

# 人工智能与语言学关系的流变—人工智能视角

## The Evolution of the Relationship between AI and Linguistics — from the Perspective of AI

刘群 LIU Qun

Huawei Noah's Ark Lab

人文+人工智能交叉学科沙龙，复旦大学，黄大年茶思屋

2024.06.30



NOAH'S ARK LAB



HUAWEI

# Content

## 动机与背景

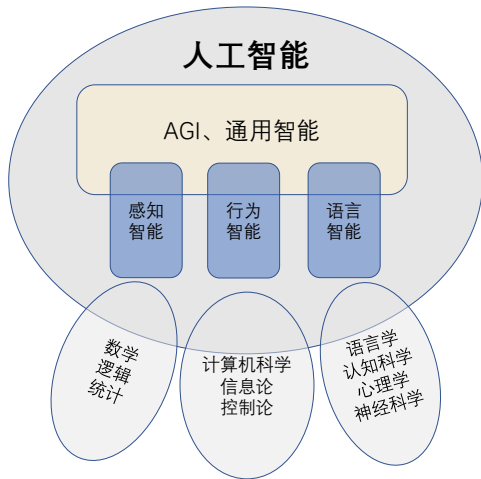
语言学对人工智能的影响

人工智能对语言学的影响

总结：大模型时代人工智能与语言学关系现状与展望

# 背景与动机

- ▶ 语言是人类智能的高级表现形式，语言智能也是人工智能的重要组成部分。
- ▶ 语言学被认为是人工智能的重要理论基础之一。
- ▶ 在人工智能发展过程中，语言学一直是深度参与，起到了重要的推动作用。
- ▶ 大模型时代，有必要重新审视和理解语言学在人工智能发展中所起到的作用。
- ▶ 本报告是本人作为在人工智能、特别是自然语言处理领域工作了数十年的长期从业者，对此做的一点粗浅的尝试。



# 时间线

- 
- 1950,图灵测试被提出
  - 1954,第一次机器翻译实验
  - 1957,Firth提出了分布式语义学的基本思想,成为现代NLP的基本假设
  - 1957,Chomsky的《句法结构》出版,转换生成语法被提出
  - 1959,Tesnière发表《结构句法基础》提出依存语法
  - 1962,达特茅斯会议,人工智能诞生
  - 1965,Chomsky的《句法理论的各个方面》出版
  - 1966,ALPAC报告,机器翻译研究受到打击
  - 1967,Brown Corpus公布
  - 1970,1970年代到1980年代初专家系统蓬勃发展
  - 1971,最早的词性标注研究
  - 1978,ARIAN78法国格勒诺布尔大学分析-转换-生成三阶段机器翻译系统
  - 1984,CYC大百科全书知识库项目启动
  - 1985,WordNet发布
  - 1985,GPSG语法提出
  - 1987,HPSG和LFG语法提出
  - 1987,首次MUC会议,信息抽取任务提出
  - 1992,Penn Treebank发布
  - 1993,Penn Discourse Treebank发布
  - 1994,随机上下文无关语法SCFG被提出 1994
  - 1997,IBM深蓝战胜人类国际象棋关键卡斯帕罗夫
  - 2000,FrameNet发布
  - 2002,语义角色标注任务被提出
  - 2003,基于短语的统计机器翻译
  - 2005,PropBank发布
  - 2006,基于句法的统计机器翻译方法
  - 2007,Dbpedia和Freebase发布
  - 2011,IBM Watson在电视智力竞赛中击败人类
  - 2013,Word Embedding发明
  - 2013,Seq2Seq神经机器翻译出现
  - 2016,AlphaGo战胜李世石
  - 2017,Transformer模型提出
  - 2018,预训练语言模型BERT、GPT等
  - 2020,GPT-3千亿参数大语言模型
  - 2022,ChatGPT发布

# Content

动机与背景

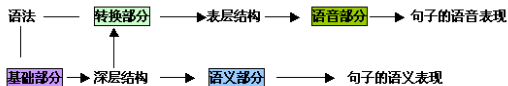
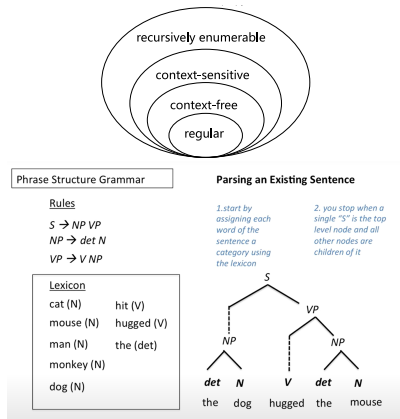
语言学对人工智能的影响

人工智能对语言学的影响

总结：大模型时代人工智能与语言学关系现状与展望

# Chomsky语言学理论

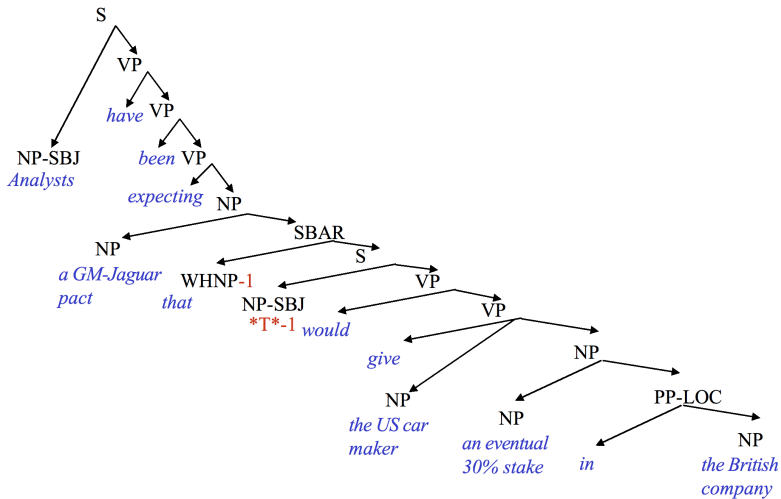
- ▶ 标准理论及其扩展、修正：
  - ▶ 短语结构语法、Chomsky层级
  - ▶ 深层结构、浅层结构、转换生成语法
  - ▶ 次范畴化、X-bar理论
  - ▶ 论元和语义角色
  - ▶ 位移、空语类、代词脱落
- ▶ 原则与参数理论
  - ▶  $\bar{X}$  Theory
  - ▶  $\theta$  Theory
  - ▶ Case Theory
  - ▶ Binding Theory
  - ▶ Bounding Theory
  - ▶ Control Theory
  - ▶ Government Theory
- ▶ 最简方案



## 宾州树库及其衍生语料库

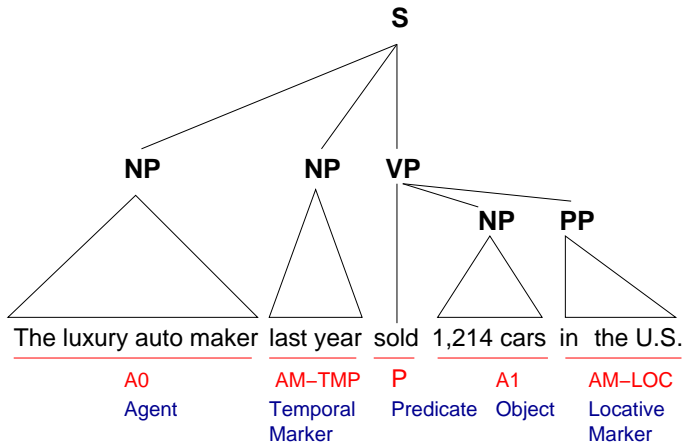
1992	Penn Treebank	1992年第一次发布，包含约1百万词的华尔街日报文本，标注了句法结构。
2002	RST Discourse Treebank	修辞结构理论（RST）话语树库，包含来自Penn Treebank的 385 篇文章，并在 RST 框架中注释了话语结构，以及与源文档相关的人工生成的摘录和摘要。
2002	Penn Chinese Treebank	2002年发布Building a Large-Scale Annotated Chinese Corpus，基于Penn Treebank的句法标注方法，对中文文本进行句法分析。
2004	NomBank	2004年发布，为名词短语提供语义角色标注。
2005	PropBank	2005年发布，为英语动词提供语义角色标注。
2006	TimeBank	2006年发布，为时间表达式提供详细的语义标注。
2008	Penn Discourse Treebank (PDTB) 2.0	2008年发布，包含对话文本的语料库，提供了话语层面的句法和语义结构标注。
2015	Universal Dependencies	2015年发布1.0版，基于多种语言的句法标注项目，部分基于Penn Treebank。
2015.179	Penn Treebank	1992年第一次发布，包含约2百万词的华尔街日报文本，标注了句法结构。

# 宾州树库

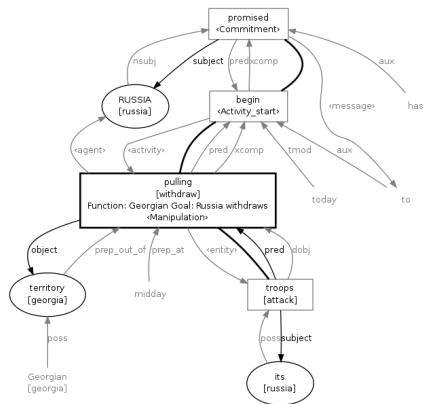
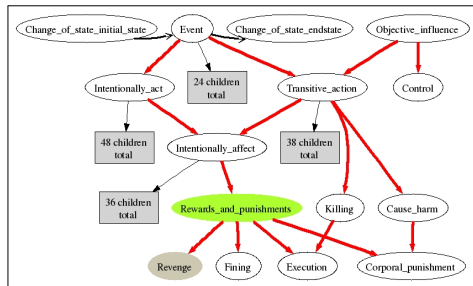




# PropBank



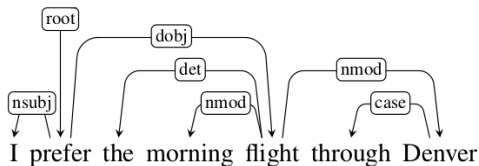
# FrameNet



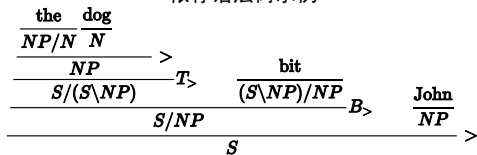
RUSSIA has promised to begin pulling its troops out of Georgia at midday today.

# 依存语法、配价语法、组合范畴语法（CCG）

- ▶ 依存语法是形式上最简单的语法：只需要建立词语之间的依存关系，无需对词语或者短语进行语言学标记。
- ▶ 配价语法将化学中“价”的概念引入语言，对词语语义的描述具有很好的可解释性。对词语的配价描述可以作为对依存语法很好的补充。
- ▶ 组合范畴语法为每个词语赋予一个复杂的范畴表示，而词语之间的组合只有简单的集中范畴组合操作，非常简单。
- ▶ 以上三种语法都属于词汇化语法，即对语言的描述只需要对词语进行刻画即可，无需像短语结构语法那样构造复杂的组合规则。
- ▶ 这几种语法因为形式简单，在自然语言处理中都有较多的研究与应用。特别是依存语法，是使用最为广泛的语言分析方法之一。



依存语法树示例

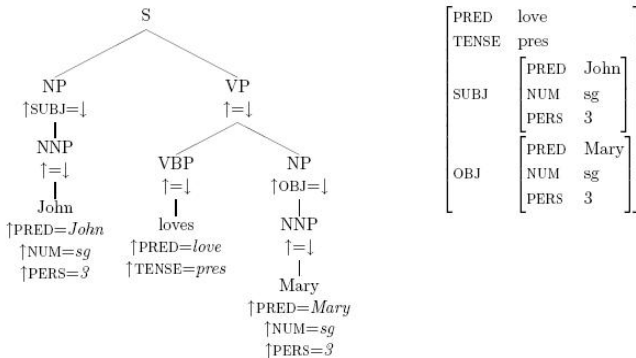


组合范畴语法示例

# 基于合一的语法

- ▶ 1980-1990年代，计算语言学界陆续提出了一批新的语法理论，包括词汇功能语法（LFG），功能合一语法（FUG），广义短语结构语法（GPSG），头驱动短语结构语法（HPSG）。
- ▶ 这些语法一个共同的特点是都采用复杂特征集+合一运算的形式，所以又叫做“基于合一的语法”（unification-based grammars）。
- ▶ 与依存语法类似，基于合一的语法也无需使用复杂的组合规则，只需要对词语进行描述。复杂特征集的使用可以对词语的特征进行细致的描述，而合一运算具有顺序无关、递增性等优点。这类语法一度收到非常多的关注，影响很大。

FIGURE 1: C-structure annotated with f-structure equations and the resulting f-



structure for the sentence *John Loves Mary*.

词汇功能语法（LFG）示例

# 句法分析算法

	CFG	Dependency
确定性（编译器用）	递归下降LL（自顶向下） 移进归约LR（自底向上）	
不确定性、无概率	递归下降、移进归约、 Chart、CYK、Tomita	
有概率	Viterbi（PCFG推理） Inside-Outside（PCFG训练）	基于转换（Yamada、Nivre） 基于图（MST最大生成树）

Book	the	flight	through	Houston
S, VP, Verb Nominal, Noun [0,1]		S,VP,X2		S,VP,X2
	[0,2] Det	[0,3] NP	[0,4]	[0,5] NP
	[1,2]	[1,3] Nominal, Noun	[1,4]	[1,5] Nominal
		[2,3]	[2,4] Prep	[2,5] PP
			[3,4]	[3,5] NP, Proper- Noun
				[4,5]

CYK Parsing

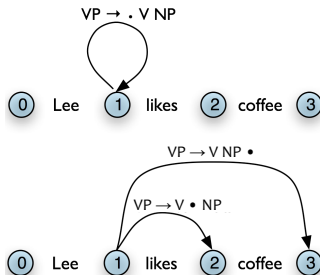


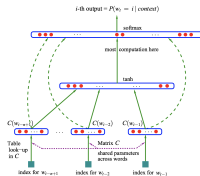
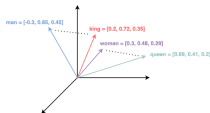
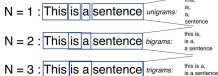
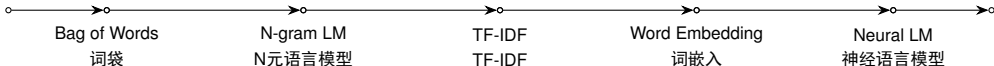
Chart Parsing

step	action	rule	stack	coverage
0				○ ○ ○ ○ ○ ○
1	S	$r_3$	[The President will]	● ● ○ ○ ○ ○
2	S	$r_1$	[The President will] [visit]	● ● ○ ○ ○ ●
3	$R_l$		[The President will visit]	● ● ○ ○ ○ ●
4	S	$r_4$	[The President will visit] [London in April]	● ● ● ● ● ●
5	$R_r$		[The President will visit London in April]	● ● ● ● ● ●

Shift-Reduce Dependency Parsing

# 分布式语义学及其影响

Firth, 1957: *You shall know a word by the company it keeps.* 你可以通过其伴随词来了解一个词的意思。



## 统计方法的兴起

- ▶ 基于语言学的方法（通常称为规则方法）在面对复杂的真实环境的语言数据时，系统性能遇到了瓶颈，难以继续提高；
- ▶ 1990年代初，IBM公司最早开始将语音识别的一些技术用在机器翻译上，开展了统计机器翻译的研究，产生了巨大的影响：
  - ▶ 当时IBM公司机器翻译负责人Fred Jelinek说了一句著名的话："Every time I fire a linguist, the performance of the speech recognizer goes up"（1998）。
  - ▶ 这句话影响很大，当然争议也很大。Fred Jelinek本人后来也给出了一些背景情况的解释。
- ▶ 统计方法为NLP带来了性能上的快速提高，但也很快遇到了瓶颈。人们再次希望引入语言学来提高系统的性能。所以这个阶段出现了更多深度语言学标注的语料库和一些更复杂的统计和语言学相结合的方法。

# 统计和语言学相结合的自然语言处理方法

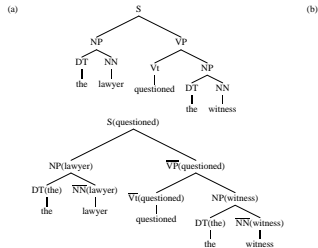


Figure 5: (a) A conventional parse tree as found for example in the Penn treebank. (b) A lexicalized parse tree for the same sentence. Note that each non-terminal in the tree now includes a single lexical item. For clarity we mark the head of each rule with an underline: for example for the rule  $NP \rightarrow DT\ NN$  the child  $NN$  is the head, and hence the  $NN$  symbol is marked as  $\overline{NN}$ .

## Lexicalized PCFG

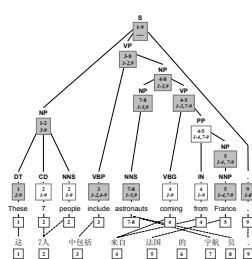


Figure 1: **Spans and complement-spans** determine what rules are extracted. Constituents in gray are members of the **frontier set**; a minimal rule is extracted from each of them.

## String-to-Tree SMT

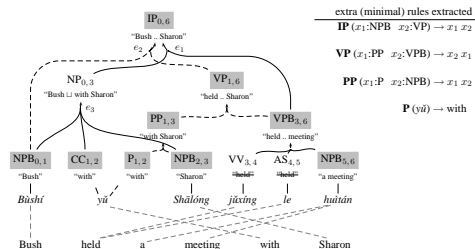


Figure 4: Forest-based rule extraction. Solid hyperedges correspond to the 1-best tree in Figure 3, while dashed hyperedges denote the alternative parse interpreting  $y\bar{u}$  as a preposition in Figure 5.

## Forest-to-String SMT



# 大语言模型的句法能力

- ▶ 我们提出在BERT中通过屏蔽一个词来观察其他词隐状态的变化，来判断一个词对另一个词的影响力。我们观察发现，词语影响力矩阵实际上已经隐含了丰富的句法结构信息。
- ▶ 最近西湖大学等单位研究发现，仅仅利用大语言模型的输出层隐状态，通过三种简单的方法即可得到接近SotA的句法分析准确率，并且具有非常好的跨领域性能。

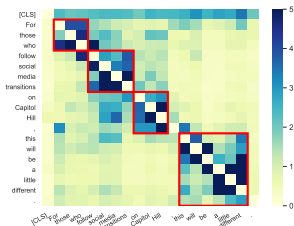


Figure 1: Heatmap of the impact matrix for the sentence "For those who follow social media transitions on Capitol Hill, this will be a little different."

Model		LR	LP	F1
Non-LLM	SEPar <sup>♡</sup>	95.56	95.89	95.72
	SAPar <sup>♡</sup>	<b>96.19</b>	<b>96.61</b>	<b>96.40</b>
	TGCN <sup>♣</sup>	96.13	96.55	96.34
	LSTM <sup>★</sup>	-	-	88.30
	Transformer <sup>★</sup>	-	-	91.20
	GPT-2 <sup>★</sup>	93.68	93.79	93.73
LLM	OPT-6.7B <sup>★</sup>	94.63	94.52	94.58
	LLaMA-7B <sup>★</sup>	95.50	95.12	95.31
	LLaMA-13B <sup>★</sup>	95.73	95.25	95.49
	LLaMA-33B <sup>★</sup>	96.05	95.56	95.81
LLM <sup>[IT]</sup>	LLaMA-65B <sup>★</sup>	<u>96.09</u>	<u>95.72</u>	<u>95.90</u>
	Alpaca-7B <sup>★</sup>	95.40	94.99	95.20
	Vicuna-7B <sup>★</sup>	95.37	94.93	95.16

Table 2: Fine-tuning results on PTB. LR: labeled recall. LP: labeled precision. <sup>♡</sup> means chart-based models. <sup>♣</sup> means transition-based models. <sup>★</sup> means sequence-based models. [IT] means instruction-tuned LLMs. The best results among all methods are **bolded** and the best sequence-based results are underlined.

# 大语言模型时代语言学对人工智能还有用吗？

- ▶ 预训练语言模型，特别是大语言模型，表现出强大的自然语言理解和生成能力，以至于现在已经很少有人再求助于基于语言学的方法来改进NLP的性能。
- ▶ 虽然大语言模型在模型设计的技术层面不再需要语言学直接接入，但我们认为，在大模型时代，语言学可以发挥重要的作用：
  - ▶ 大语言模型的数据工程：语言模型预训练数据和指令微调数据对大语言模型的能力起到决定性的作用。但语言模型的数据工程仍然处于经验性摸索阶段，缺乏明确的理论指导，在这方面语言学应该可以发挥作用；
  - ▶ 大语言模型的评价：大模型的能力评价是多方面多维度的，语言能力的评价也是其中重要组成部分，语言学应该可以在其中发挥作用；
  - ▶ 大语言模型的应用：大语言模型能力的发挥越来越取决于提示词的设计，提示词工程成为大语言模型应用的重要手段，特别是在用基于大语言模型的智能体（Agent）来解决复杂问题的时候，需要综合使用大语言模型的规划、记忆、反思、搜索、工具使用等复杂能力，在这方面，语言学可以大有可为；
  - ▶ 基于大语言模型的多智能体应用：多智能体在处理一些复杂问题的时候表现出独特的优势，但多智能体直接如何沟通协作，是影响多智能体解决问题能力的重要制约因素，在这方面，语言学应该可以起到重要的作用。

# 基于情境语义学的大模型常识推理方法

## SMART: A Situation Model for Algebra Story Problems via Attributed Grammar

Yining Hong, Qing Li, Ran Gong, Daniel Ciao, Siyuan Huang, Song-Chun Zhu

University of California, Los Angeles, USA.

yininghong@cs.ucla.edu, {liqing, nikepupu, danielciao, huangsiyuan}@ucla.edu, sczhu@stat.ucla.edu

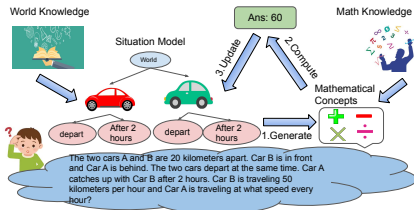


Figure 1: The process of human solving algebra story problems: We first hallucinate a situation model from the text and then perform arithmetic reasoning on the situation model to compute an answer. If we fail to generate a correct solution, we can adjust our situation model accordingly.

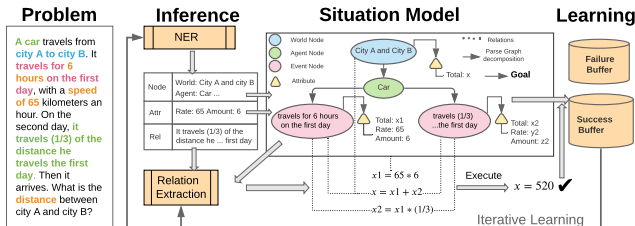


Figure 2: Overview of our SMART model. The Named Entity Recognition (NER) module extracts the spans of nodes, attributes, as well as relations from the text, and construct a parse graph using Attributed Grammar. The Relation Extraction module uses the relation spans and the parse graph already constructed to embed some relations into the parse graph. In the updated graph parser, Relation Extraction corresponds to Seq2Seq. The relations are then executed to get the final answer. If the answer is correct, it is added to the buffer of pseudo-gold parse graphs to train NER and Seq2Seq. If not, it is added to the failure set to be updated in the following iterations.

# Chomsky对ChatGPT的批评

- ▶ 诺姆·Chomsky对ChatGPT代表的人工智能研究提出了激烈的批评：

人类的思维不像 *ChatGPT* 及其同类产品那样，是一个笨拙的模式匹配统计引擎，它吞噬了数百 *TB* 的数据，并推断出最有可能的对话响应或最可能的科学问题答案。相反，人类的思维是一个令人惊讶的高效甚至优雅的系统，它使用少量信息运行；它不寻求推断数据点之间的粗暴相关性，而是创建解释。让我们停止称它为人工智能，而称它为：抄袭软件。它不会创造任何东西，只是复制艺术家的现有作品并对其进行充分修改以逃避版权法。这是自美洲原住民登陆以来欧洲定居者盗窃财产的最大事件。

- ▶ 现代人工智能领军人物Georfery Hilton对Chomsky的观点进行了明确的反驳，他的观点是“疯狂”的。
- ▶ 著名在自然语言处理学者Christopher Manning也对Chomsky的观点表示失望，认为他落后于时代了。

# Content

动机与背景

语言学对人工智能的影响

人工智能对语言学的影响

总结：大模型时代人工智能与语言学关系现状与展望

# 基于大数据的语言学研究

## 刘海涛教授计量语言学报告

发布者: xujiajin [发表时间]: 2019-06-22 [来源]: [浏览次数]: 1106

2019年6月18日下午16:00-18:00, 浙江大学刘海涛教授在110期语料库沙龙上做了题为“大数据时代语言研究的思考与践行”的学术报告。

刘教授以大数据为人类生活创造了前所未有的可量化维度为背景, 提出大数据带给语言研究的机遇。计量语言学越来越受到学界青睐。同时, 刘教授也向大家展示了基于语言大数据研究的挑战, 即自然语言处理对于语言学家作用及贡献的争议。随后, 刘教授介绍了基于多语言依存距离进行的相关研究, 展示了大数据时代计量语言学研究的最新成果, 包括通过对依存距离的计算, 揭示认知规律及语言普遍性、生态多样性与语言多样性之间的关系等。

报告后, 刘教授与现场师生就大数据分析在商务文本中的应用、翻译文本质量测量、以及依存距离与短时记忆之间关系的研究等问题展开了讨论与互动。



4/6/27 00:14

Haitao Liu - Google 学术搜索



### Haitao Liu

Professor of Linguistics, Zhejiang University  
Quantitative Linguistics  
Digital Humanities  
Dependency Grammar  
Language Planning  
Interlinguistics

	总计	2019 年至今
引用	4543	2564
h 指数	32	25
i10 指数	82	56

2 篇文章

5 篇文章

无法查看的文章

可查看的文章

根据资助方的强制性开放获取政策

标题	引用次数	年份
Dependency distance as a metric of language comprehension difficulty H Liu Journal of Cognitive Science 9 (2), 159-191	403	2008
Approaching human language with complex networks J Cong, H Liu Physics of life reviews 11 (4), 598-618	275	2014
Dependency distance: A new perspective on syntactic patterns in natural languages H Liu, C Xu, J Liang Physics of life reviews 21, 171-193	274	2017
Dependency direction as a means of word-order typology: A method based on dependency treebanks H Liu Lingua 120 (6), 1567-1578	173	2010
The effects of sentence length on dependency distance, dependency direction and the implications—based on a parallel English–Chinese dependency treebank J Jiang, H Liu Language Sciences 50, 93-104	164	2015

# Dependency direction as a means of word-order typology: A method based on dependency treebanks

	VS	SV	VO	OV	NAdj	AdjN	WALS
Arabic (ara)	61.4 (2153)	38.6 (1351)	91 (5313)	9 (524)	95.9 (3953)	4.1 (167)	VS-VO-NAdj
Bulgarian (bul)	18.5 (3,036)	81.5 (13,417)	90.1 (6224)	9.9 (682)	1.6 (180)	98.4 (11,212)	?-VO-AdjN
Catalan (cat)	18.5 (4584)	81.5 (20,221)	85.5 (19,080)	14.5 (3239)	99.2 (1680)	0.8 (14)	?-VO-NAdj
Chinese (chi)	1.3 (19)	98.7 (1400)	98 (1679)	2 (34)	0.4 (2)	99.6 (461)	SV-VO-AdjN
Czech (cze)	27.4 (34,273)	72.6 (90,841)	72.9 (74,583)	27.1 (27,735)	8.6 (11,521)	91.4 (122,004)	SV-VO-AdjN
Danish (dan)	19.8 (1015)	80.2 (4122)	99.1 (8739)	0.9 (81)	60 (1683)	40 (1124)	SV-VO-AdjN
Dutch (dut)	28.7 (13,258)	71.3 (33,000)	82.5 (71,030)	17.5 (15,085)	7.4 (2024)	92.6 (25,207)	SV-?-AdjN
Greek (ell)	34.7 (1609)	65.3 (3029)	80.5 (3437)	19.5 (834)	8.4 (400)	91.6 (4345)	?-VO-AdjN
English (eng)	3.2 (1116)	96.8 (33,916)	93.5 (28,219)	6.5 (1959)	2.6 (661)	97.4 (24,801)	SV-VO-AdjN
Basque (eus)	20.4 (765)	79.6 (2990)	12.8 (381)	87.2 (2589)	78 (1234)	22 (349)	SV-OV-NAdj
German (ger)	33.2 (17,382)	66.8 (34,938)	36.8 (9447)	63.2 (16,237)	37.1 (15,355)	62.9 (26,016)	SV-?-AdjN
Hungarian (hun)	26.6 (1764)	73.4 (4862)	47.8 (2600)	52.2 (2843)	2.3 (339)	97.7 (14,239)	SV-?-AdjN
Italian (ita)	24.5 (869)	75.5 (2681)	82.3 (2090)	17.7 (451)	60.9 (2374)	39.1 (1523)	?-VO-NAdj
Japanese (jpn)	0	100 (5509)	0	100 (27,553)	0	100 (3820)	SV-OV-AdjN
Portuguese (por)	15.7 (1899)	84.3 (10,190)	85.1 (9447)	14.9 (1656)	70.1 (5858)	29.9 (2495)	SV-VO-NAdj
Romanian (rum)	21.9 (648)	78.1 (2313)	88.3 (1568)	11.7 (208)	66.9 (2905)	33.1 (1439)	SV-VO-NAdj
Slovenian (slv)	38.9 (658)	61.1 (1035)	74.5 (2375)	25.5 (815)	11 (189)	89 (1534)	SV-VO-AdjN
Spanish (spa)	21.5 (1107)	78.5 (4032)	77.3 (3417)	22.7 (1006)	98 (431)	2 (9)	?-VO-NAdj
Swedish (swe)	22.7 (4296)	77.3 (14,589)	94.6 (10,411)	5.4 (596)	0.4 (26)	99.6 (6656)	SV-VO-AdjN
Turkish (tur)	8.1 (284)	91.9 (3208)	4 (255)	96 (6175)	0.3 (11)	99.7 (3514)	SV-OV-AdjN

<sup>a</sup> In fact, here we use *dominant word order* unlike the definition in Croft (2002:60), and closer to the understanding of basic word order in Whale (1997:100). In other words, it only shows that one of the word order types is more frequent (or dominant) in language use. Dryer (2008a) points out that WALS also uses the *dominant word order* in this meaning, to emphasize that priority is given to the criterion of what is more frequent in language use

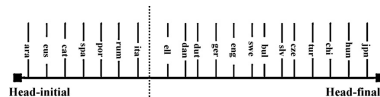


Fig. 5. 20 languages in Tesnière's typological classification system.

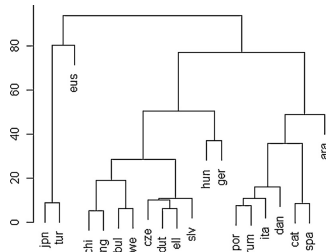


Fig. 10. Clustering of observations for 20 languages.

# 基于概率模型的古文字破解

## A Computational Approach to Deciphering Unknown Scripts

Kevin Knight

USC/Information Sciences Institute  
4676 Admiralty Way  
Marina del Rey, CA 90292  
knight@isi.edu

Kenji Yamada

USC/Information Sciences Institute  
4676 Admiralty Way  
Marina del Rey, CA 90292  
kyamada@isi.edu



Figure 1: The Phaistos Disk (c. 1700BC). The disk is six inches wide, double-sided, and is the earliest known document printed with a form of movable type.

B → b or v	r → r
D → d	t → t
G → g	tS → c h
J → ñ	u → u or ú
L → l l or y	x → j
a → a or á	nothing → h
b → b or v	T (followed by a, o, or u) → z
d → d	T (followed by e or i) → c or z
e → e or é	T (otherwise) → c
f → f	k (followed by e or i) → q u
g → g	k (followed by s) → x
i → i or í	k (otherwise) → c
l → l	rr (at beginning of word) → r
m → m	rr (otherwise) → rr
n → n	s (preceded by k) → nothing
o → o or ó	s (otherwise) → s
p → p	

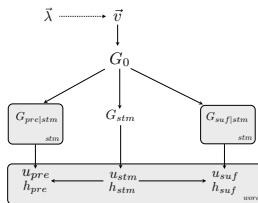
## A Statistical Model for Lost Language Decipherment

Benjamin Snyder and Regina Barzilay  
CSAIL

Massachusetts Institute of Technology  
{bsnyder, regina}@csail.mit.edu

Kevin Knight  
ISI

University of Southern California  
knight@isi.edu



在本文中，我们提出了一种自动破译丢失语言的方法。给定一个已知相关语言的非平行语料库，我们的模型既产生字母映射，又产生单词到其相应同源词的翻译。我们采用非参数贝叶斯框架来同时捕获低级字符映射和高级词素对应。当应用于古代闪米特语乌加里特语时，该模型正确地将30个字母中的29个映射到它们的希伯来语对应物，并推断出60%的乌加里特语单词的正确希伯来语同源词，这些单词在希伯来语中有同源词。



# 基于多智能体的语言发生和演化

## Emergence and evolution of language in multi-agent systems

Dorota Lipowska<sup>a,\*</sup>, Adam Lipowski<sup>b</sup>

<sup>a</sup> Faculty of Modern Languages and Literature, Adam Mickiewicz University, Poznań, Poland

<sup>b</sup> Faculty of Physics, Adam Mickiewicz University, Poznań, Poland

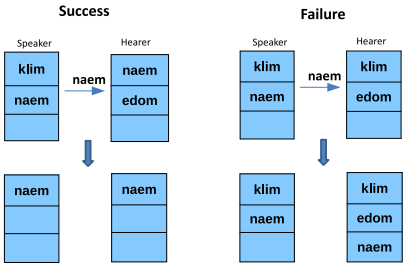


Fig. 1. An elementary step in the single-object version of the naming game.

Naming Game

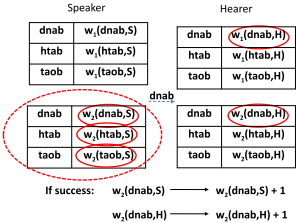


Fig. 2. An elementary step in a 2-object version of the signaling game model with reinforcement learning (Lipowska and Lipowski, 2018). The speaker randomly chooses an object (the corresponding section of the inventory is encircled by a dotted line). Using the relevant weights (in solid circles), the speaker selects one of its words (here: "dnab"). Next the hearer tries to guess the object the speaker is talking about, taking into account the weights of the communicated word (in circles). If the hearer's guess is correct, both agents increase their corresponding weights by 1. Otherwise, the weights remain unchanged.

Signaling Game with Reinforcement Learning

# Content

动机与背景

语言学对人工智能的影响

人工智能对语言学的影响

总结：大模型时代人工智能与语言学关系现状与展望

## 总结：大模型时代人工智能与语言学关系现状与展望

- ▶ 早期对人工智能影响重大的一些语言学思想，和人工智能的诞生几乎同步；
- ▶ 一些重要的语言学思想和理论，对人工智能，特别是计算语言学的发展产生了深刻的影响；
- ▶ 自然语言处理领域产生了一大批语言学驱动的语料库、算法和应用，包括树库、句法分析、信息抽取、机器翻译等；
- ▶ 统计方法兴起以后，一方面语言学的作用被削弱，另一方面，语言学的作用也更被强调，用来解决统计方法难以奏效的复杂问题；
- ▶ 神经网络，特别是大语言模型，表现出强大的语言理解和生成能力：
  - ▶ 一方面，传统的基于语言学的方法在人工智能研究中能起到的作用越来越弱，甚至逐步被边缘化了；
  - ▶ 另一方面，我们认为大语言模型的发展也为语言学在人工智能中的作用提供了很多新的机会，包括大语言模型的数据工程、大语言模型的评价、基于大语言模型的多智能体应用等；
- ▶ 以Chomsky为代表的部分语言学家对以ChatGPT为代表的现代人工智能方法持强烈的反对态度，认为这些方法并没有揭示语言的本质；
- ▶ 人工智能的飞速发展，为语言学研究提供的新的手段和强大的工具，在多方面推动了语言学的发展。

# Thank you!

把数字世界带入每个人、每个家庭、  
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and organization  
for a fully connected, intelligent world.

Copyright©2018 Huawei Technologies Co., Ltd.  
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

