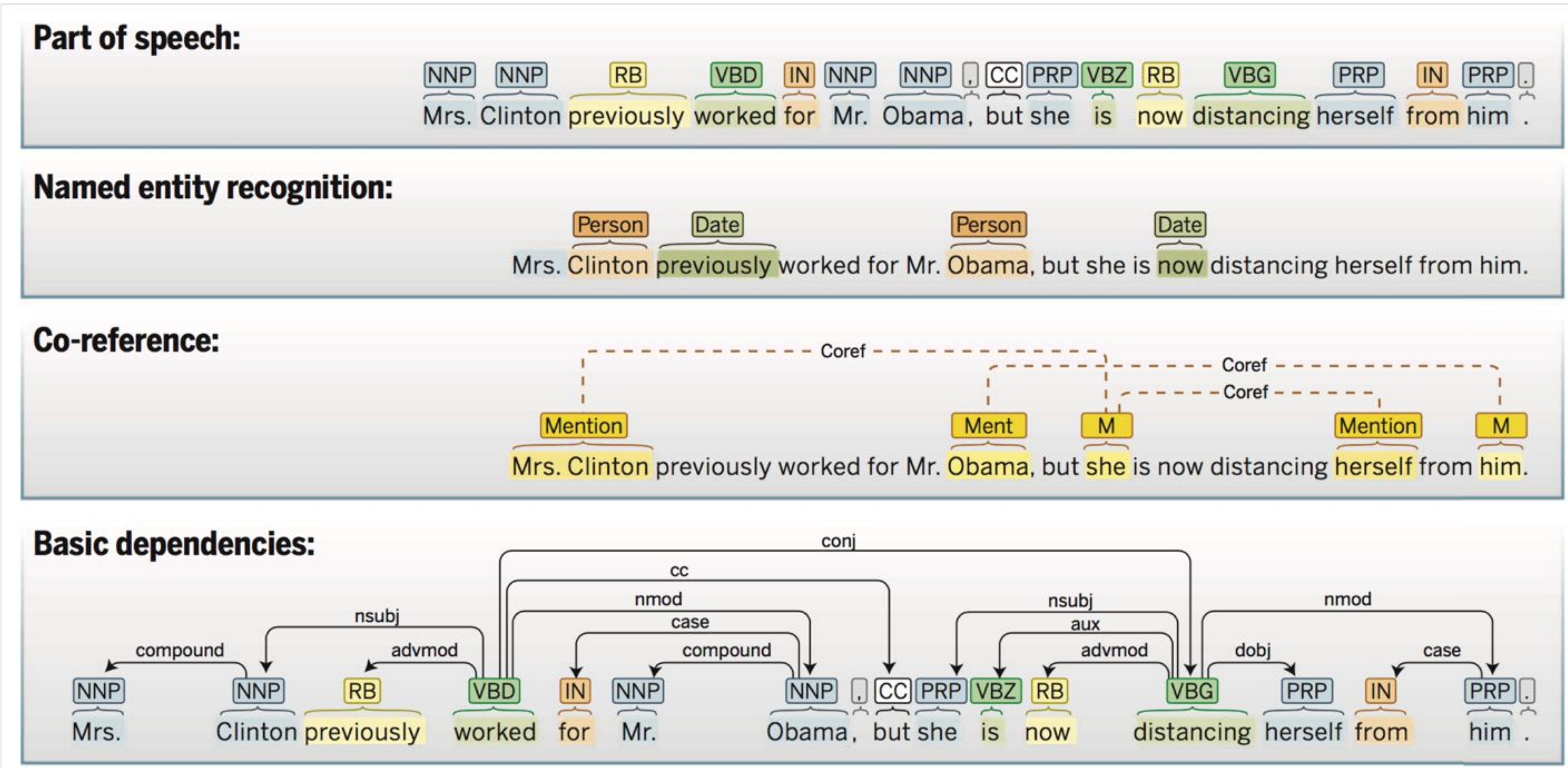
THU-CUHK-NWPU
The International Doctoral Forum 2018

Knowledge-Guided Natural Language Processing

THUNLP
Zhiyuan Liu

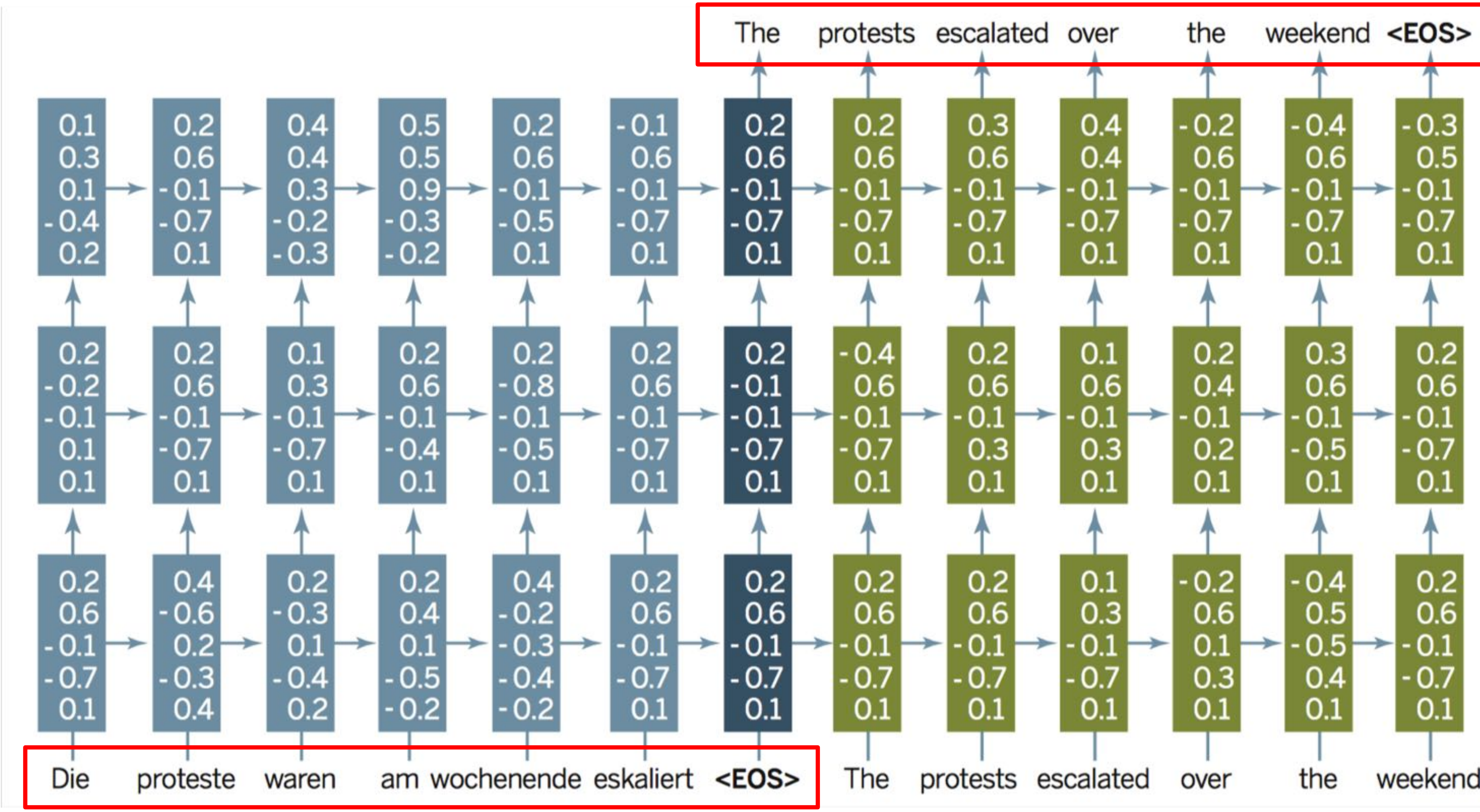
Natural Language Processing

- NLP aims to understand human language
- Nature of NLP is structure prediction



Advances in Natural Language Processing. Science 2015.

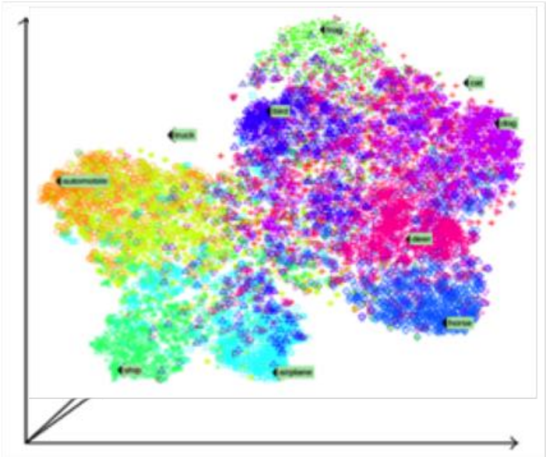
Deep Learning for NLP

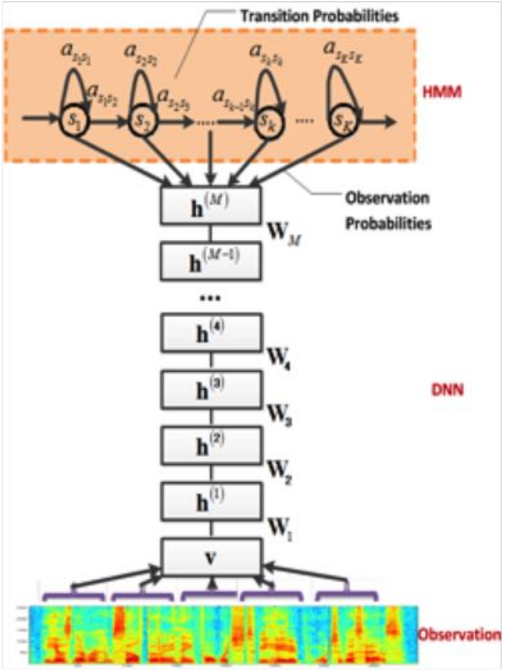


Advances in Natural Language Processing. Science 2015.

Characteristics of DL

- Distributed representation
  - Embeddings
  - Dense, real-valued, low-dimensional vectors
- Hierarchical structure
  - Corresponding to world hierarchy
  - Generalization
- Data-driven approach
  - Learn from large-scale training data





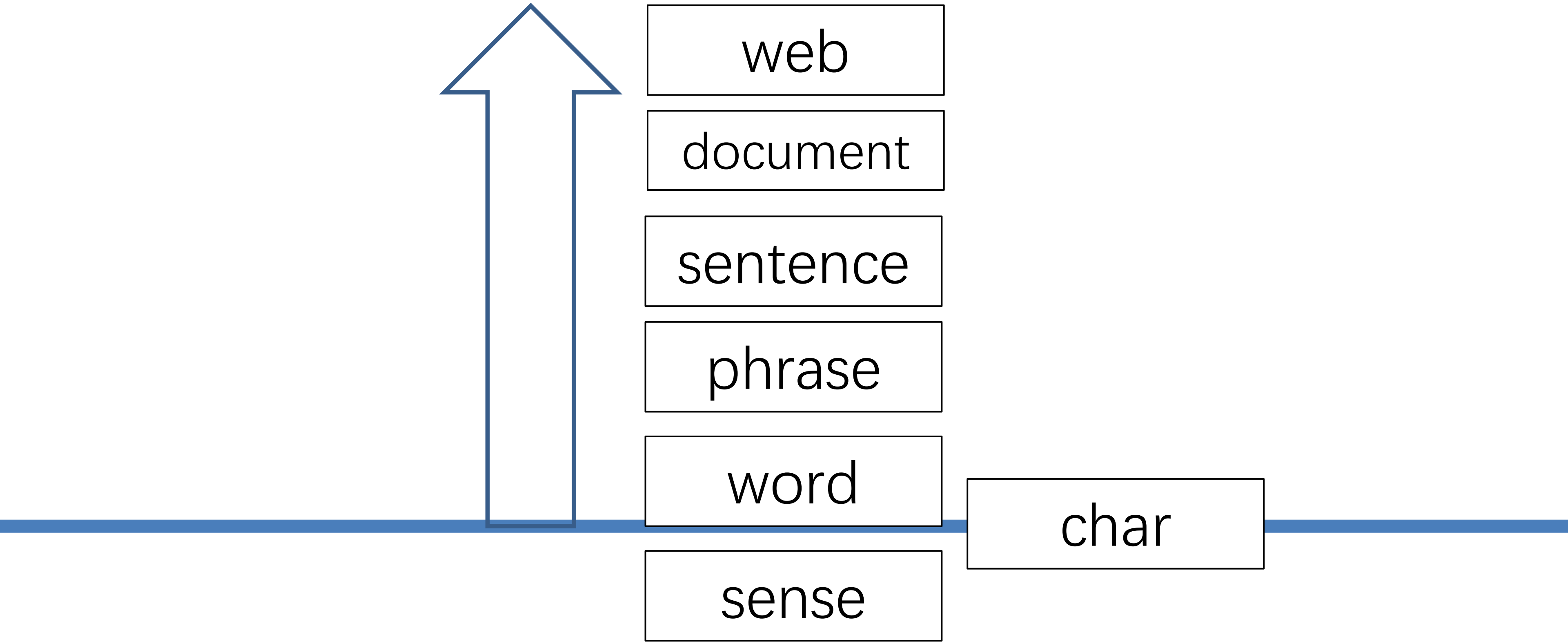
Challenges of DL for NLP

… 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.

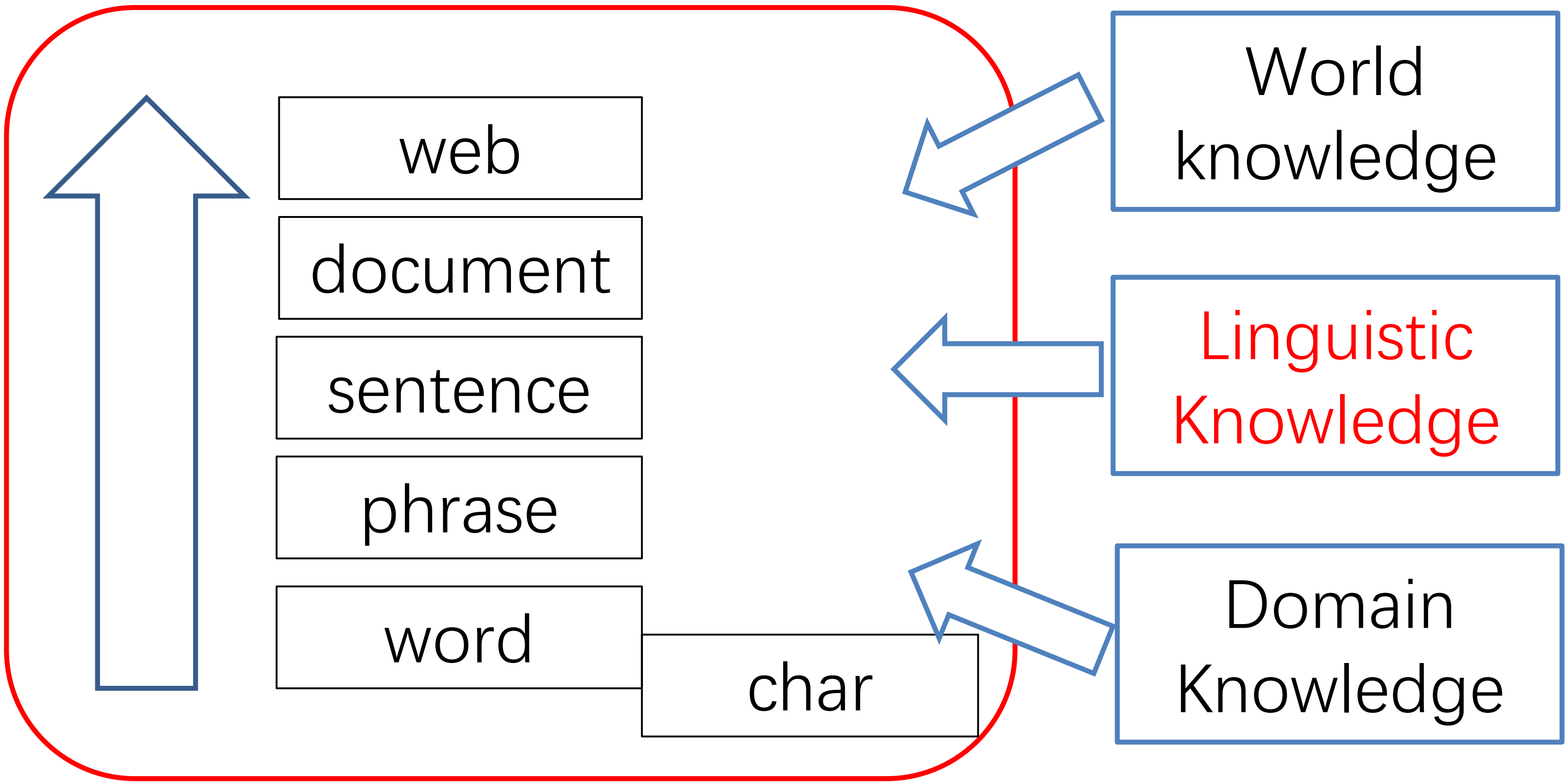
Characteristics of Natural Language

- There are multiple-grained units in languages
- Words/Chinese characters are minimal units of usages, but not minimal units of semantics



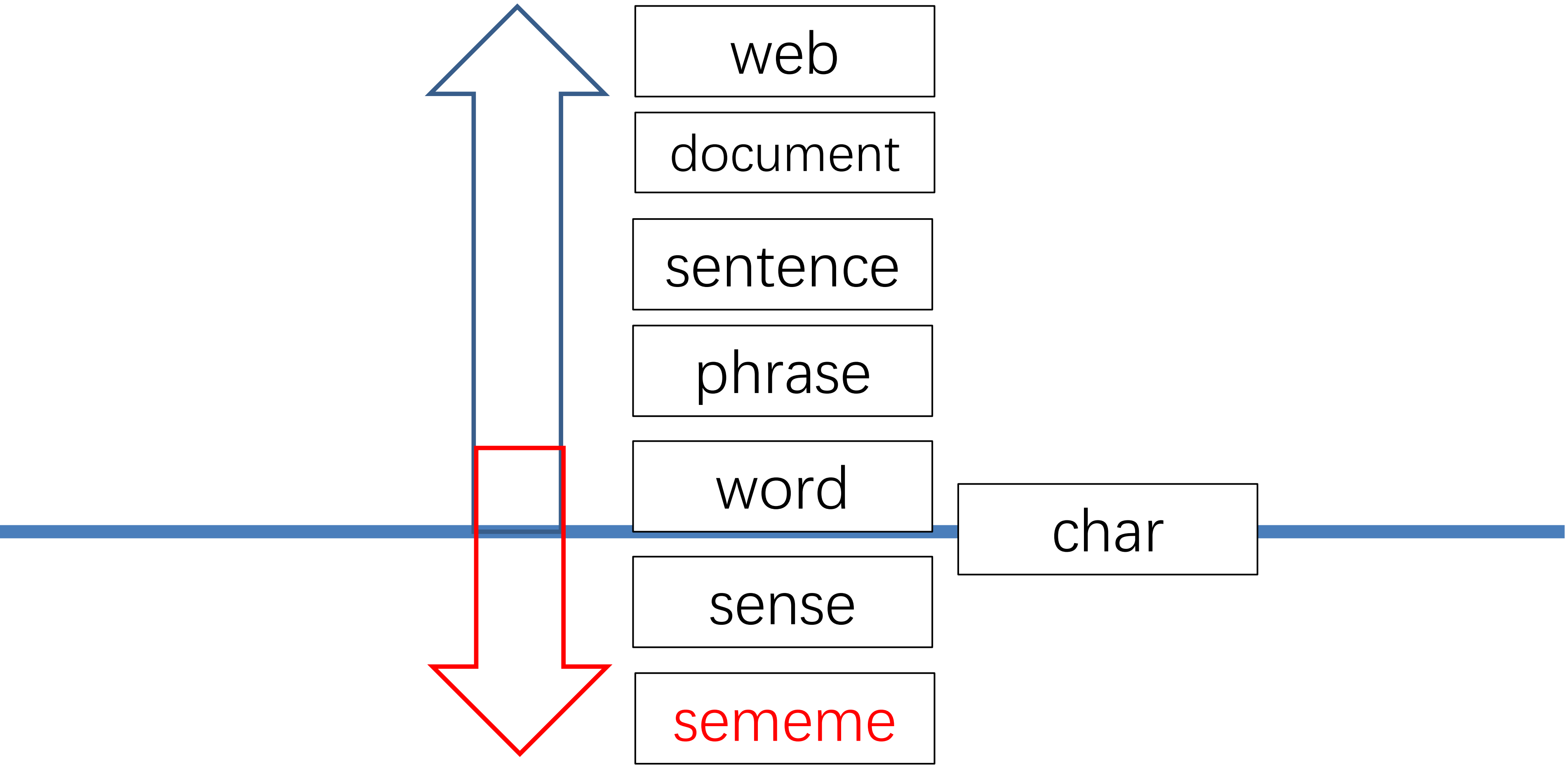
Characteristics of Natural Language

- There are rich knowledge in text



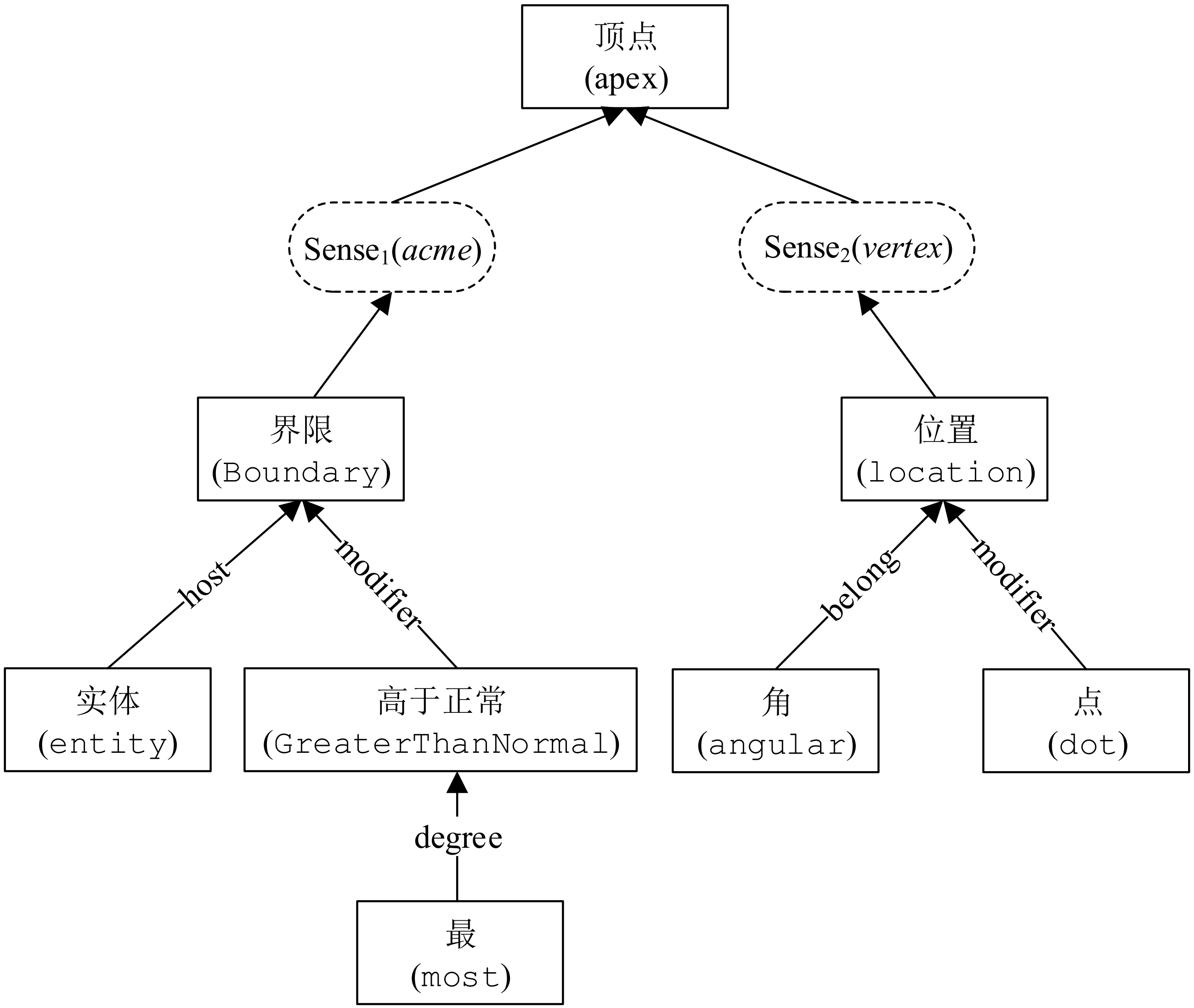
Use Sememes to Break Word Boundary

- 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 annotat3 ~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

Guide

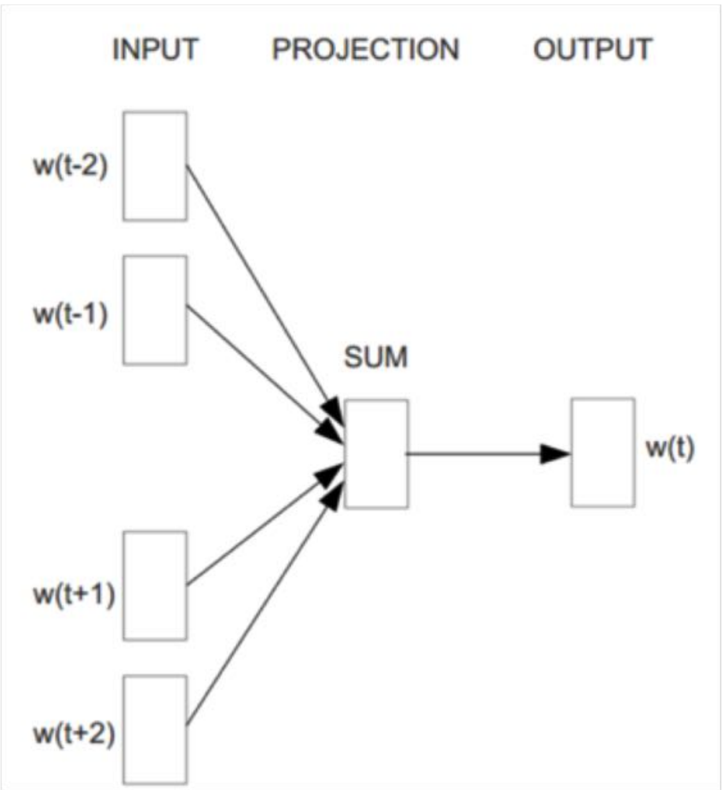
Data-Driven
DL

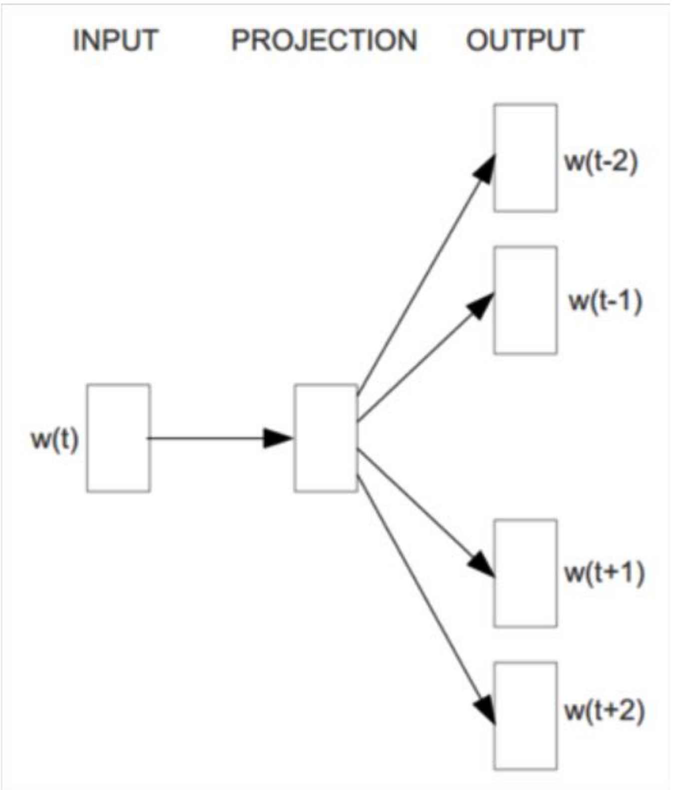
Symbol-based
Sememe Knowledge

WORD EMBEDDING WITH SEMEMES

Word Embedding

- Learn low-dimensional semantic representations for words



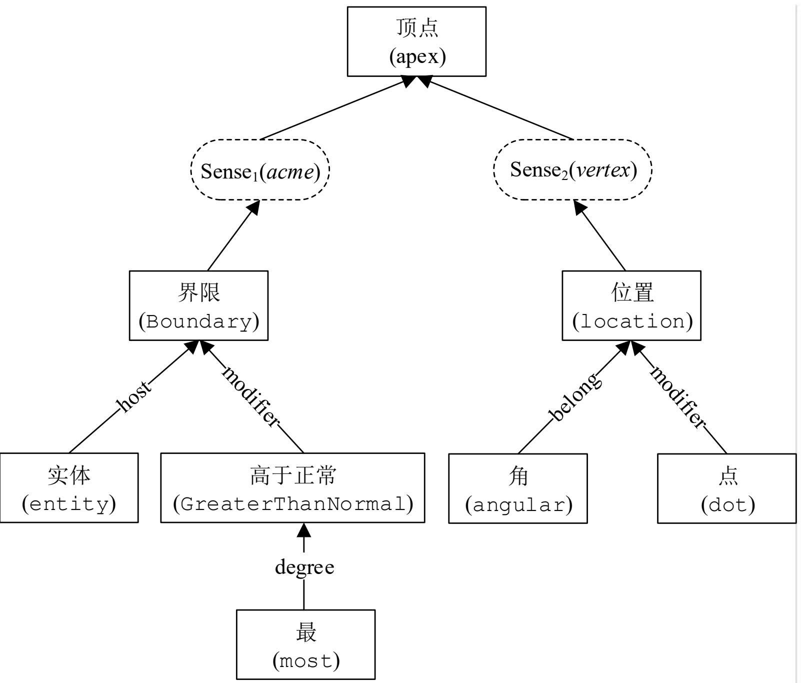


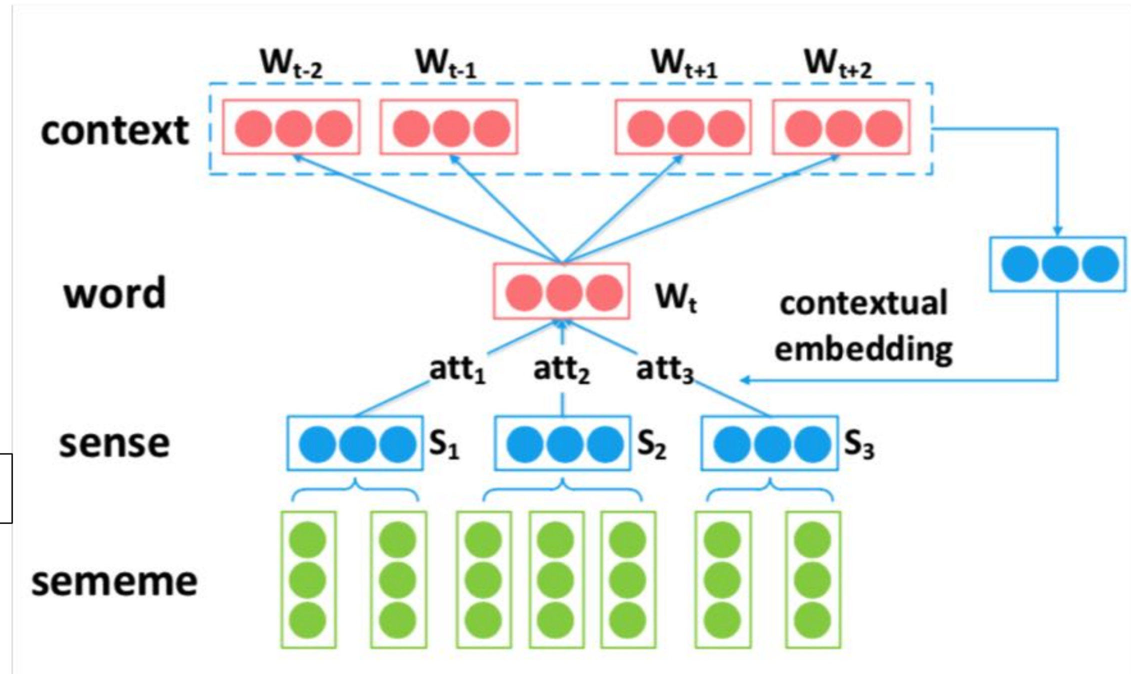
word2vec

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

Word Embedding with Sememes

- 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

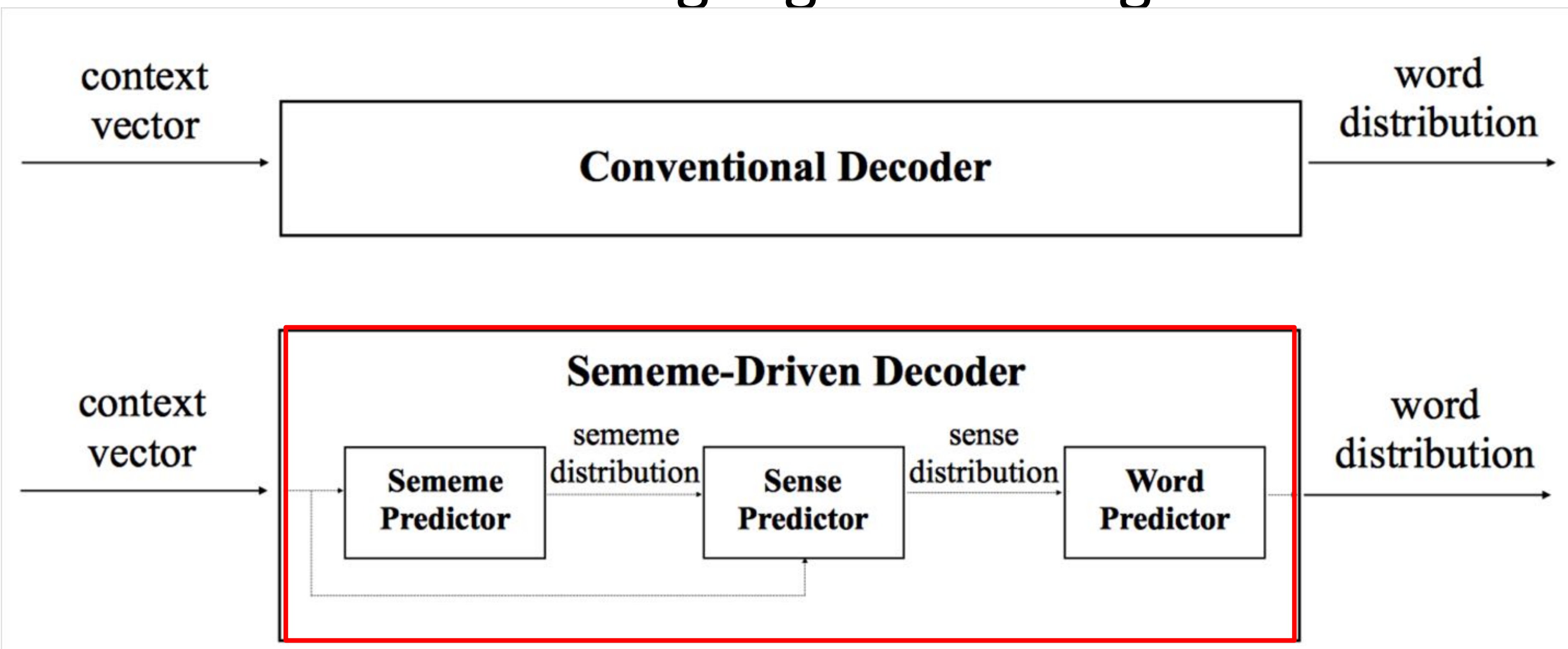
LANGUAGE MODELING WITH SEMEMES

Language Modeling

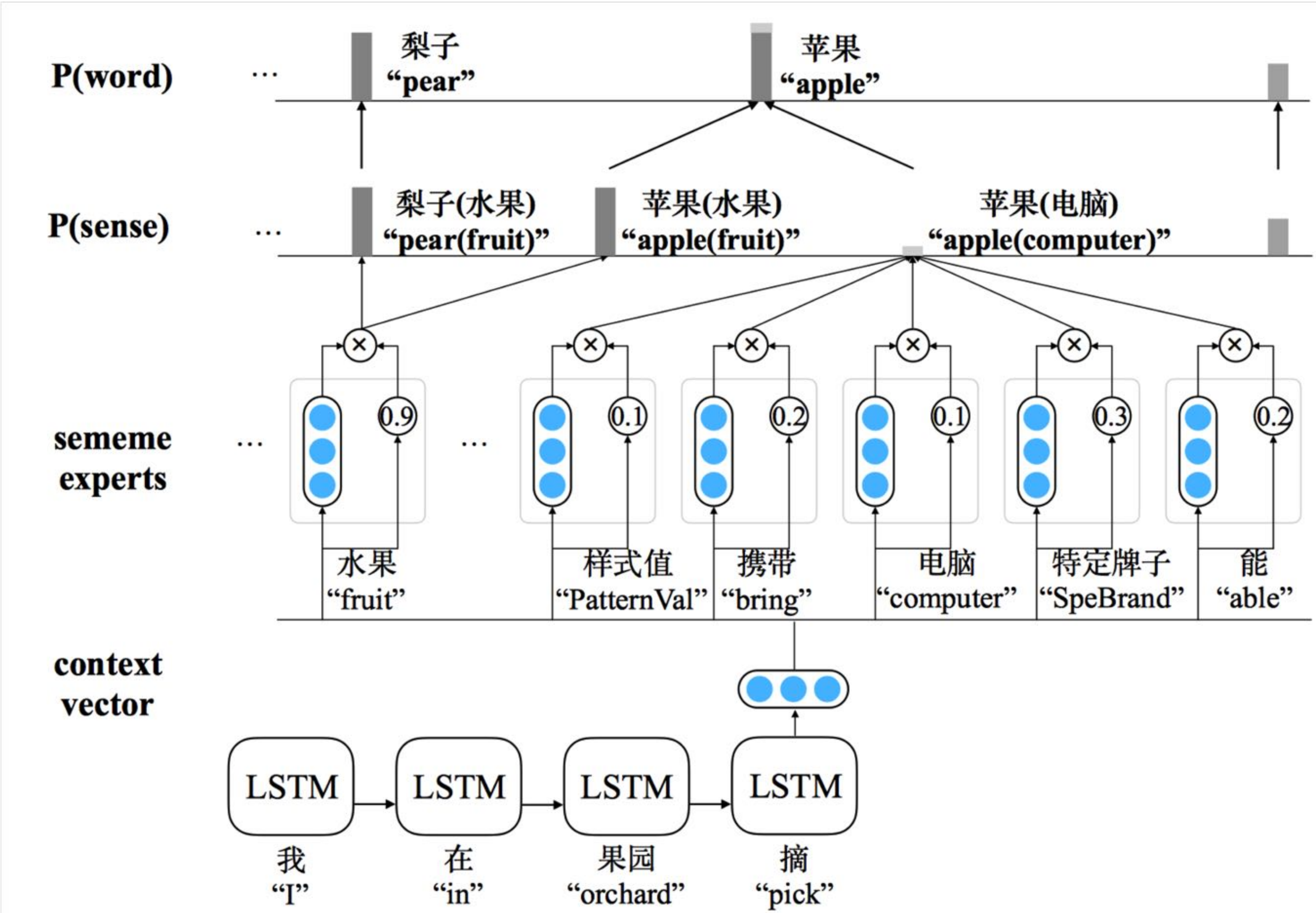
- Modeling word sequence with Markov property

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

- Sememe-Driven Language Modeling



Sememe-Driven Neural 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”	

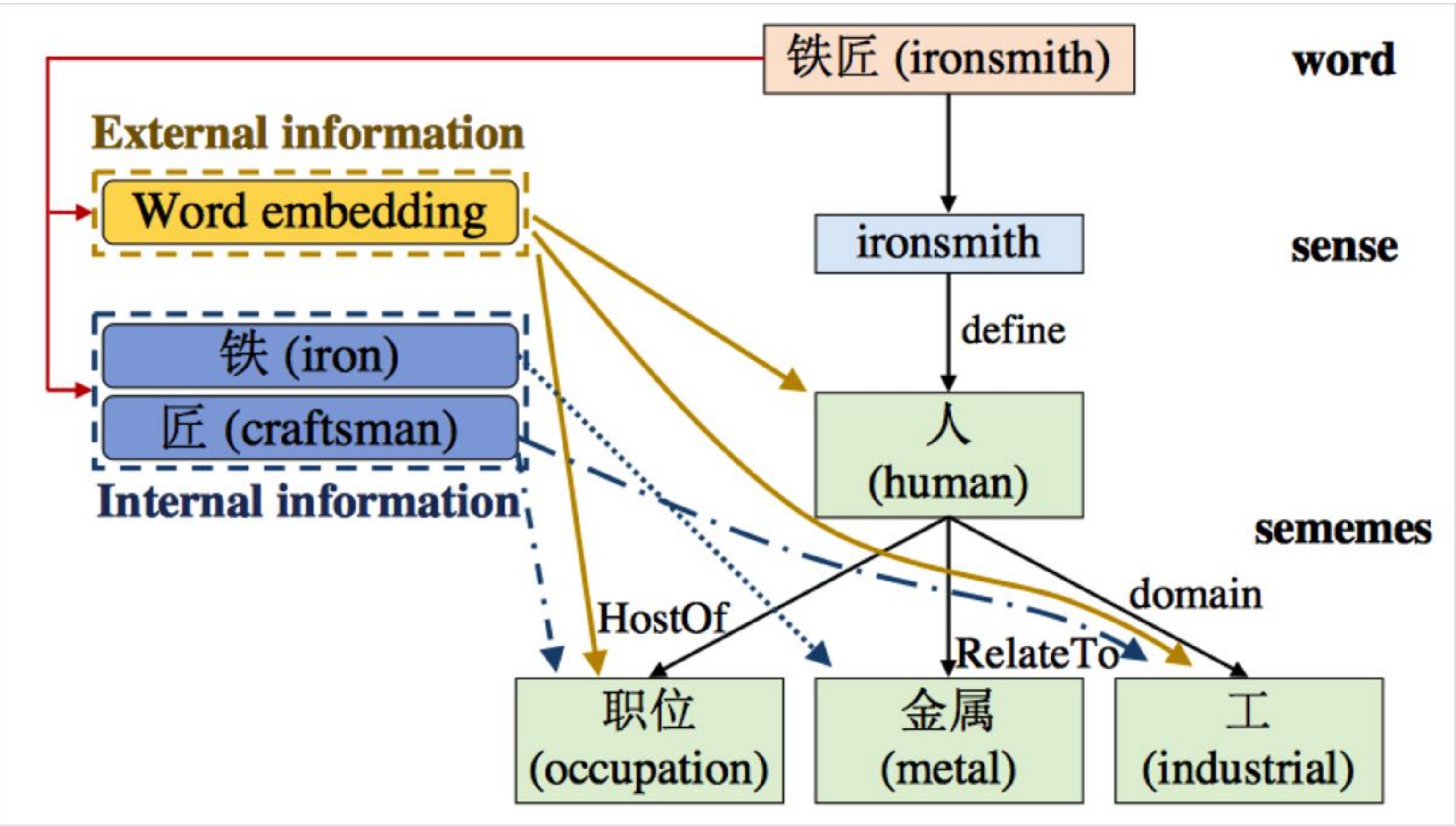
Prediction

Data-Driven
DL

Symbol-based
Sememe Knowledge

Sememe Prediction

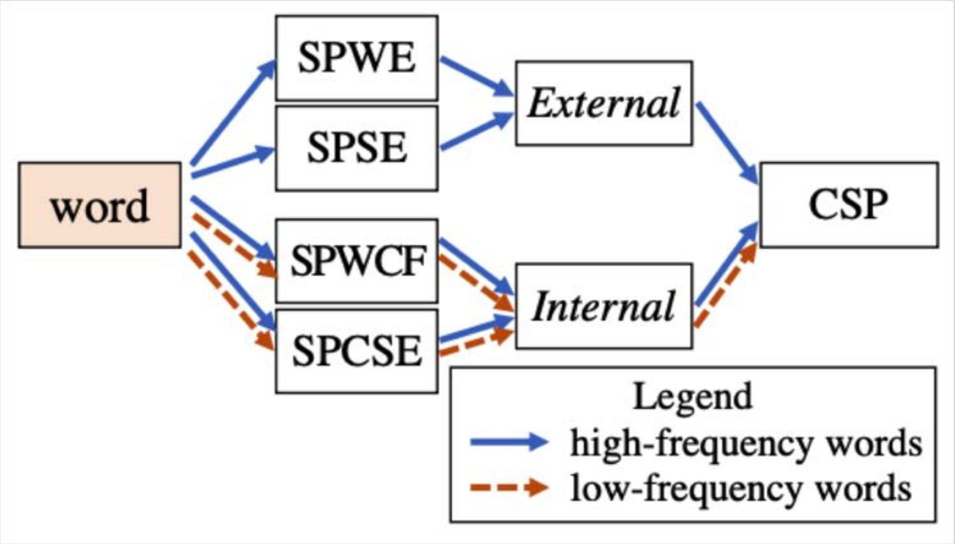
- Use both external and internal information to predict sememes



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**

Experiment Examples

- Both internal and external information can help sememe prediction

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)

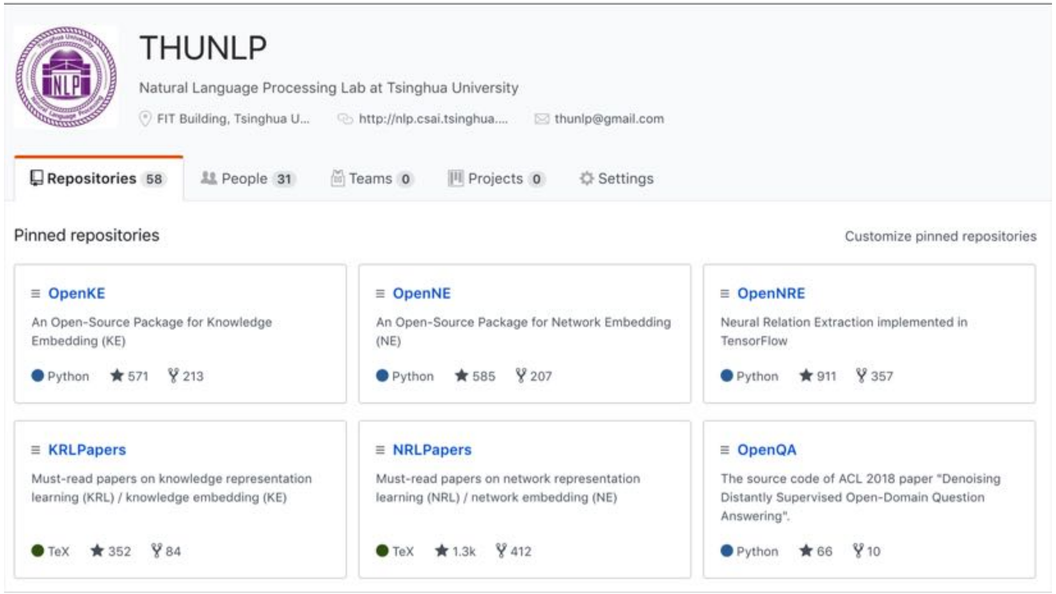
Related Papers

- 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.

Open Source

- Packages for representation and acquisition of linguistic and world knowledge
- The projects obtain 10000+ stars on GitHub

https://github.com/thunlp

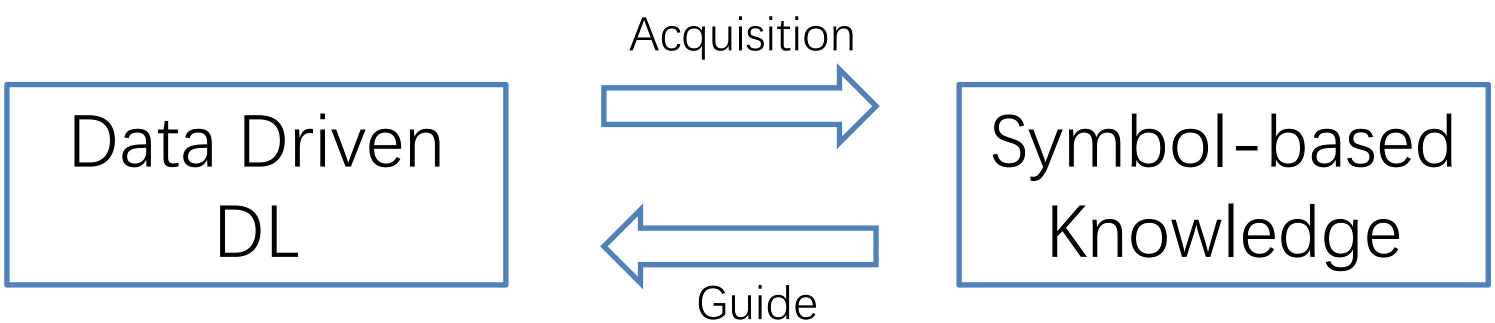


Summary

- Linguistic knowledge of lexical sememes can break word boundary for language modeling, and improve interpretability of neural language models

NLP/AI = Data-Driven + Knowledge-Guide

- DL methods for NLP can also be used for knowledge acquisition



THANKS!

liuzy@tsinghua.edu.cn