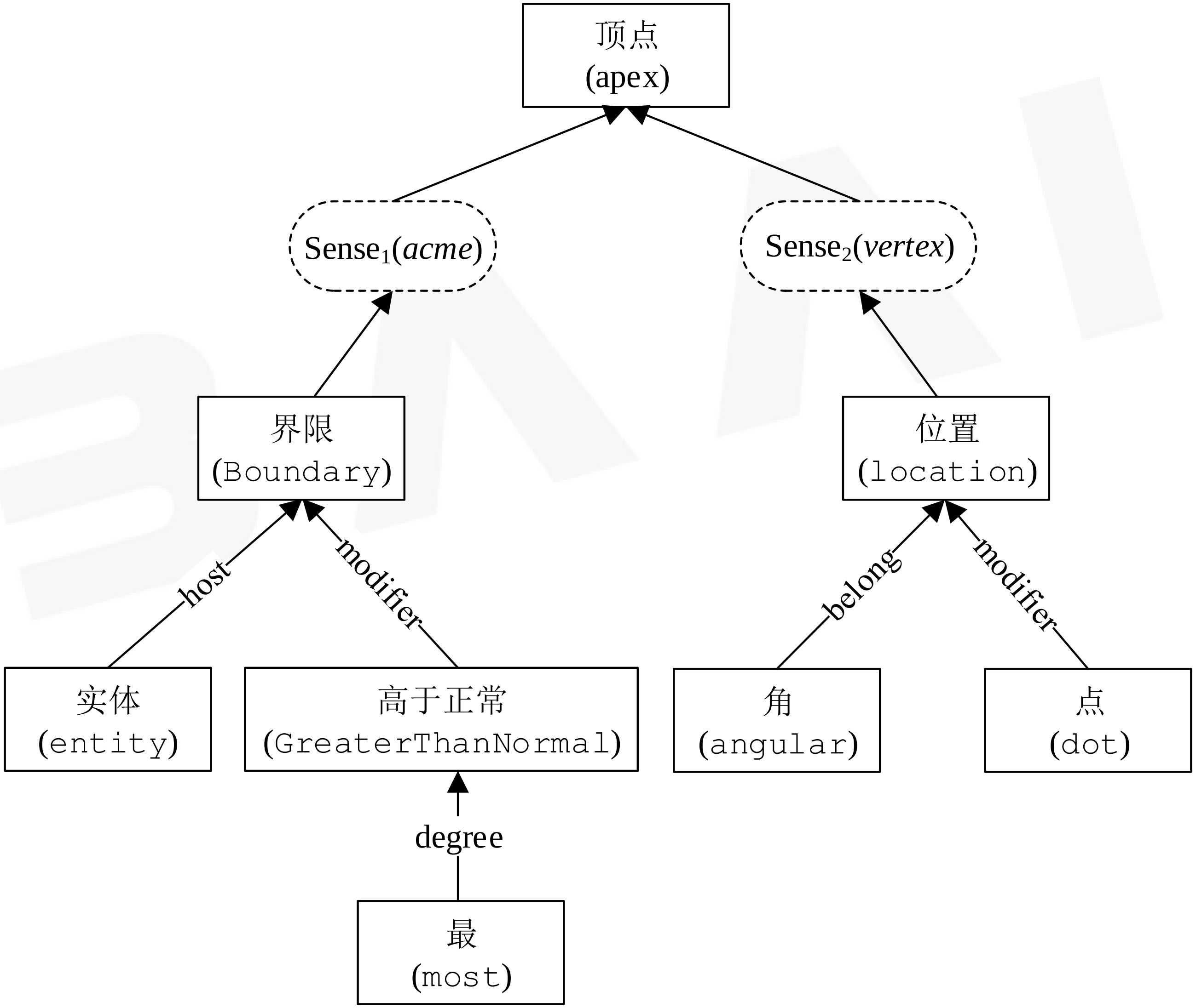
Use Sememes to Break Word Boundary

- Lexical sememes: minimal units of semantics



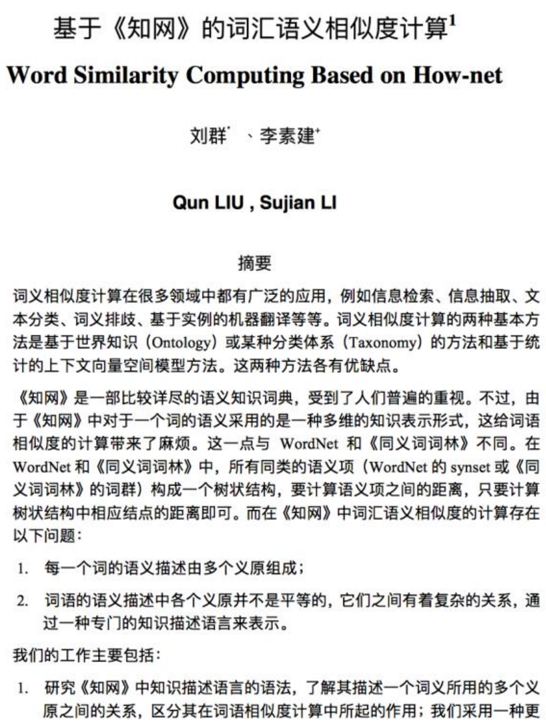
Linguistic Knowledge with Lexical Sememes

- Lexical sememes: minimal units of semantics



HowNet

- Linguistic knowledge base of lexical sememes, released in 1999
- Manually create ~2,000 sememes
- Manually annotate ~100,000 words with sememes



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：

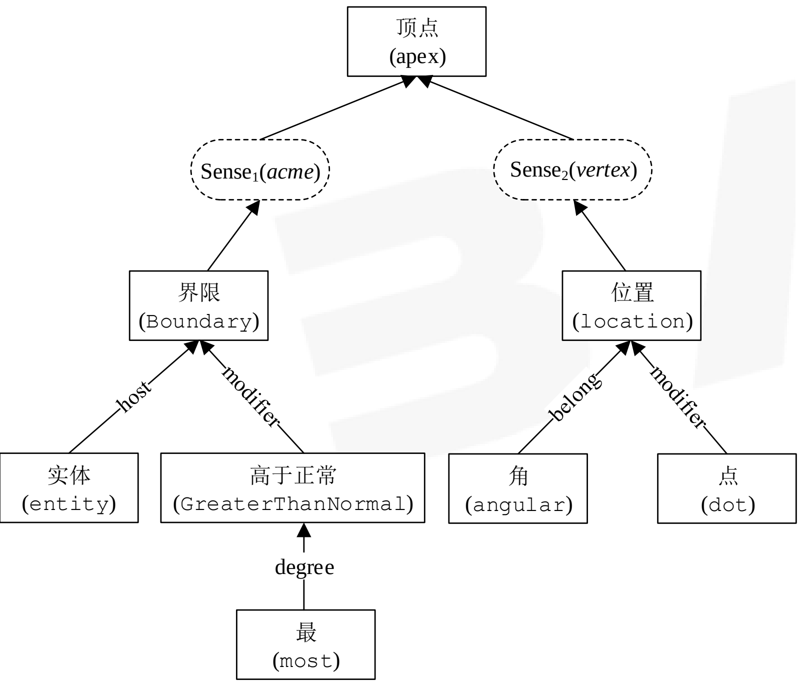
1. 每一个词的语义描述由多个义原组成；
2. 词语的语义描述中各个义原并不是平等的，它们之间有着复杂的关系，通过一种专门的知识描述语言来表示。

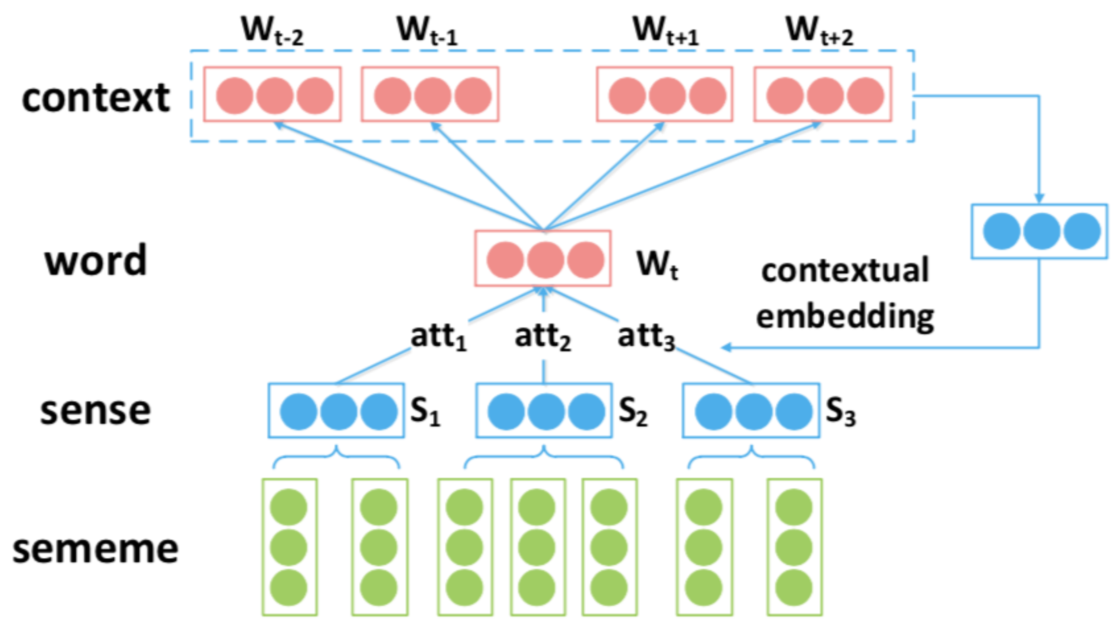
我们的工作主要包括：

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

Sememe-Guided Word Embedding

- Incorporate sense-sememe knowledge into word embeddings





Sememe-Sense-Word Joint Model

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

Experiment Results

- The enhanced word embeddings perform better on the tasks of analogy reasoning and word similarity

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**

Experiment Examples

- The model can conduct sense disambiguation based on sememes and contexts

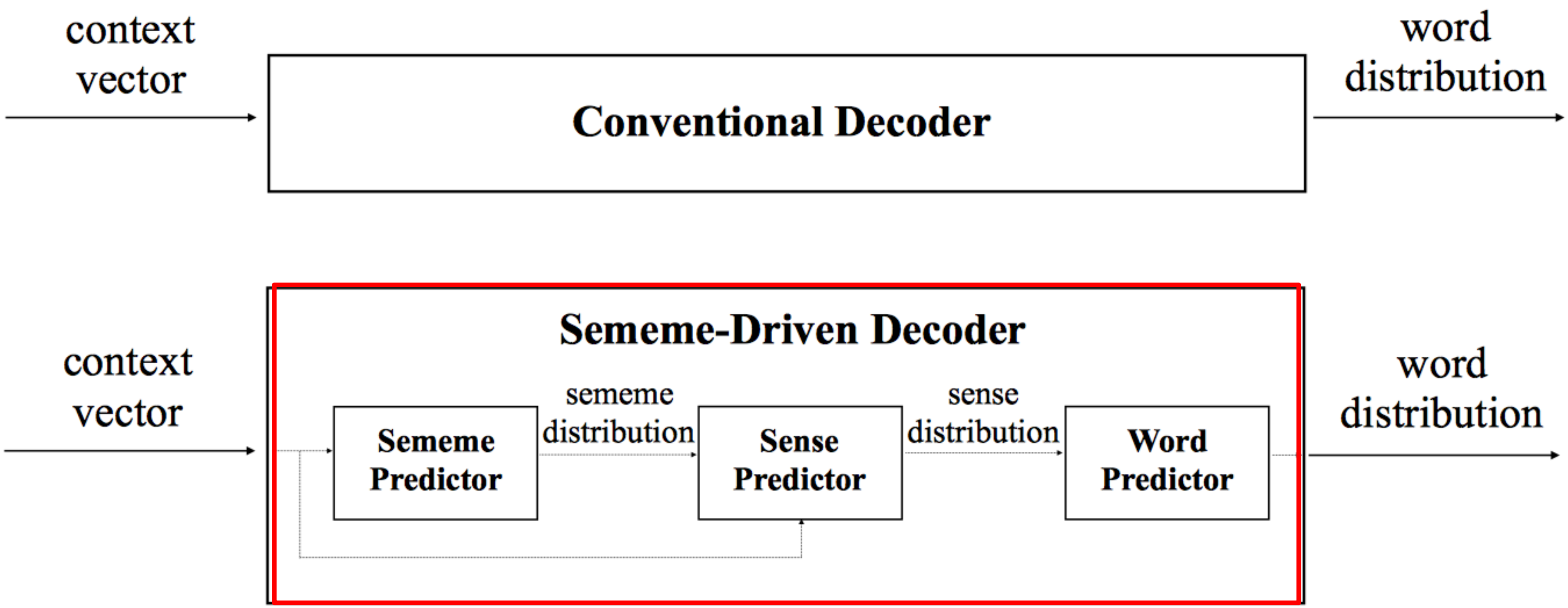
Word: 苹果("Apple brand/apple") sense1: *Apple brand* (computer, PatternValue, able, bring, SpeBrand) sense2: *duct* (fruit)		
苹果 素有果中王美称（**Apple** is always famous as the king of fruits） **苹果** 电脑无法正常启动（The **Apple brand** computer can not startup normally）	*Apple brand*: 0.28 *Apple brand*: 0.87	*apple*: 0.72 *apple*: 0.13
Word: 扩散("proliferate/metastasize") sense1: *proliferate* (disperse) sense2: *metastasize* (disperse, disease)		
防止疫情**扩散**（Prevent epidemic from **metastasizing**） 不**扩散**核武器条约（Treaty on the Non-**Proliferation** of Nuclear Weapons）	*proliferate*: 0.06 *proliferate*: 0.68	*metastasize*: 0.94 *metastasize*: 0.32
Word: 队伍("contingent/troops") sense1: *contingent* (community) sense2: *troops* (army)		
八支**队伍**进入第二阶段团体赛（Eight **contingents** enter the second stage of team competition） 公安基层**队伍**组织建设（Construct the organization of public security's **troops** in grass-roots unit）	*contingent*: 0.90 *contingent*: 0.15	*troops*: 0.10 *troops*: 0.85

Sememe-Guided Language Modeling

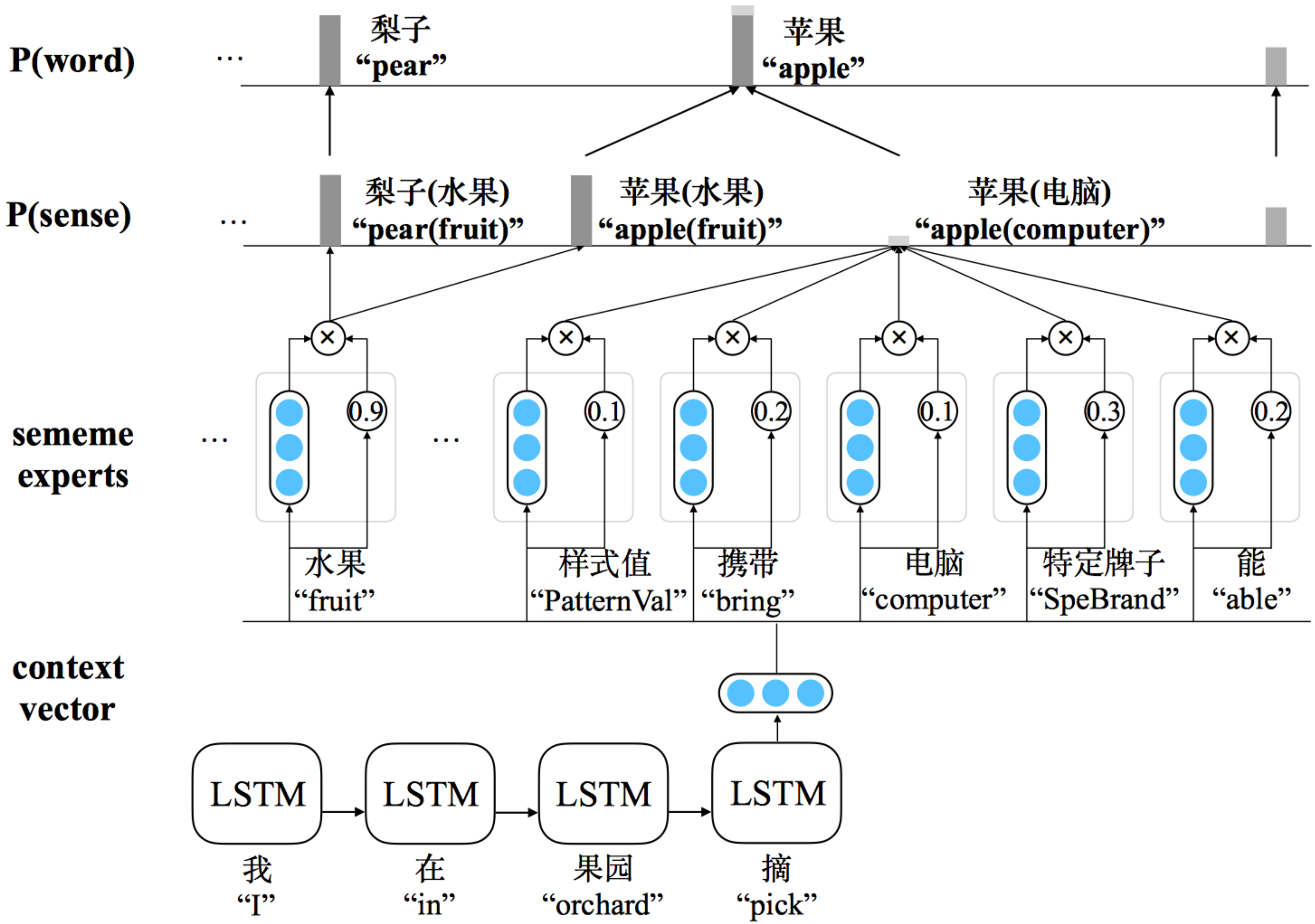
- Modeling word sequence with Markov property

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

- Sememe-Guided Language Modeling



Sememe-Guided Language Modeling



Experiment Results

- Sememe knowledge can significantly reduce the perplexity of language models

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**

Experiment Examples

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"	

Semantic Composition

农民 (peasant)

起义 (uprising)

农民起义 (peasant uprising)

画 (draw)

句号 (a period)

画句号 (draw a period)

Sememe-Guided SC Modeling

- A preliminary experiment of semantic composition degree of Multi-word Expressions (MWEs)

SCD	Our Computation Formulae	Examples	
		MWEs and Constituents	Sememes
3	$S_p = S_{w_1} \cup S_{w_2}$	农民起义(peasant uprising) 农民 (peasant) 起义(uprising)	事情|fact,职位|occupation,政|politics,暴动|uprise,人|human,农|agricultural 职位|**occupation**,人|**human**,农|**agricultural** 暴动|**uprise**,事情|**fact**,政|**politics**
2	$S_p \subsetneq (S_{w_1} \cup S_{w_2})$	几何图形(geometric figure) 几何 (geometry; how much) 图形(figure)	数学|math,图像|image 数学|**math**,知识|knowledge,疑问|question,功能词|funcword 图像|**image**
1	$S_p \cap (S_{w_1} \cup S_{w_2}) \neq \emptyset$ $\wedge\ S_p \not\subset (S_{w_1} \cup S_{w_2})$	应考(engage a test) 应 (deal with; echo; agree) 考(quiz; check)	考试|exam,从事|engage 处理|handle,回应|respond,同意|agree,遵循|obey,功能词|funcword,姓|surname 考试|**exam**,查|check
0	$S_p \cap (S_{w_1} \cup S_{w_2}) = \emptyset$	画句号(end) 画 (draw) 句号(period)	完毕|finish 画|draw,部件|part,图像|image, 文字|character,表示|express 符号|symbol,语文|text

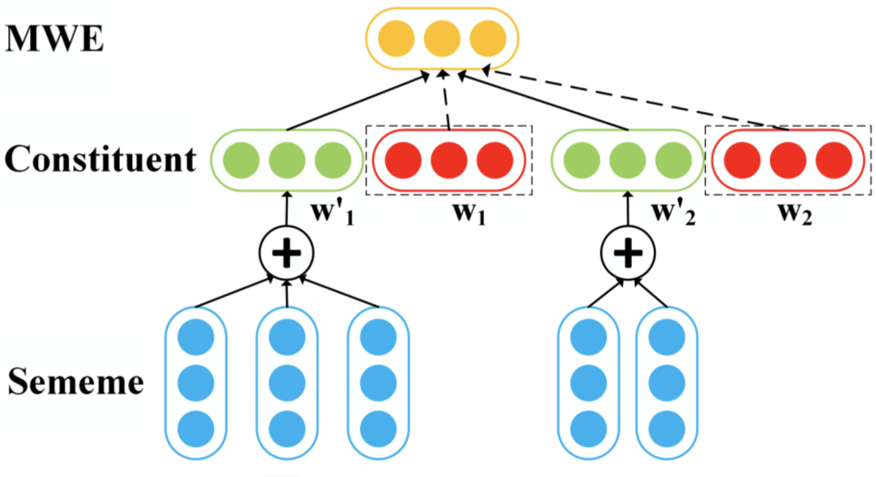
S_p, S_{w_1} and S_{w_2} : sememe sets of an MWE, its first constituent and second constituent.

Pearson's correlation with human evaluation: **0.75**

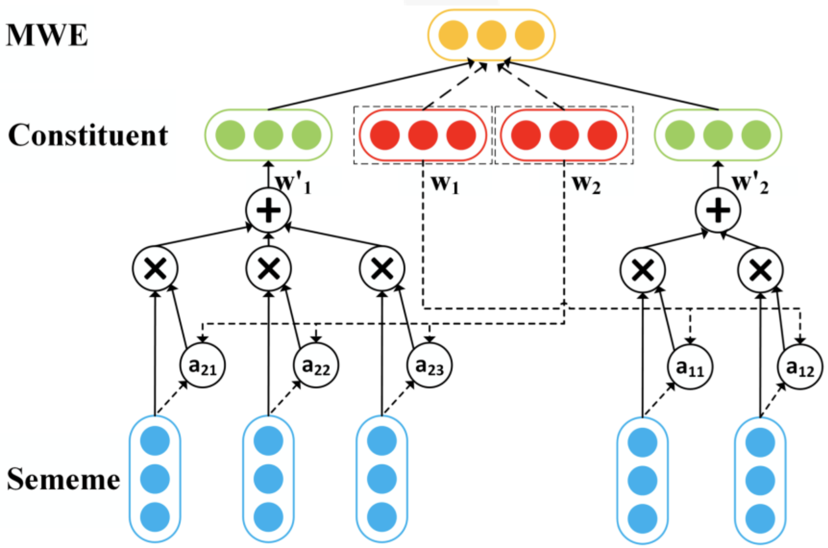
Fanchao Qi, Junjie Huang, Chenghao Yang, Zhiyuan Liu, Xiao Chen, Qun Liu, Maosong Sun. Modeling Semantic Compositionality with Sememe Knowledge. ACL 2019.

Sememe-Guided SC Modeling

- Sememe-incorporated SC models



SC with Aggregated Sememe Model (**SC-AS**)



SC with Mutual Sememe Attention Model (**SC-MSA**)

Fanchao Qi, Junjie Huang, Chenghao Yang, Zhiyuan Liu, Xiao Chen, Qun Liu, Maosong Sun. Modeling Semantic Compositionality with Sememe Knowledge. ACL 2019.

Experiment Results

Framework	Method	WS240	WS297	COS960
$f(\mathbf{w}_1, \mathbf{w}_2)$	ADD	50.8	53.1	49.1
	MUL	19.6	21.6	−3.9
	TIM	47.4	54.2	50.5
	RNTN	42.5	53.6	55.8
	RAE	61.3	59.9	59.6
	SCAS-S	61.4	57.0	60.1
$f(\mathbf{w}_1, \mathbf{w}_2, K)$	SCAS	60.2	60.5	61.4
	SCMSA	**61.9**	58.7	60.5
$f(\mathbf{w}_1, \mathbf{w}_2, K, R)$	SCAS+R	59.0	60.8	**61.8**
	SCMSA+R	61.4	**61.2**	60.4

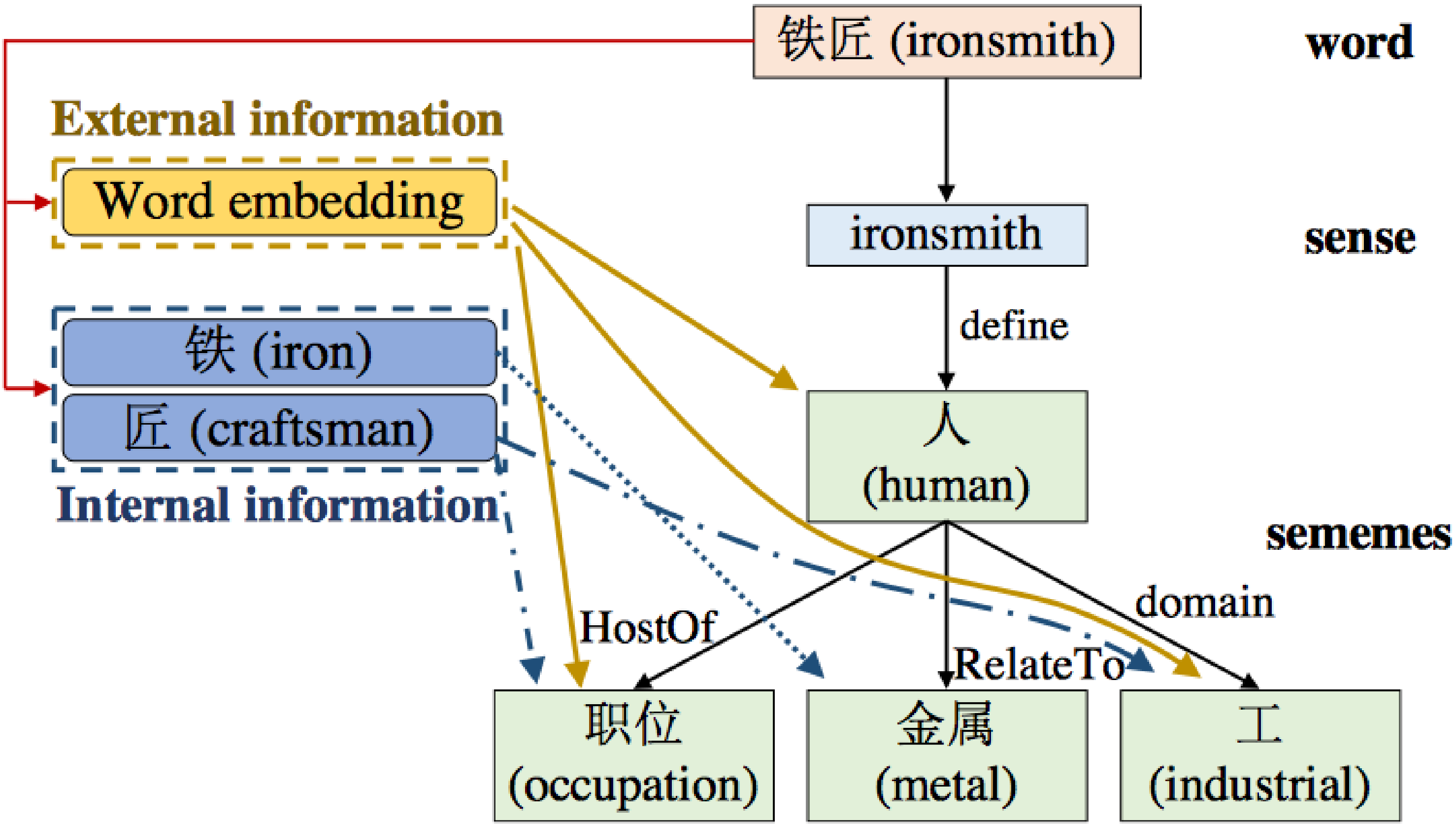
Intrinsic Evaluation
(MWE Similarity)

Framework	Method	Sememe Prediction	
		MAP	F1 Score
$f(\mathbf{w}_1, \mathbf{w}_2)$	ADD	40.7	23.2
	MUL	11.2	0.3
	TIM	46.8	35.3
	RNTN	47.7	35.3
	RAE	44.0	30.8
	SCAS-S	39.0	27.9
$f(\mathbf{w}_1, \mathbf{w}_2, K)$	SCAS	52.2	41.3
	SCMSA	55.1	43.4
$f(\mathbf{w}_1, \mathbf{w}_2, K, R)$	SCAS+R	56.8	**46.1**
	SCMSA+R	**58.3**	46.0

Extrinsic Evaluation
(MWE Sememe Prediction)

Sememe Prediction

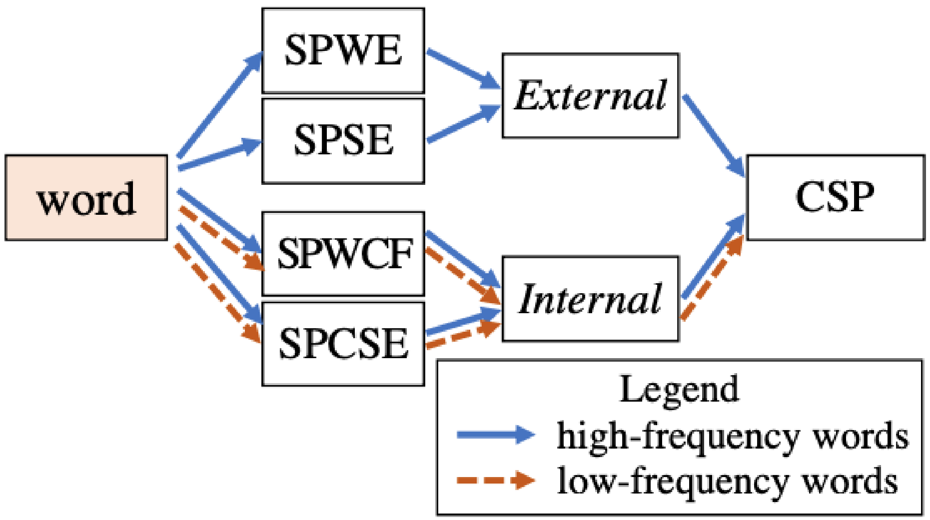
- Use both external and internal information to predict sememes



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.

Experiment Results

- We propose several models for sememe prediction with either internal and external information

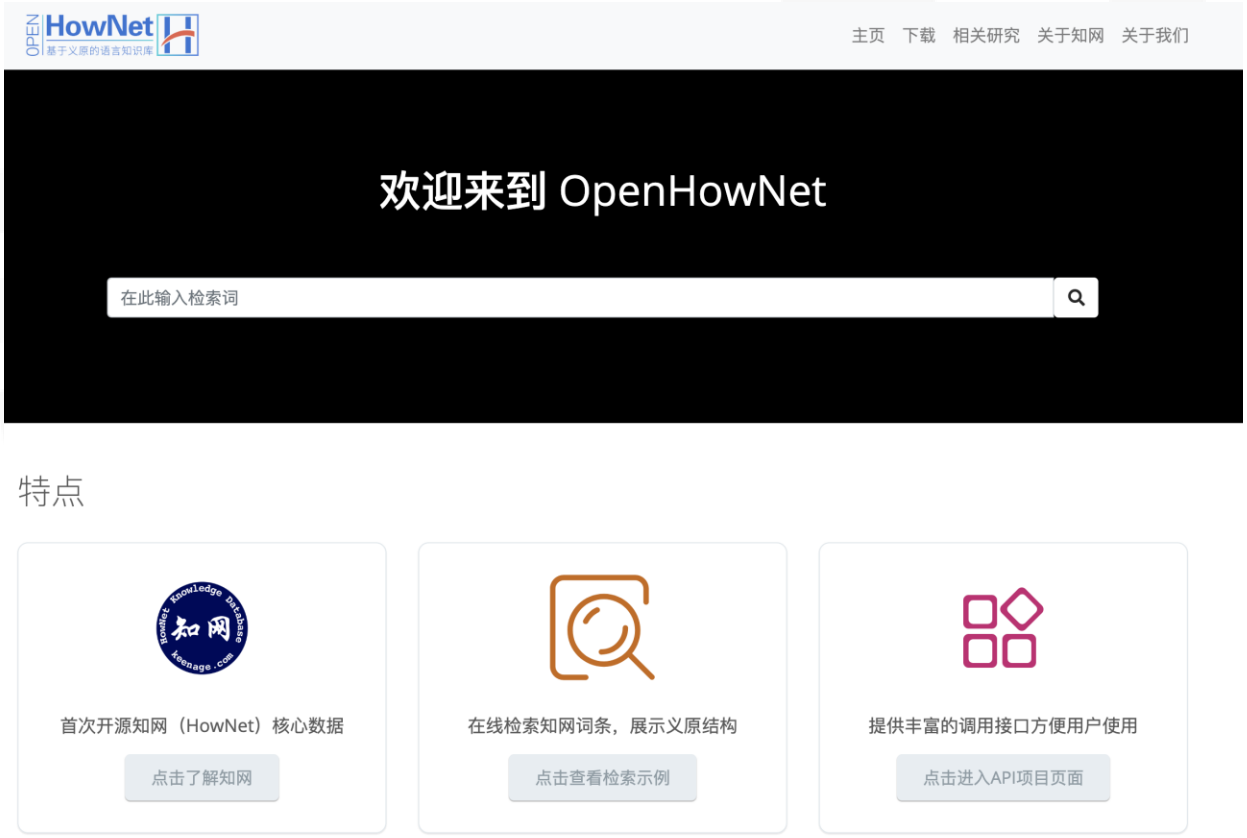


Method	MAP
SPSE	0.411
SPWE	0.565
SPWE+SPSE	0.577
SPWCF	0.467
SPCSE	0.331
SPWCF + SPCSE	**0.483**
SPWE + fastText	0.531
CSP	**0.654**

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)

OpenHowNet

https://openhownet.thunlp.org/



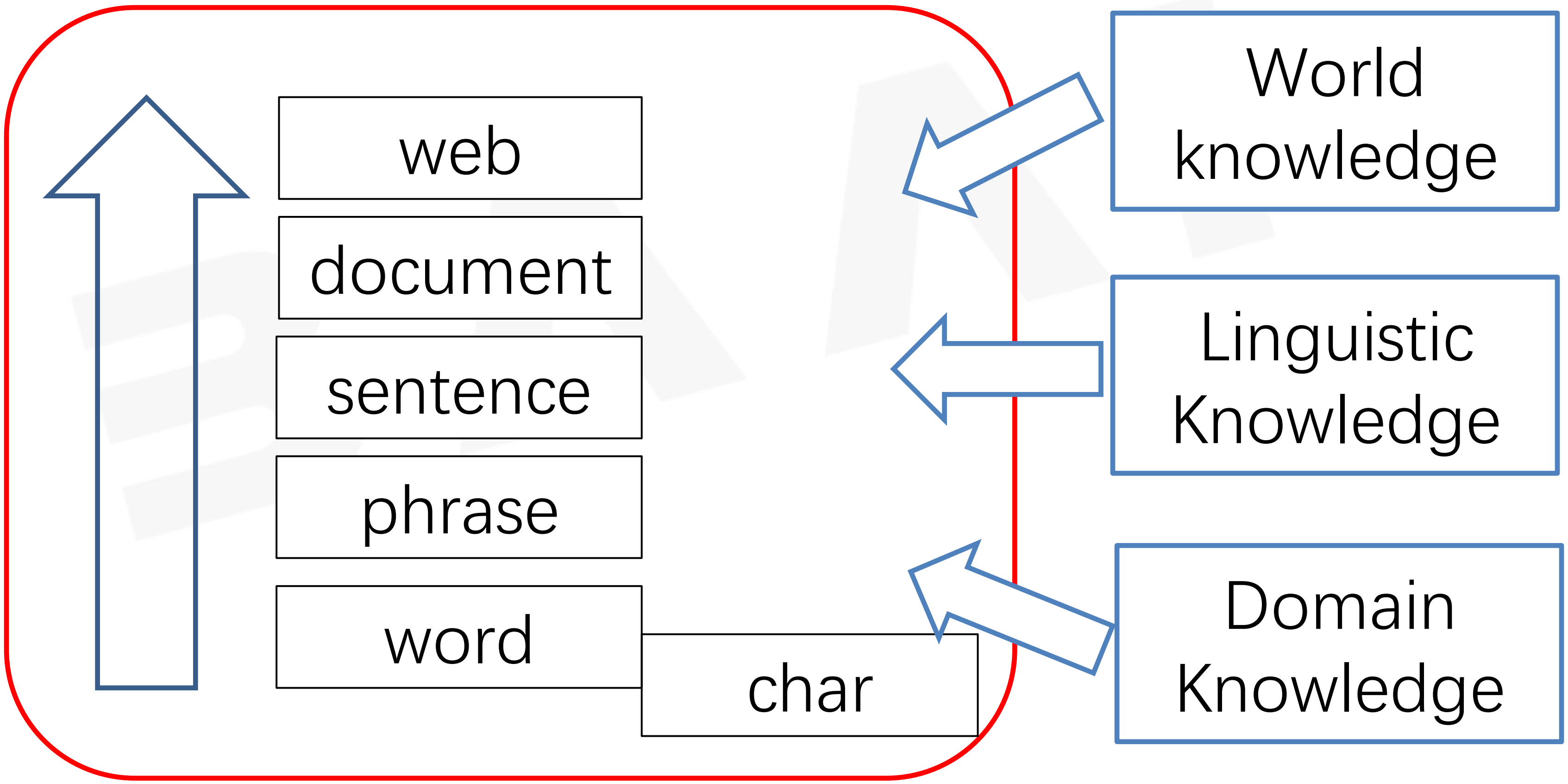
Sememe Computation Paper List

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.

Characteristics of Natural Language - 3

- There are rich knowledge in text



From Language to Knowledge

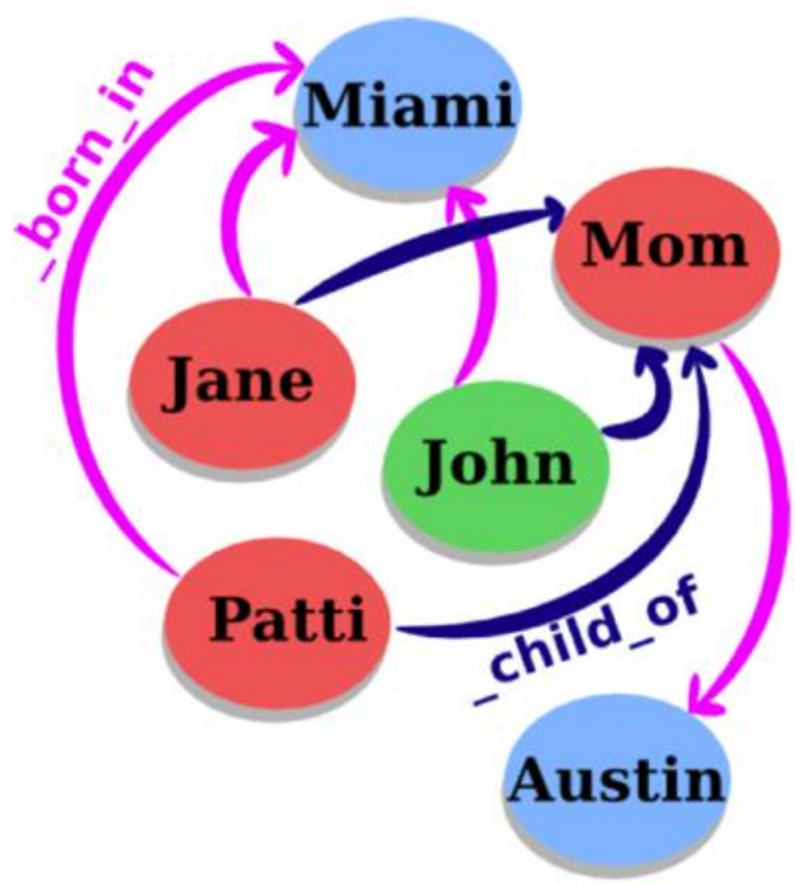
author

Shakespeare

Romeo and Juliet

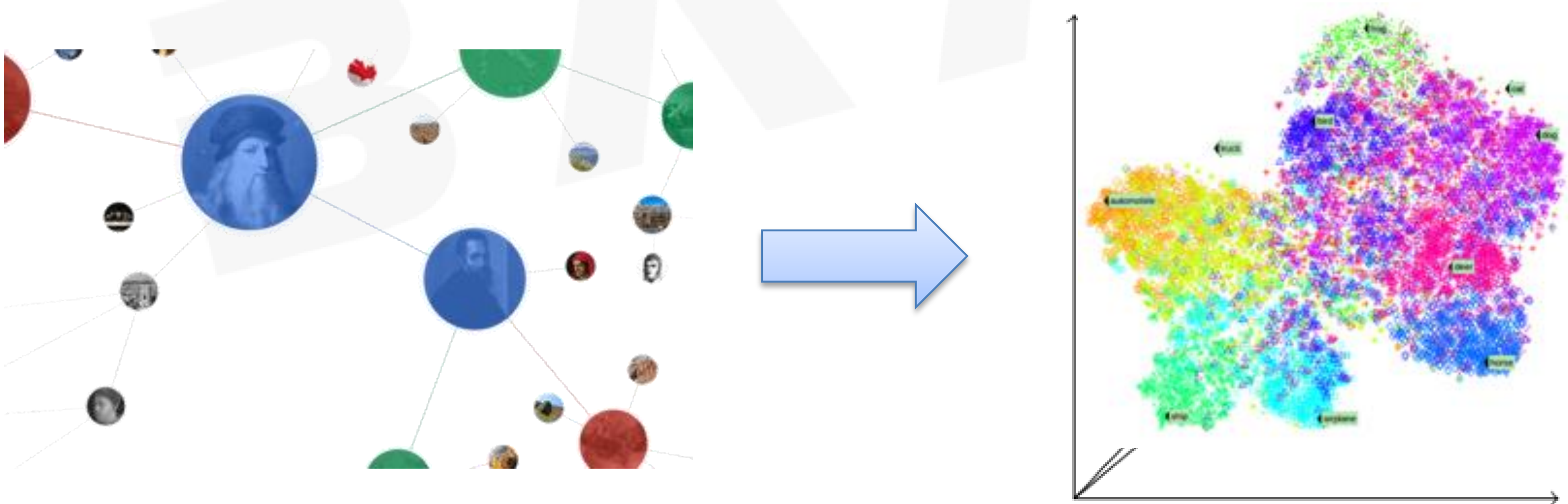
Knowledge Graph

- Entity as vertices and relations as edges
- Facts as triples
  - (head, *relation*, tail)
- Typical KG
  - Lexical KG: WordNet
  - World KG: Freebase



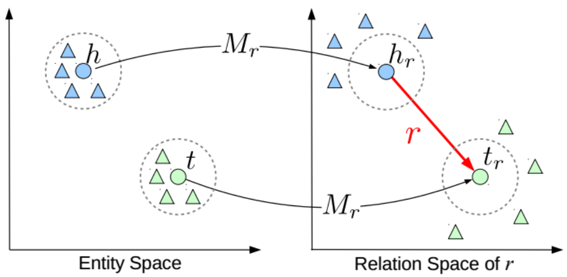
Knowledge Representation

- Symbol-based knowledge representation can not well compute semantic relations of entities
- Solution: project knowledge into low-dimensional space

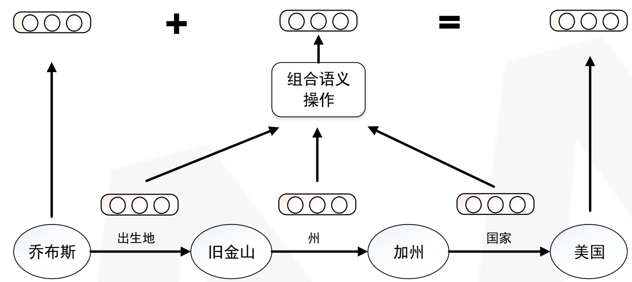


Knowledge Representation Learning

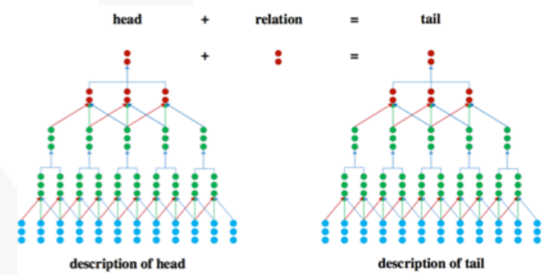
- Incorporate rich information in KG (such as description, class and images) for KRL



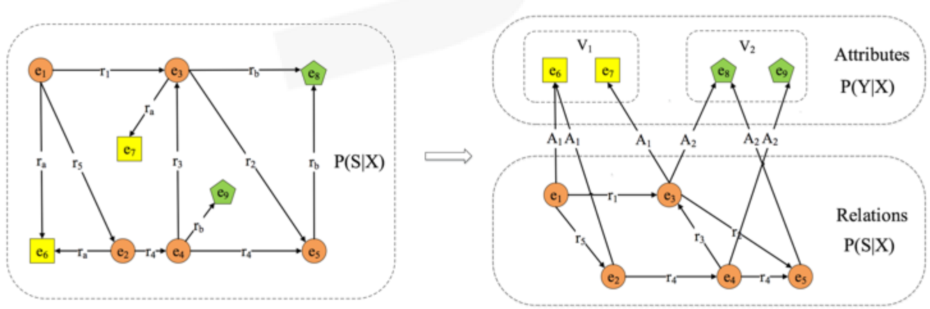
KRL with Complex Relations
TransR (AAAI 2015)



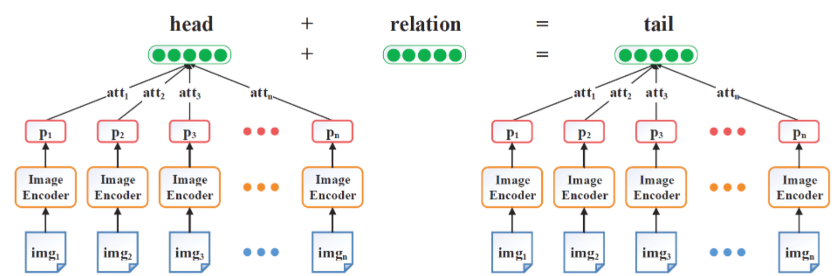
KRL with Relation Paths
PTransE (EMNLP 2015)



KRL with Entity Descriptions
DKRL (AAAI 2016)



KRL with Entities, Relations and Attributes
KR-EAR (IJCAI 2016)



KRL with Entity Images
IKRL (IJCAI 2017)

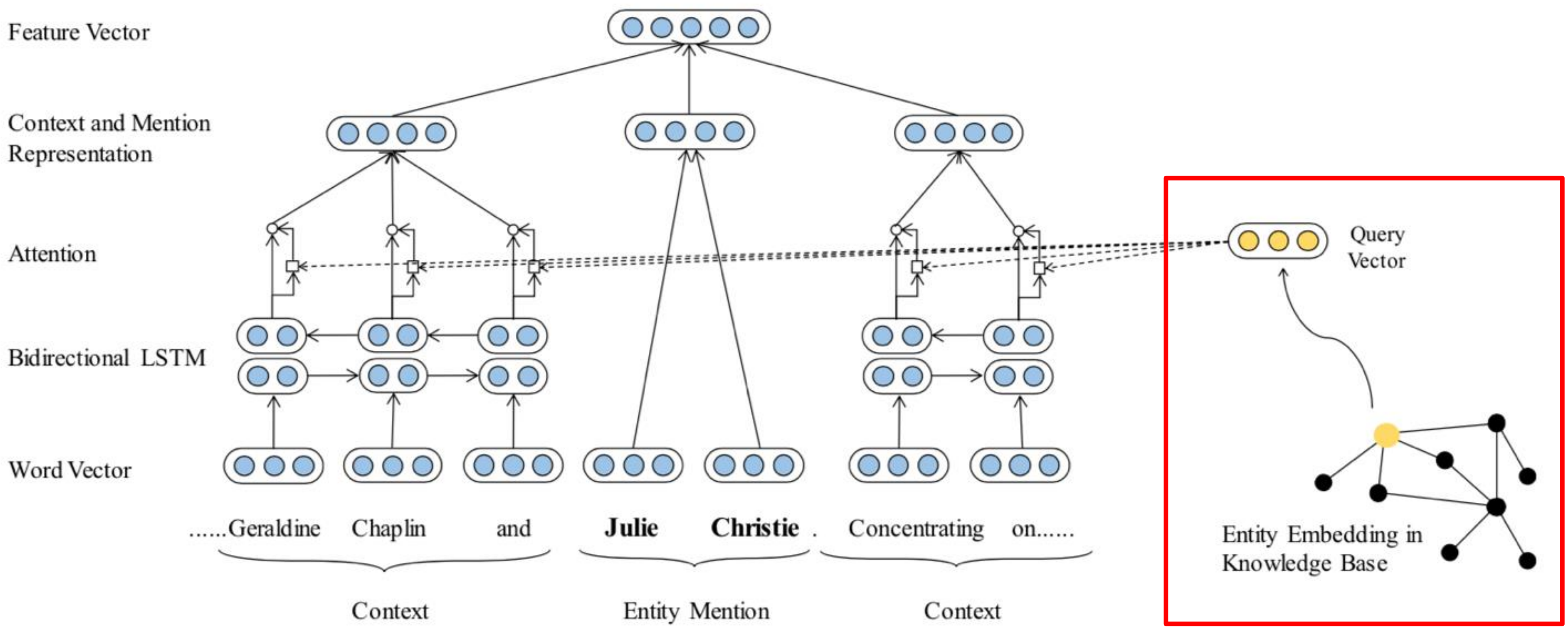
Knowledge Representation Learning Paper List

https://github.com/thunlp/KRLPapers

- Xin Lv, Lei Hou, Juanzi Li, Zhiyuan Liu. **Differentiating Concepts and Instances for Knowledge Graph Embedding**. EMNLP 2018.
- Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin. **Does William Shakespeare REALLY Write Hamlet? Knowledge Representation Learning with Confidence**. AAAI 2018.
- Ruobing Xie, Zhiyuan Liu, Huanbo Luan, Maosong Sun. **Image-embodied Knowledge Representation Learning**. IJCAI 2017.
- Yankai Lin, Zhiyuan Liu, Maosong Sun. **Knowledge Representation Learning with Entities, Attributes and Relations**. IJCAI 2016.
- Ruobing Xie, Zhiyuan Liu, Maosong Sun. **Representation Learning of Knowledge Graphs with Hierarchical Types**. IJCAI 2016.
- Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, Maosong Sun. **Representation Learning of Knowledge Graphs with Entity Descriptions**. AAAI 2016.
- Yankai Lin, Zhiyuan Liu, Huanbo Luan, Maosong Sun, Siwei Rao, Song Liu. **Modeling Relation Paths for Representation Learning of Knowledge Bases**. EMNLP 2015.
- Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, Xuan Zhu. **Learning Entity and Relation Embeddings for Knowledge Graph Completion**. AAAI 2015.

Knowledge-Guided Entity Typing

- Fine-grained entity typing
- Based on KG embeddings, propose Knowledge attention for better context understanding



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).

Experiment Results

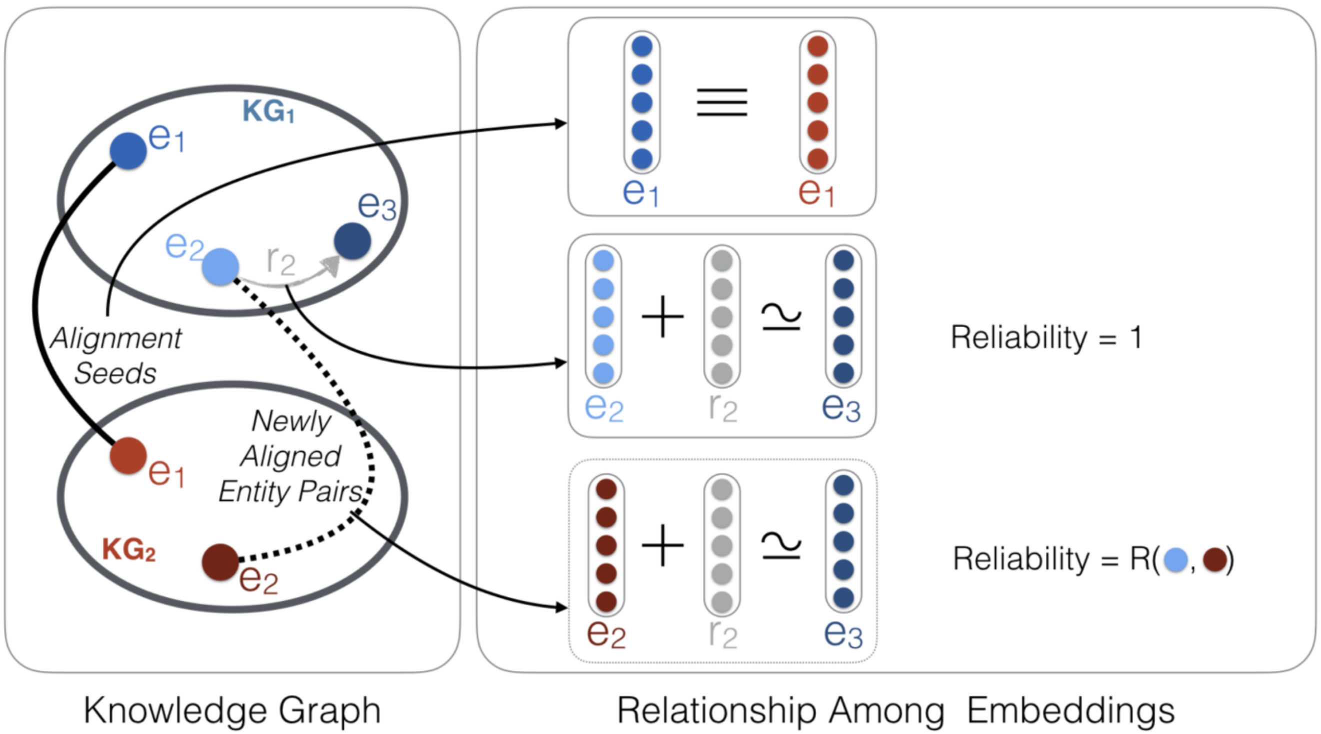
- KA and KA+D outperform all baselines, which indicates the effectiveness of knowledge
- KA+D means KA with Disambiguation

Dateset	WIKI-AUTO						
Metrics	Strict	Macro			Micro		
	Acc	Pre	Rec	F1	Pre	Rec	F1
AFET	20.32	67.00	45.82	54.75	69.29	42.40	52.61
KB-ONLY	35.12	69.65	71.35	70.49	54.85	74.99	63.36
HNM	34.88	68.09	61.03	64.37	72.80	64,48	68.39
SA	42.77	75.33	69.69	72.40	77.35	72.63	74.91
MA	41.58	73.64	71.71	72.66	75.94	75.52	75.72
KA	45.49	74.82	72.46	73.62	76.96	75.49	76.22
KA+D	**47.20**	**75.72**	**74.03**	**74.87**	**77.96**	**77.87**	**77.92**

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).

Knowledge-guided Entity Alignment

- The solid line and the dashed line between KGs denote alignment seeds and newly aligned entity pairs during iterative learning



Hao Zhu, Ruobing Xie, Zhiyuan Liu, Maosong Sun. Iterative Entity Alignment via Joint Knowledge Embeddings. IJCAI 2017.

Experiment Results

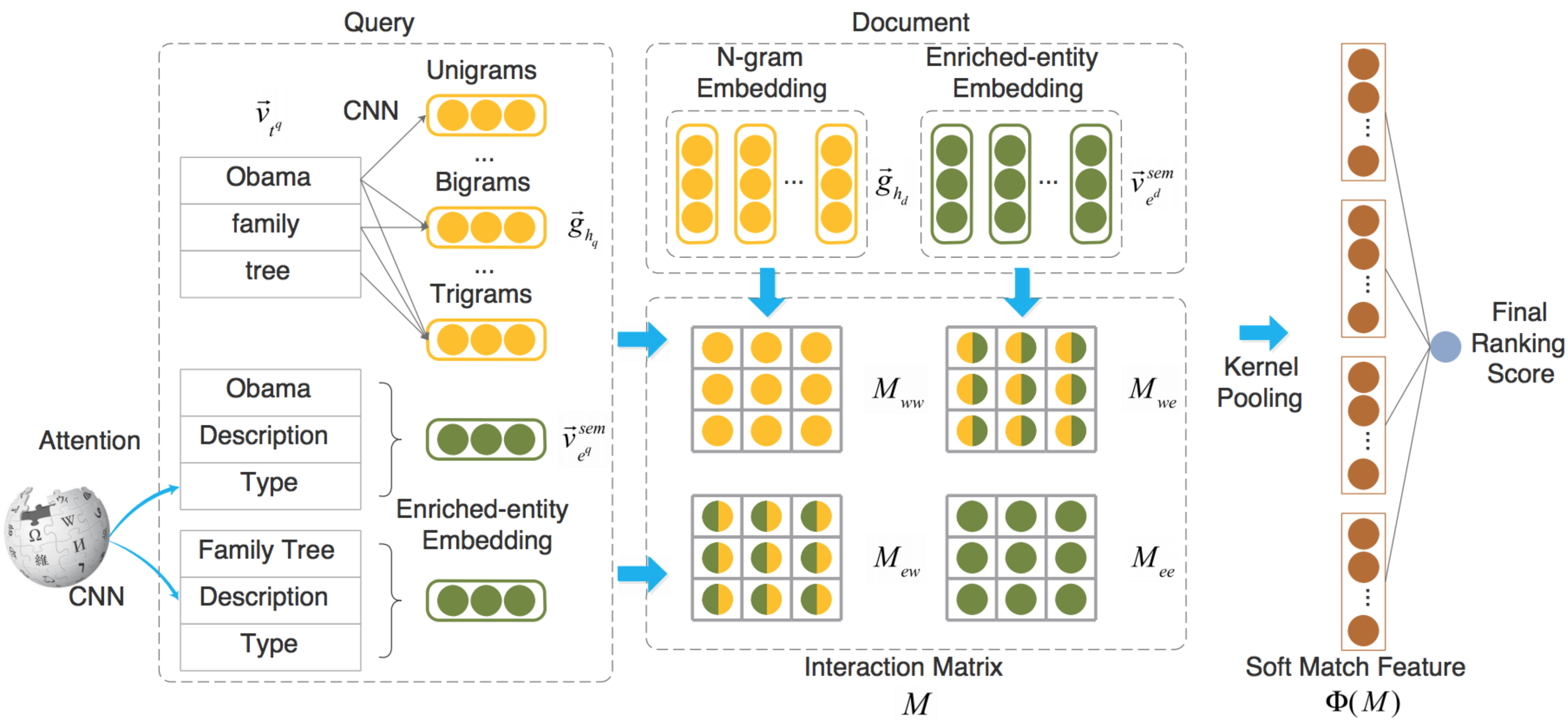
- Build three datasets based on FB15K (DFB-1,2,3)
- Knowledge-guided Entity Alignment achieves the best performance

指标	DFB-1			DFB-2			DFB-3		
	Hits@1 (%)	Hits@10 (%)	MR	Hits@1 (%)	Hits@10 (%)	MR	Hits@1 (%)	Hits@10 (%)	MR
MTransE (LT)	38.9	61.0	237.7	12.3	33.8	419.2	6.5	22.0	699.8
MTransE (TB)	13.6	35.1	547.7	13.9	35.4	675.7	4.5	16.1	1255.5
TransE + PS	61.9	79.2	105.2	41.1	67.0	154.9	12.2	34.6	431.9
ITransE (HA)	62.6	78.9	100.0	41.2	66.9	151.9	12.3	33.7	432.3
ITransE (SA)	**67.1**	**83.1**	**80.1**	**57.7**	**77.7**	**109.3**	**16.2**	**40.9**	**367.2**
PTransE + PS	65.8	83.4	62.9	46.3	72.1	96.8	15.8	40.2	346.9
IPTransE (HA)	66.1	83.3	59.1	46.2	72.6	94.2	15.1	39.7	337.6
IPTransE (SA)	**71.7**	**86.5**	**49.0**	**63.5**	**82.2**	**67.5**	**20.4**	**47.4**	**281.0**

Hao Zhu, Ruobing Xie, Zhiyuan Liu, Maosong Sun. Iterative Entity Alignment via Joint Knowledge Embeddings. IJCAI 2017.

Knowledge-Guided Neural Ranking

- Introduce world knowledge from KGs into KNRM



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.

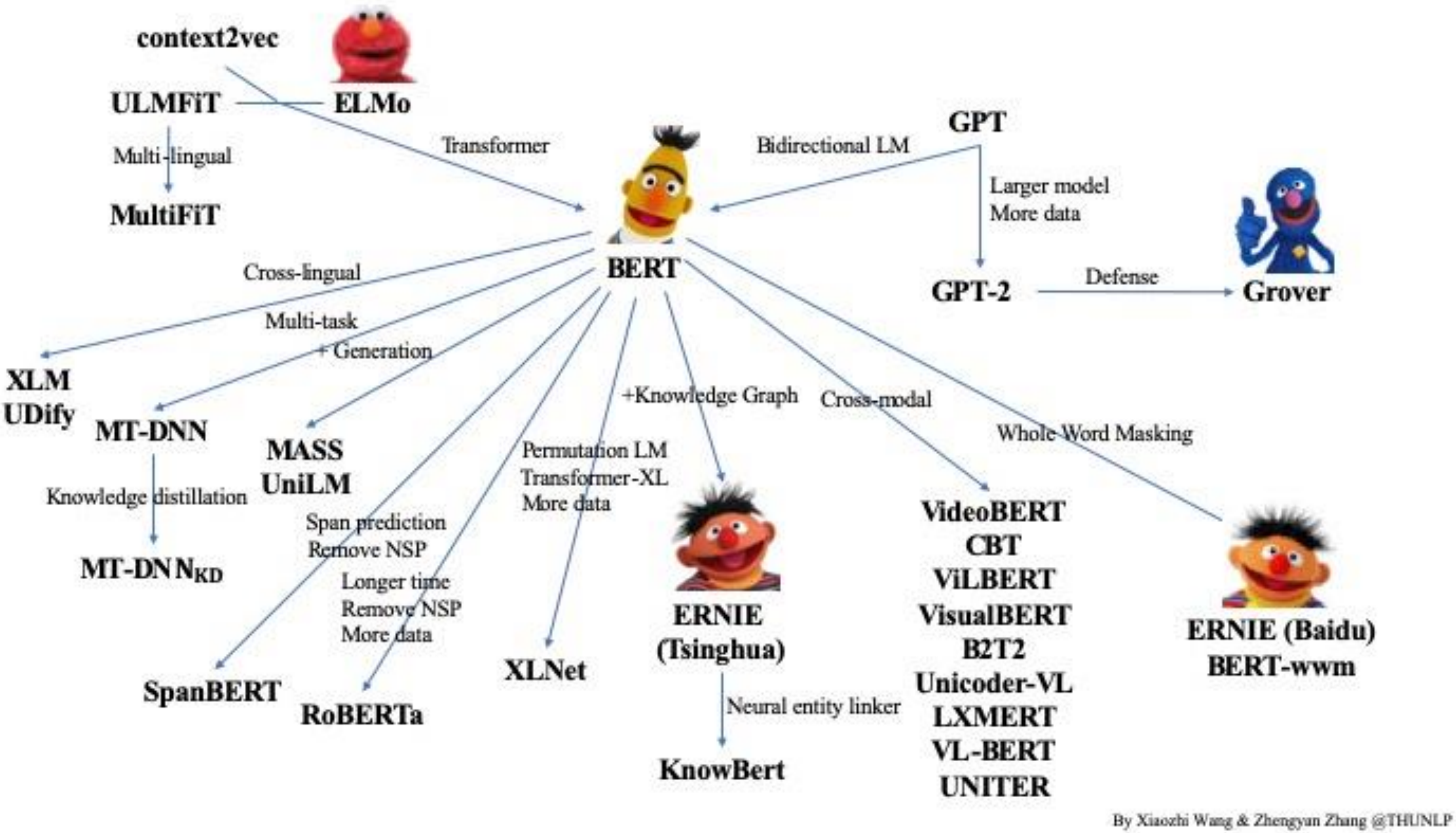
Experiment Results

- Knowledge-guided models achieve significant improvement on KNRM

	Testing-SAME				Testing-DIFF				Testing-RAW	
Method	NDCG@1		NDCG@10		NDCG@1		NDCG@10		MRR	
BM25	0.142	−46%	0.287	−32%	0.163	−46%	0.325	−23%	0.228	−34%
RankSVM	0.146	−45%	0.309	−26%	0.170	−43%	0.352	−17%	0.224	−35%
Coor-Ascent	0.159	−40%	0.355	−15%	0.209	−30%	0.378	−11%	0.242	−30%
DRMM	0.137	−48%	0.313	−25%	0.213	−29%	0.359	−15%	0.234	−32%
CDSSM	0.144	−46%	0.333	−21%	0.183	−39%	0.353	−16%	0.231	−33%
MP	0.218	−17%	0.379	−10%	0.197	−34%	0.345	−18%	0.240	−30%
K-NRM	0.265	–	0.420	–	0.300	–	0.423	–	0.345	–
Conv-KNRM	0.336	27%	0.481	15%	0.338	13%	0.432	2%	0.358	4%
EDRM-KNRM	0.310	17%	0.455	8%	0.333	11%	0.434	3%	0.362	5%
EDRM-CKNRM	**0.340**	28%	**0.482**	15%	**0.371**	24%	**0.451**	7%	**0.389**	13%

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.

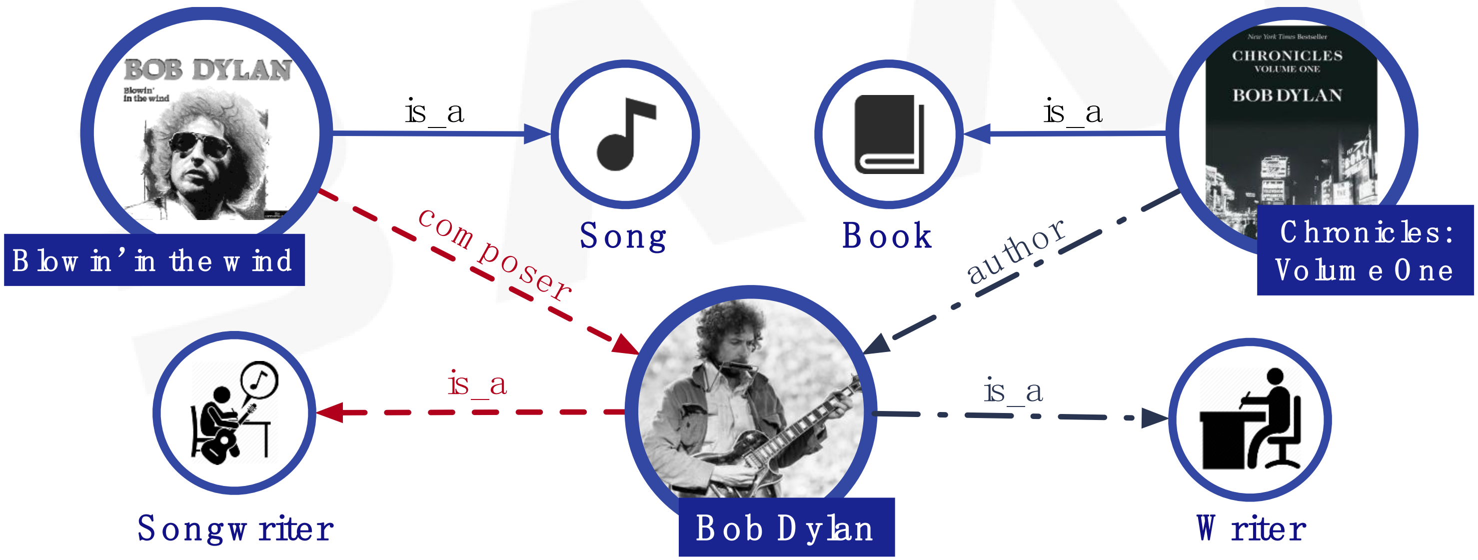
Pretrained Language Model



https://github.com/thunlp/PLMpapers

Knowledge-Guided PLM

- Intuitively, external knowledge information can effectively benefit language understanding
  - Low resource entities
  - Implicit background knowledge

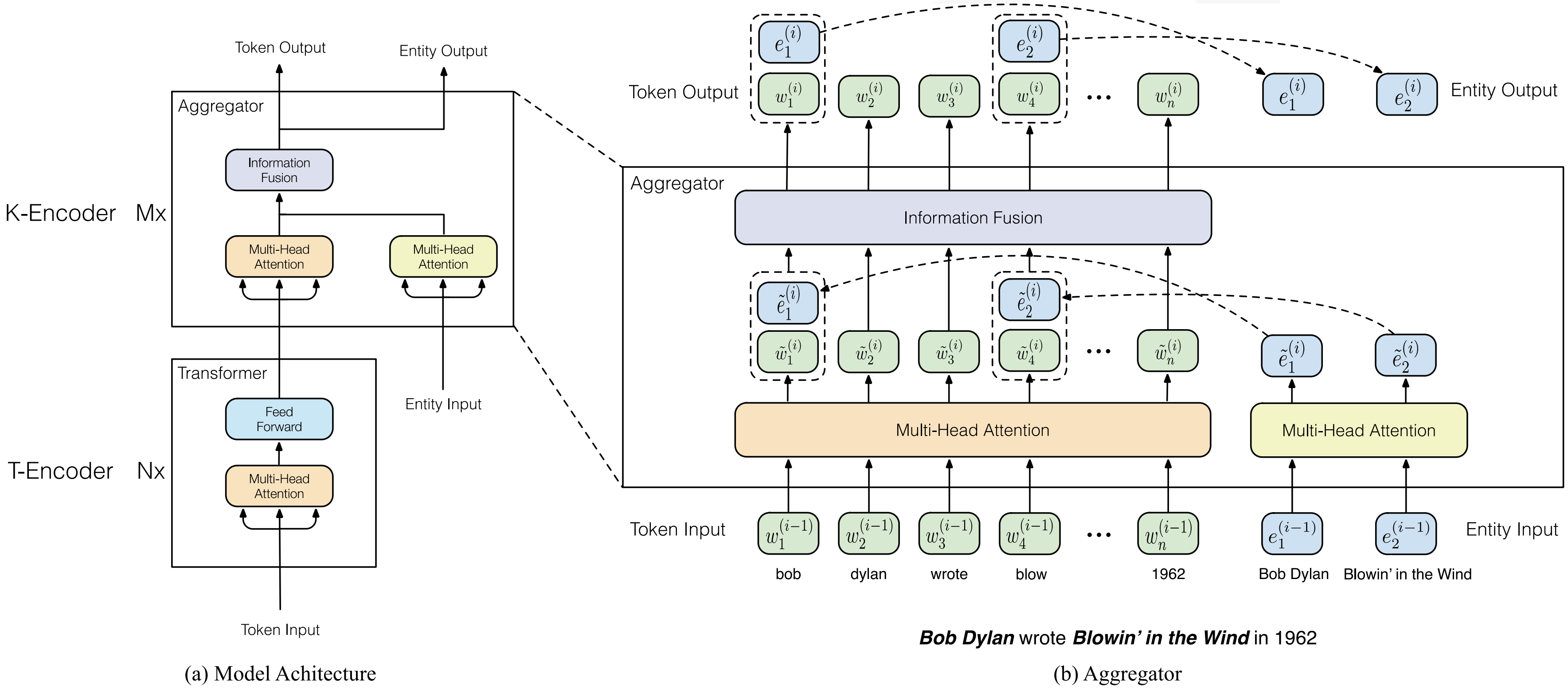


Bob Dylan wrote *Blowin' in the Wind* in 1962, and wrote *Chronicles: Volume One* in 2004.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, Qun Liu. ERNIE: Enhanced Language Representation with Informative Entities. ACL 2019.

Knowledge-Guided PLM

- The architecture of ERNIE
  - Lower layers for text
  - Higher layers for knowledge integration



Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, Qun Liu. ERNIE: Enhanced Language Representation with Informative Entities. ACL 2019.

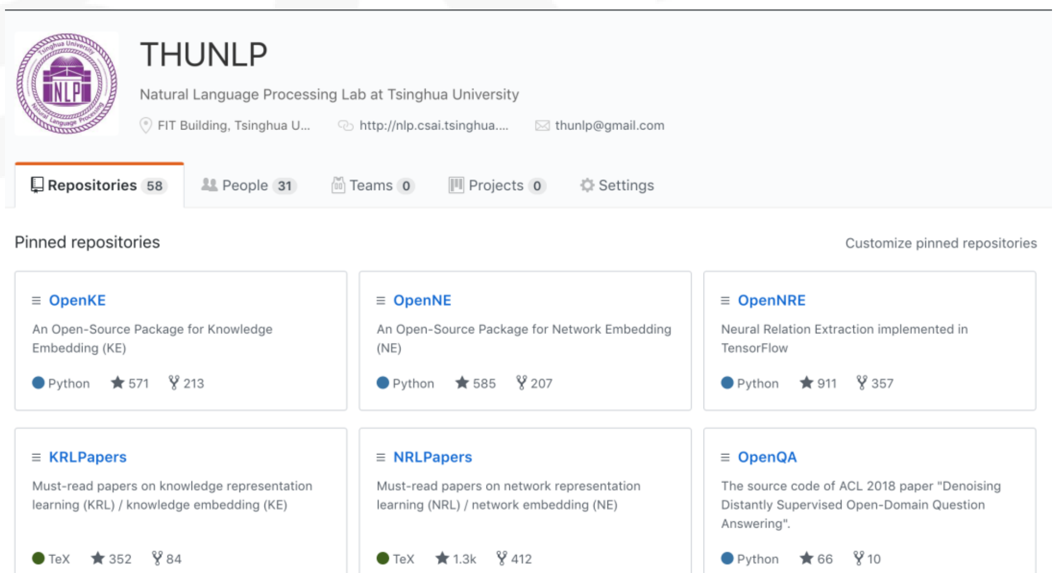
World Knowledge Guided NLP Paper List

- 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.

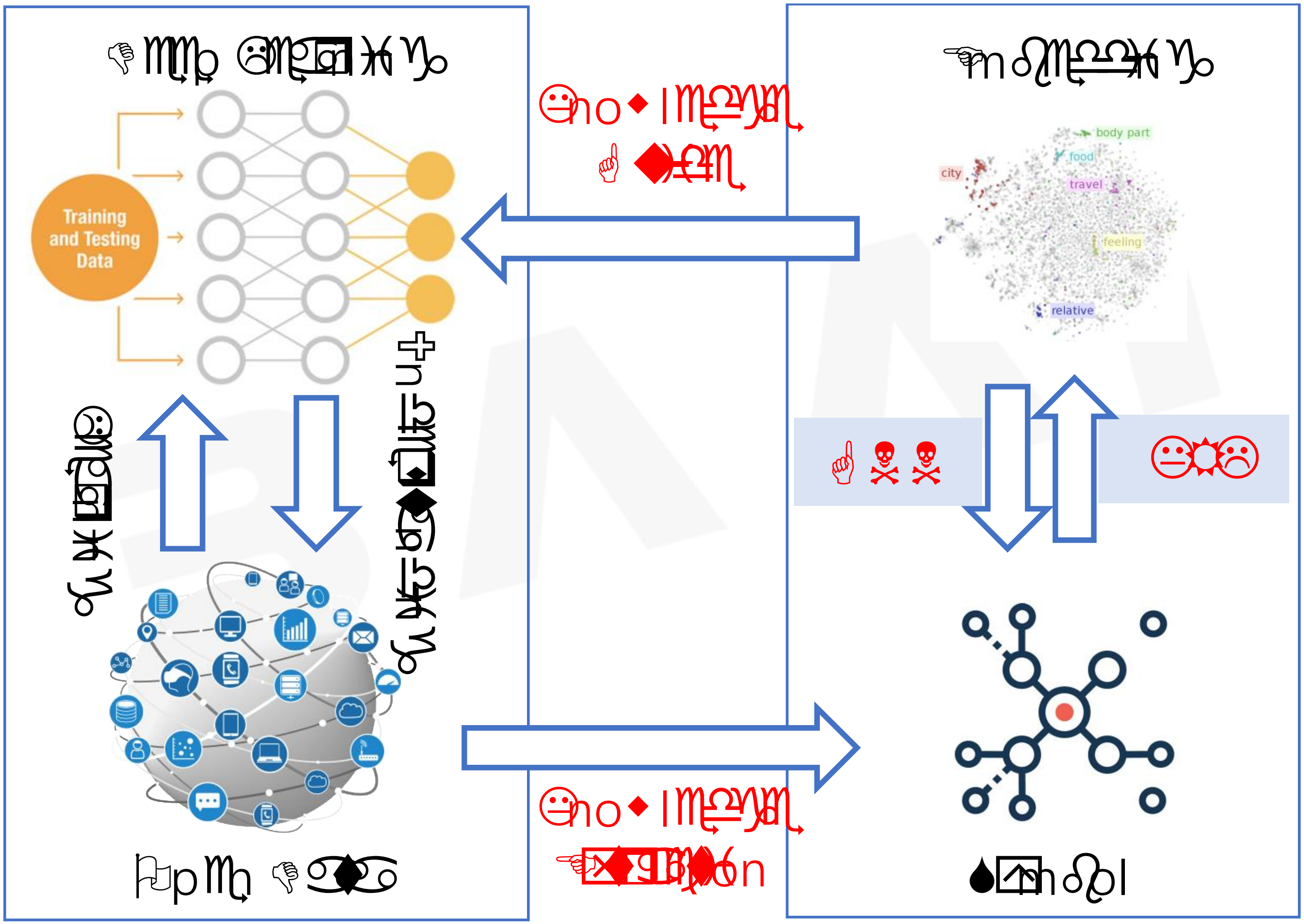
Open Source

- Packages for representation and acquisition of linguistic and world knowledge
- The projects obtain 23000+ stars on GitHub

https://github.com/thunlp



Summary: Knowledge-Guided NLP



Deep Learning

Knowledge Graph

THANK YOU!

liuzy@tsinghua.edu.cn

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

BAAI CONFERENCE

2019北京智源大会

数据、知识与认知推理专题论坛

微信扫码加入

智源大会现场微信交流群