2017全球人工智能技术大会

自然语言处理中的技术评测

刘　挺
哈尔滨工业大学社会计算与信息检索研究中心
2017年5月22日

提纲

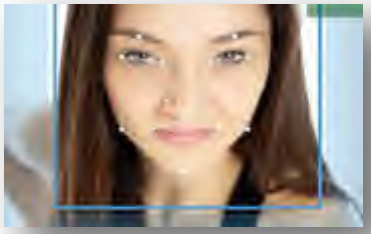
- **概述**
- 句法语义分析评测
- 阅读理解评测
- 人机对话评测
- 总结

自然语言处理（NLP）是典型的认知智能

认知智能

“预测决策”智能

“创新”智能



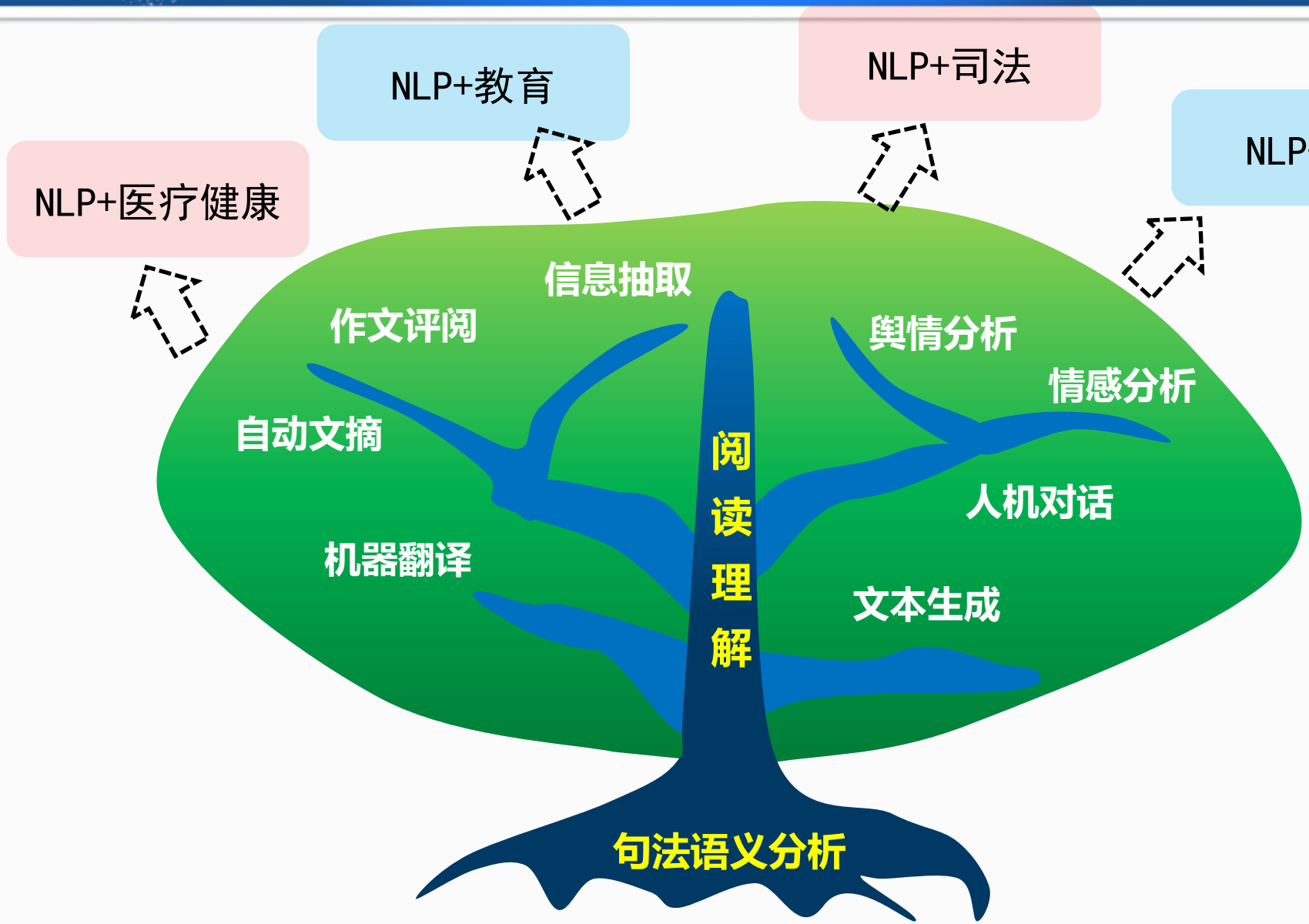
感知智能

运算智能

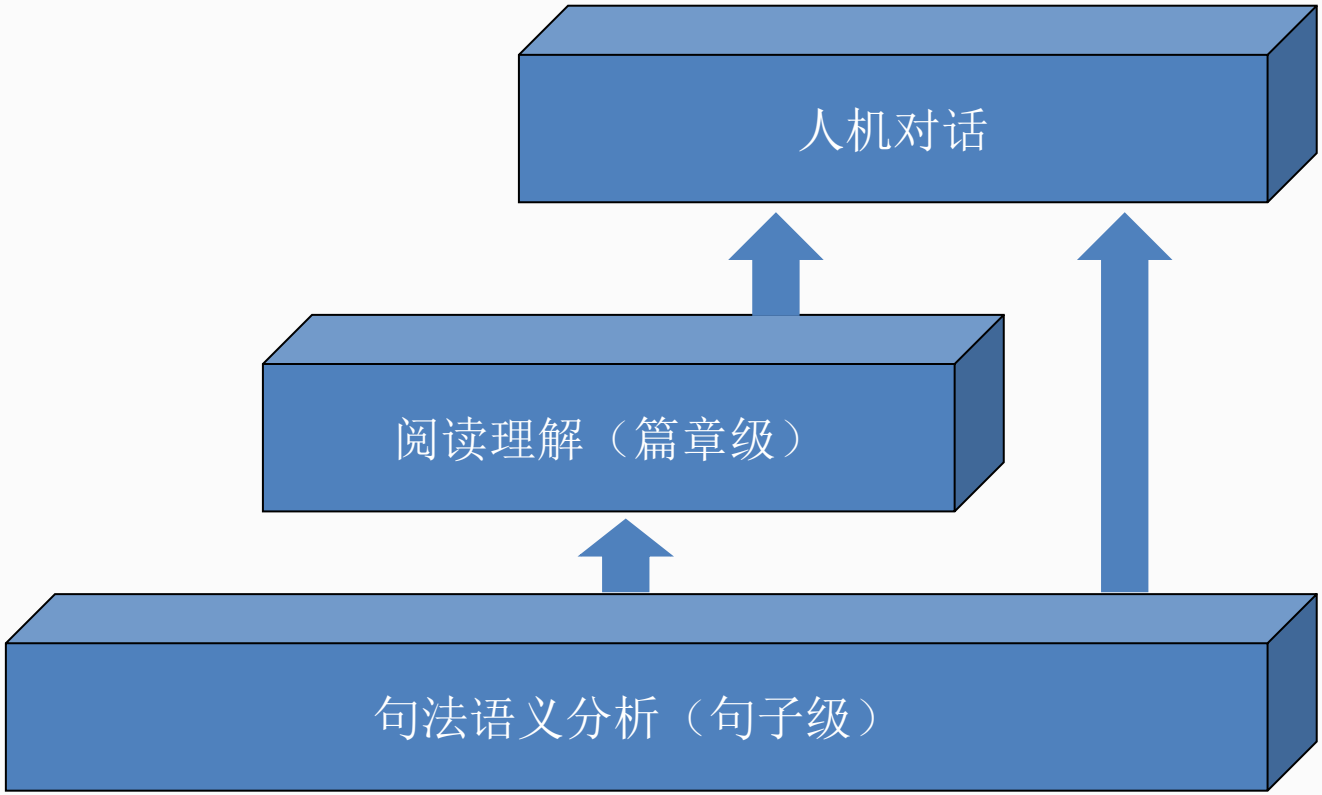
自然语言处理问题的特点

- 与感知智能中的语音处理相比，自然语言处理有以下特点：
  - （1）问题纷繁复杂
  - （2）难度大，触及常识、推理等认知能力
  - （3）存在大量可以利用的先验知识
  - （4）与行业深入结合
  - （5）部分课题评测难度高

自然语言处理课题的分布



NLP领域三项重要技术



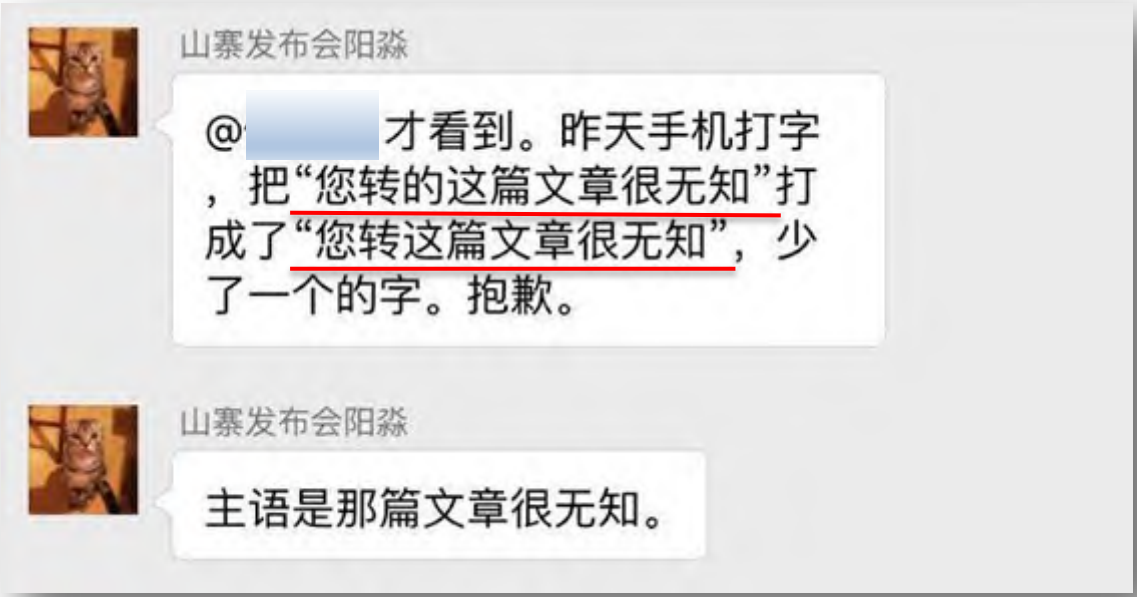
评测驱动技术进步

- 技术评测一直是驱动自然语言处理技术发展的重要手段
- 美国：NIST、DARPA
- 日本：NTCIR
- 中国：
  - 国家863计划（2005年以前）
  - 学会： CWMT、CCIR、NLPCC、SMP、CCL
- 我的一个理念：**中文评测由中国人引导**

提纲

- 概述
- **句法语义分析评测**
- 阅读理解评测
- 人机对话评测
- 总结

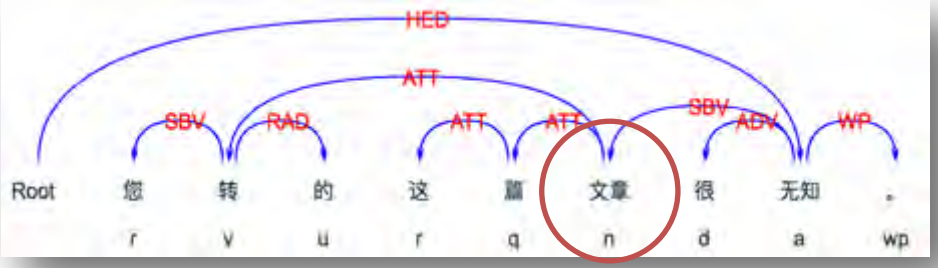
为什么需要句法语义分析？



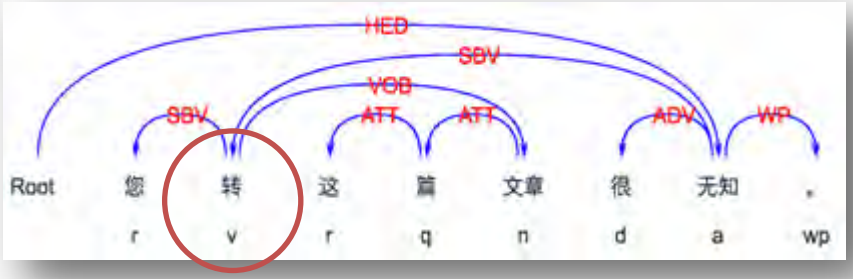
- 很多应用需要考虑词语之间的组合关系
- 句法语义分析很好的刻画了这种关系

句法语义分析

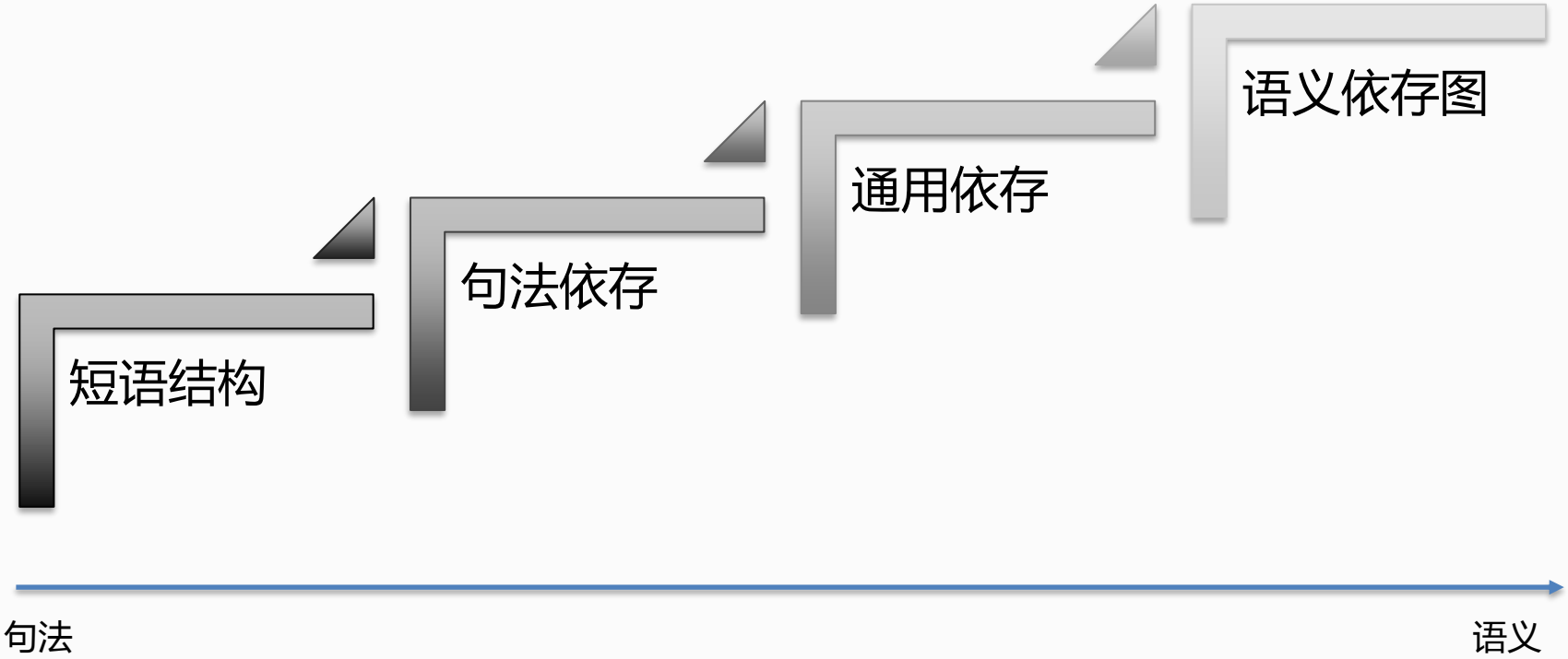
- 按照某种语法体系，将句子从词语的序列形式转化为图结构（通常为树结构）
- 刻画句子内部的句法语义关系（如主谓宾等）



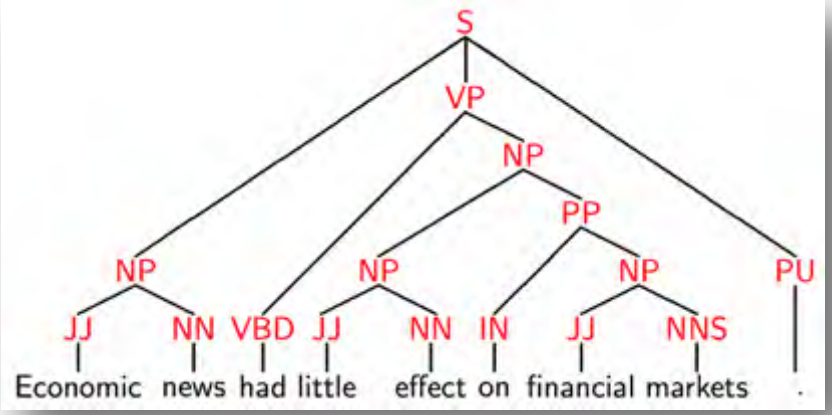
http://www.ltp-cloud.com/demo/



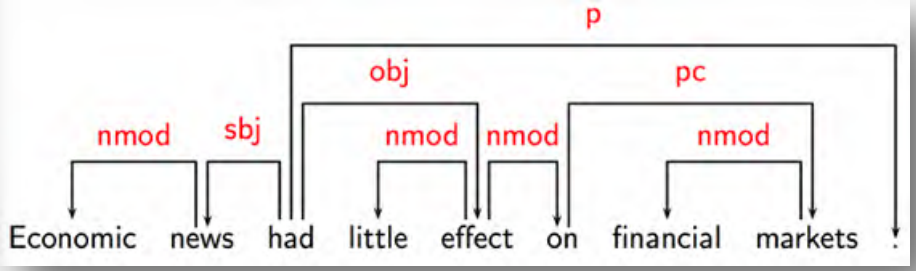
句法语义分析进展



短语结构与句法依存



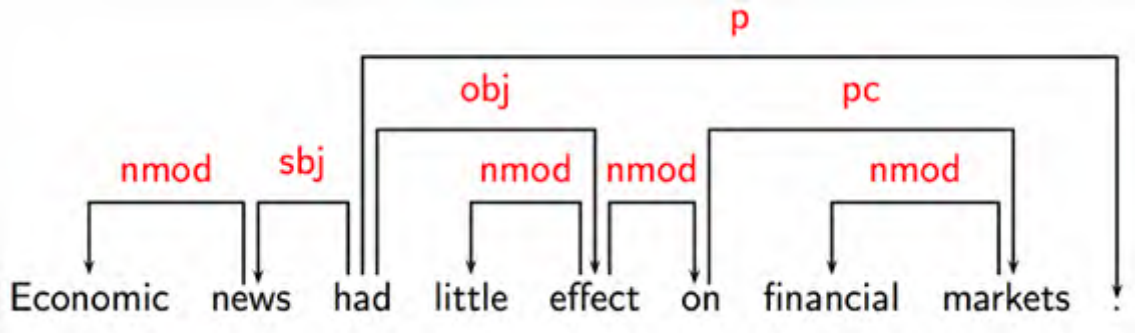
短语结构

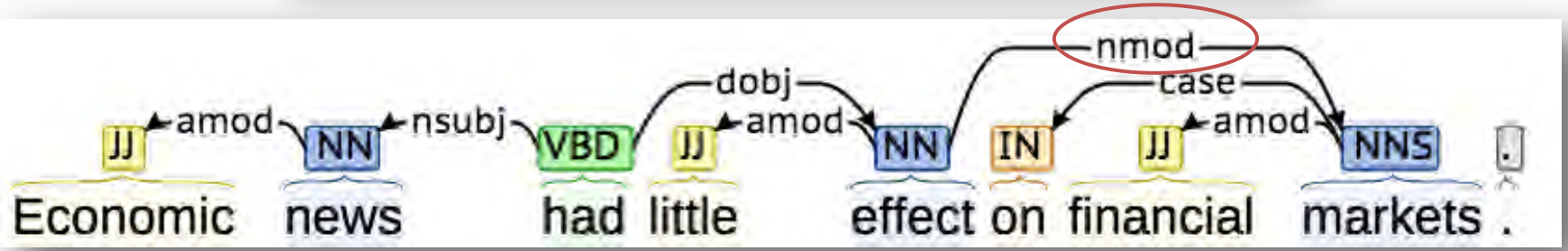


句法依存

- 依存分析表示更直观，标注更方便

句法依存与通用依存

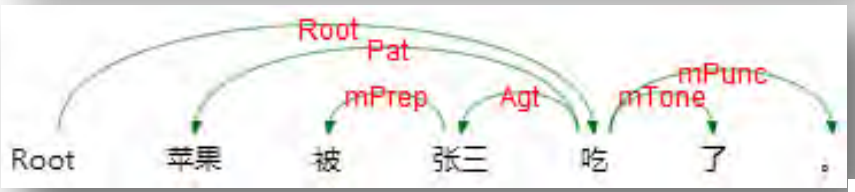


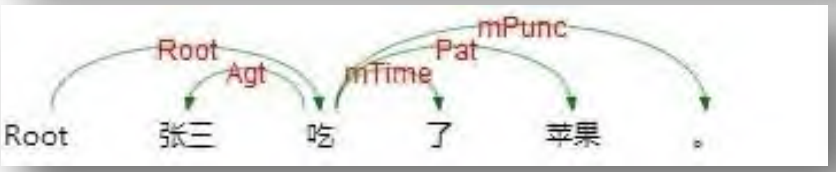


http://nlp.stanford.edu/software/stanford-dependencies.shtml

- 通用依存（Universal Dependencies）直接标注实词之间的关系

语义依存树/图





语义依存树

语义依存图

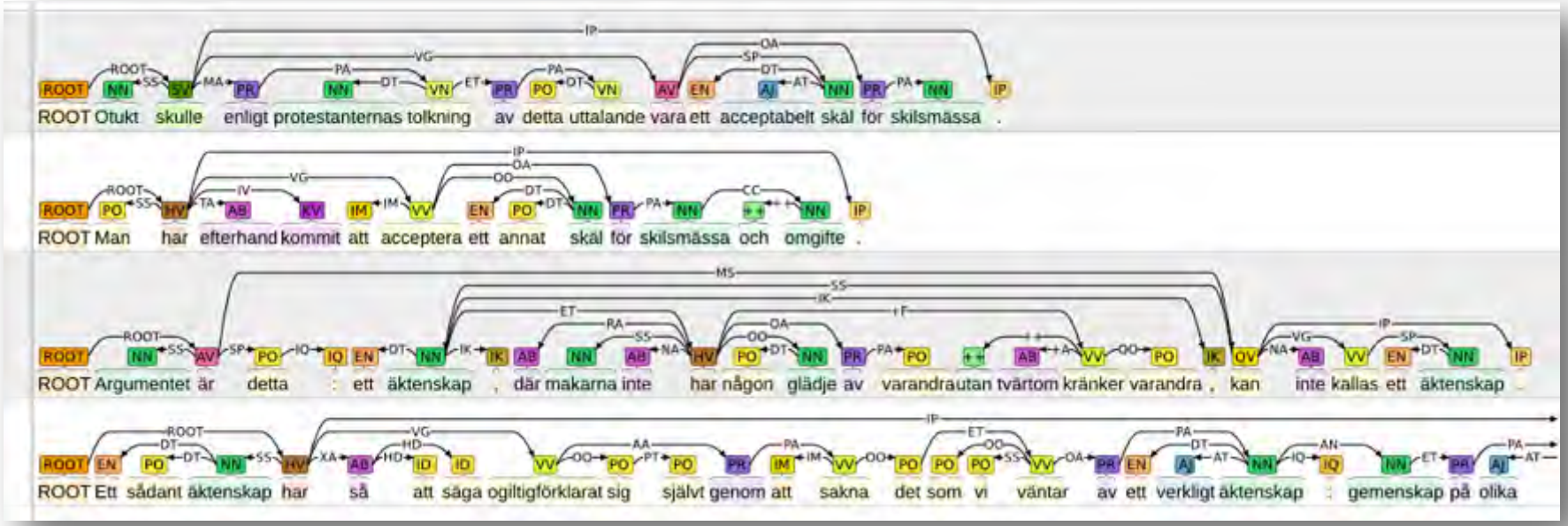
Wanxiang Che, Meishan Zhang, Yanqiu Shao, Ting Liu. **SemEval-2012 Task 5**: Chinese Semantic Dependency Parsing.

Wanxiang Che, Yu Ding, Yanqiu Shao, Ting Liu. **SemEval-2016 Task 9**: Chinese Semantic Dependency Parsing.

相关评测

多语依存句法分析

- CoNLL 2006，2007 评测
  - http://ilk.uvt.nl/conll/
  - 多语依存句法分析
  - 10 - 12 种语言



多语句法和语义联合分析

- CoNLL 2009 评测
  - http://ufal.mff.cuni.cz/conll2009-st/
  - 多语依存句法分析和语义角色标注
  - 7 种语言
  - 哈工大获得总成绩第一名

Rank	System	Average	Catalan	Chinese	Czech	English	German	Japanese	Spanish
1	Che	82.64	81.84	76.38	83.27	87.00	82.44	85.65	81.90
2	Chen	82.52	83.01	76.23	80.87	87.69	81.22	85.28	83.31
3	Merlo	82.14	82.66	76.15	83.21	86.03	79.59	84.91	82.43
4	Bohnet	80.85	80.44	75.91	79.57	85.14	81.60	82.51	80.75
5	Asahara	78.43	75.91	73.43	81.43	86.40	69.84	84.86	77.12
6	Brown	77.27	77.40	72.12	75.66	83.98	77.86	76.65	77.21
7	Zhang	76.49	75.00	73.42	76.93	82.88	73.76	78.17	75.25
8	Dai	73.98	72.09	72.72	67.14	81.89	75.00	80.89	68.14
9	Lu Li	73.97	71.32	65.53	75.85	81.92	70.93	80.49	71.72
10	Lluís	71.49	56.64	66.18	75.95	81.69	72.31	81.76	65.91
11	Vallejo	70.81	73.75	67.16	60.50	78.19	67.51	77.75	70.78
12	Ren	67.81	59.42	75.90	60.18	77.83	65.77	77.63	57.96
13	Zeman	51.07	49.61	43.50	57.95	50.27	49.57	57.69	48.90

多领域依存句法分析

- Syntactic Analysis of Non-Canonical Language (SANCL) 2012 评测
  - https://sites.google.com/site/sancl2012/
  - 非规范语言句法分析（Google Web Treebank）
  - Google 组织
  - 我们获得第二、三名

Dependency Parsing Results:

	Domain A (answers)			Domain B (newsgroups)			Domain C (reviews)			Domain D (wsj)			Average (A-C)		
Team	LAS	UAS	POS	LAS	UAS	POS	LAS	UAS	POS	LAS	UAS	POS	LAS	UAS	POS
*Zhang&Nivre**	*76.60*	*81.59*	*89.74*	*81.62*	*85.19*	*91.17*	*78.10*	*83.32*	*89.60*	*89.37*	*91.46*	*96.84*	*78.77*	*83.37*	*90.17*
UPenn	68.54	82.28	89.65	74.41	86.10	90.99	70.17	82.88	89.02	81.74	91.99	96.93	71.04	83.75	89.89
UMass	72.51	78.36	89.42	77.23	81.61	91.28	74.89	80.34	89.90	81.15	83.97	94.71	74.88	80.10	90.20
NAIST	73.54	79.89	89.92	79.83	84.59	91.39	75.72	81.99	90.47	87.95	90.99	97.40	76.36	82.16	90.59
IMS-2	74.43	80.77	89.50	79.63	84.29	90.72	76.55	82.18	89.41	86.88	89.90	97.02	76.87	82.41	89.88
IMS-3	75.90	81.30	88.24	79.77	83.96	89.70	77.61	82.38	88.15	86.02	88.89	95.14	77.76	82.55	88.70
IMS-1	78.33	83.20	91.07	83.16	86.86	91.70	79.02	83.82	90.01	90.82	92.73	97.57	80.17	84.63	90.93
Copenhagen	78.12	82.91	90.42	82.90	86.59	91.15	79.58	84.13	89.83	90.47	92.42	97.25	80.20	84.54	90.47
Stanford-2	77.50	82.57	90.30	83.56	87.18	91.49	79.70	84.37	90.46	89.87	91.95	95.00	80.25	84.71	90.75
HIT-Baseline	80.75	85.84	90.99	85.26	88.90	92.32	81.60	86.60	90.65	**91.88**	**93.88**	**97.76**	82.54	87.11	91.32
HIT-Domain	80.79	85.86	90.99	85.18	88.81	92.32	81.92	86.80	90.65	91.82	93.83	97.76	82.63	87.16	91.32
Stanford-1	81.01	85.70	90.30	**85.85**	**89.10**	91.49	82.54	86.73	90.46	91.50	93.38	95.00	83.13	87.18	90.75
DCU-Paris13	**81.15**	**85.80**	**91.79**	85.38	88.74	**93.81**	**83.86**	**88.31**	**93.11**	89.67	91.79	97.29	**83.46**	**87.62**	**92.90**

通用依存句法分析

- CoNLL 2017 评测
  - http://universaldependencies.org/conll17/
  - Multilingual Parsing from Raw Text to Universal Dependencies
  - 45种语言、64个领域的树库
  - 统一的词性和句法标注体系

CoNLL 2017 评测结果

- 113支队伍报名
  - 包括卡内基梅隆大学、华盛顿大学、多伦多大学、牛津大学、爱丁堡大学、东京大学、IBM研究院、Facebook
  - 包括中科院、复旦大学、上海交大
- 33支队伍提交系统
- 前三名：美国斯坦福大学、美国康奈尔大学、德国斯图加特大学
- 哈工大获得第四名

语义依存分析

- 我们组织了 SemEval 2012 and 2016 评测
  - https://www.cs.york.ac.uk/semeval-2012/task5.html
  - http://alt.qcri.org/semeval2016/task9/
  - 中文语义依存分析
- SemEval 2014 and 2015 评测
  - http://alt.qcri.org/semeval2014/task8/
  - http://alt.qcri.org/semeval2015/task18/
  - 英文语义依存分析

语言技术平台

在线演示

直观了解语言技术平台能为你做什么

我们即将以昂扬的斗志迎来新的一年。

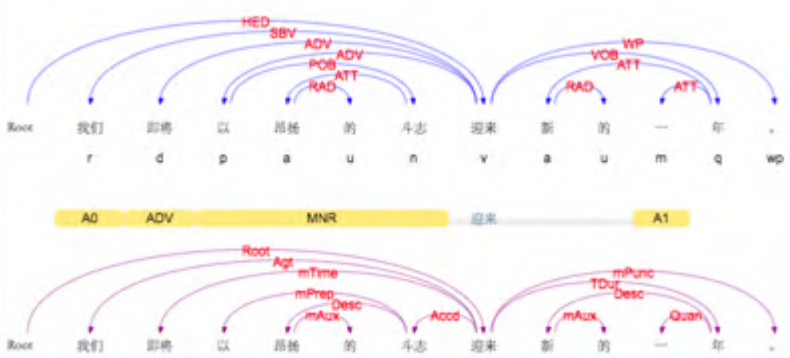
样例

分析

句子视图 篇章视图 XML视图

☑ 词性标注 ☑ 命名实体 ☑ 句法分析 ☑ 语义角色标注 ☐ 语义依存（树）分析 ☑ 语义依存（图）分析

段落1句子1:我们即将以昂扬的斗志迎来新的一年。



标签释义

语义依存分析 依存句法分析 词性标注 语义角色标注

Tag	关系类型	Description
Accd	依据角色	According
Agt	施事关系	Agent
Datv	源事关系	Dative
Desc	描写角色	Description
Dir	趋向角色	Direction
Feat	描写角色	Description
mAux	的字标记	Auxiliary
mPrep	介词标记	Preposition
mPunc	标点标记	Punctuation Marker
mTime	时间标记	Time
Quan	数量角色	Quantity
Root	根节点	Root

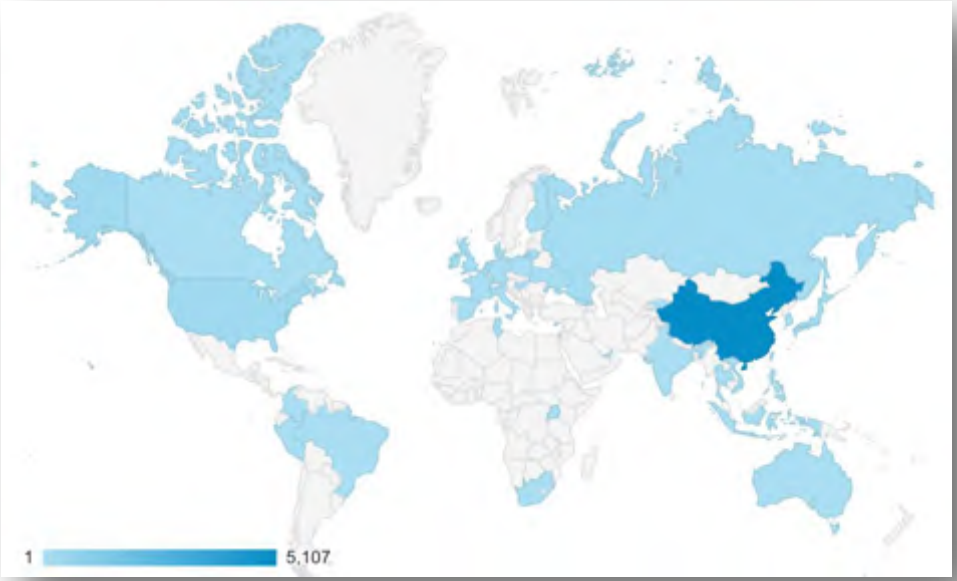
在文档中查看全部标签信息

商业授权企业

Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform. In Proceedings of the **Coling 2010**: Demonstrations. 2010.08, pp13-16, Beijing, China.

“语言云”统计

- http://www.ltp-cloud.com/
- 注册开发者：1.1万余名
- 日均处理请求：70余万次



用户分布

提纲

- 概述
- 句法语义分析评测
- **阅读理解评测**
- 人机对话评测
- 总结

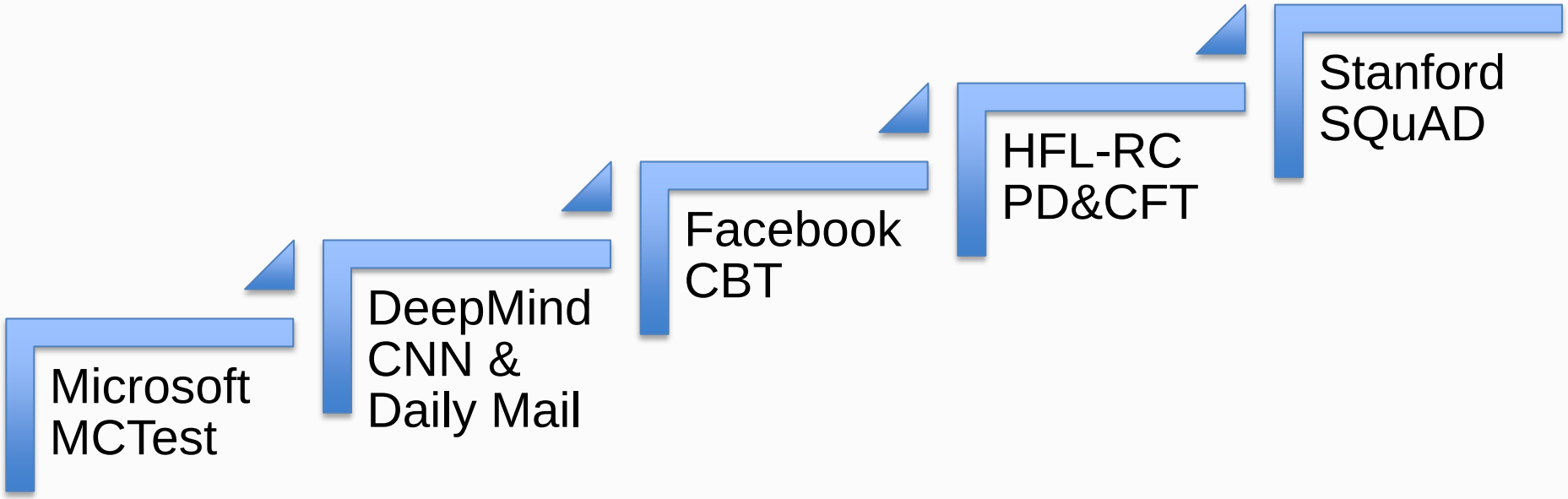
儿童阅读理解系统演示

- 哈工大讯飞联合实验室（HFL）研制

儿童阅读理解系统演示

- 哈工大讯飞联合实验室（HFL）研制

阅读理解数据集发展



国际人工智能巨头纷纷投入“阅读理解”的研究

阅读理解数据集发展

- 阅读理解数据集的变化：Microsoft MCTest

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?

A) Fries

B) Pudding

C) James

D) Jane

MCTest

- MCTest (EMNLP2013)
  - 问题类型：选择题
  - 答案类型：一段文本描述（共4个选项）
  - 数据标注：人工标注
  - 数据规模：500个篇章*4个问题 ≈ 2000个问题
  - 评价方法：准确率（与正确选项是否完全一致）

MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text

Matthew Richardson
Microsoft Research
One Microsoft Way
Redmond, WA 98052
mattri@microsoft.com

Christopher J.C. Burges
Microsoft Research
One Microsoft Way
Redmond, WA 98052
cburges@microsoft.com

Erin Renshaw
Microsoft Research
One Microsoft Way
Redmond, WA 98052
erinren@microsoft.com

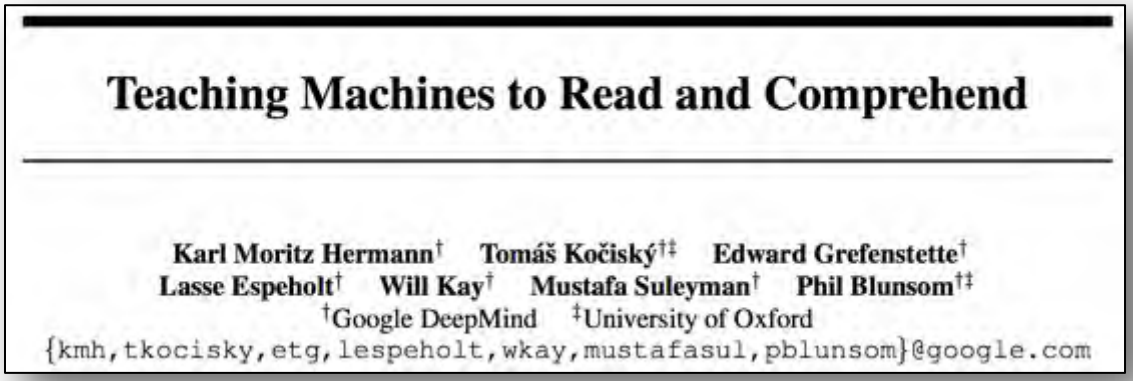
阅读理解数据集发展

- 阅读理解数据集的变化：DeepMind CNN&Daily Mail

Original Version	Anonymised Version
Context	
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ...	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " ...
Query	
Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer	
Oisin Tymon	*ent193*

CNN / Daily Mail

- DeepMind CNN / Daily Mail (NIPS2015)
  - 问题类型：完形填空类
  - 答案类型：篇章中的单个词（名实体）
  - 数据标注：机器自动
  - 数据规模：38.0万 / 87.9万 问题
  - 评价方法：准确率（与正确答案是否完全一致）


Teaching Machines to Read and Comprehend

Karl Moritz Hermann† Tomáš Kočiský†‡ Edward Grefenstette†
Lasse Espeholt† Will Kay† Mustafa Suleyman† Phil Blunsom†‡
†Google DeepMind ‡University of Oxford
{kmh,tkocisky,etg,lespeholt,wkay,mustafasul,pblunsom}@google.com

CNN / Daily Mail

	CNN News		CBTest NE		CBTest CN	
	Valid	Test	Valid	Test	Valid	Test
Deep LSTM Reader (Hermann et al., 2015)	55.0	57.0	-	-	-	-
Attentive Reader (Hermann et al., 2015)	61.6	63.0	-	-	-	-
Human (context+query) (Hill et al., 2015)	-	-	-	*81.6*	-	*81.6*
MemNN (window + self-sup.) (Hill et al., 2015)	63.4	66.8	70.4	66.6	64.2	63.0
AS Reader (Kadlec et al., 2016)	68.6	69.5	73.8	68.6	68.8	63.4
CAS Reader (Cui et al., 2016)	68.2	70.0	74.2	69.2	68.2	65.7
Stanford AR (Chen et al., 2016)	72.4	72.4	-	-	-	-
GA Reader (Dhingra et al., 2016)	73.0	73.8	74.9	69.0	69.0	63.9
Iterative Attention (Sordoni et al., 2016)	72.6	73.3	75.2	68.6	72.1	69.2
Bengio → EpiReader (Trischler et al., 2016)	**73.4**	74.0	75.3	69.7	71.5	67.4
AoA Reader	73.1	**74.4**	**77.8**	**72.0**	**72.2**	**69.4**
AoA Reader + reranking	-	-	**79.6**	**74.0**	**75.7**	**73.1**
MemNN (Ensemble)	66.2	69.4	-	-	-	-
AS Reader (Ensemble)	73.9	75.4	74.5	70.6	71.1	68.9
GA Reader (Ensemble)	76.4	77.4	75.5	71.9	72.1	69.4
EpiReader (Ensemble)	-	-	76.6	71.8	73.6	70.6
Bengio → Iterative Attention (Ensemble)	74.5	75.7	76.9	72.0	74.1	**71.0**
AoA Reader (ensemble)	-	-	**78.9**	**74.5**	**74.7**	70.8
AoA Reader (ensemble + reranking)	-	-	**80.3**	**75.6**	**77.0**	**74.1**

阅读理解数据集发展

- 阅读理解数据集的变化：Facebook CBTest

"Well, Miss Maxwell, I think it only fair to tell you that you may have trouble with those boys when they do come. Forewarned is forearmed, you know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and it is hard to corner him."

"Are the boys big ?" queried Esther anxiously.

"Yes. Thirteen and fourteen and big for their age. You can't whip 'em -- that is the trouble. A man might, but they'd twist you around their fingers. You'll have your hands full, I'm afraid. But maybe they'll behave all right after all."

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that Mr. Baxter had exaggerated matters a little.

S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 '' Are the boys big ? ''
8 queried Esther anxiously .
9 '' Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly about her school and her work , hoped she was getting on well , and said he had two young rascals of his own to send soon .
20 Esther felt relieved .

Q: She thought that Mr. _____ had exaggerated matters a little .

C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

a: Baxter

Children's Book Test

- Facebook CBTest (ICLR2016)
  - 问题类型：完形填空类
  - 答案类型：篇章中的单个词（实体、普通名词、动词、介词）
  - 数据标注：机器自动
  - 数据规模：约70万个问题
  - 评价方法：准确率（与正确答案是否完全一致）

THE GOLDILOCKS PRINCIPLE: READING CHILDREN'S BOOKS WITH EXPLICIT MEMORY REPRESENTATIONS

Felix Hill,* Antoine Bordes, Sumit Chopra & Jason Weston
Facebook AI Research
770 Broadway
New York, USA
felix.hill@cl.cam.ac.uk,{abordes,spchopra,jase}@fb.com

CHILDREN'S BOOK TEST

- 填空型阅读理解任务上大幅超过DeepMind, Facebook, IBM, CMU等机构，且超过state-of-the-art (ours)

	CBTest NE		CBTest CN	
	Valid	Test	Valid	Test
Human (Hill et al., 2015)	-	*81.6*	-	*81.6*
MemNN (Hill et al., 2015)	70.4	66.6	64.2	63.0
AS Reader (Kadlec et al., 2016)	73.8	68.6	68.8	63.4
GA Reader (Dhingra et al., 2016)	74.9	69.0	69.0	63.9
Iterative Attention (Sordoni et al., 2016)	75.2	68.6	72.1	69.2
AoA Reader (Cui et al., 2016)	77.8	72.0	72.2	69.4
NSE Adp. Com. (Munkhdalai and Yu, 2016)	78.2	73.2	74.2	71.4
GA Reader + Fine-gating (Yang et al., 2016)	79.1	*75.0*	75.3	72.0
AoA Reader + Re-ranking (Cui et al., 2016)	*79.6*	74.0	*75.7*	*73.1*
AoA Reader$^+$ (GRU)	78.0	73.8	72.8	69.8
AoA Reader$^+$ (CRU)	79.5	75.4	74.4	71.3
AoA Reader$^+$ (CRU) + Re-ranking	**80.6**	**76.1**	**76.6**	**74.5**
AS Reader (Ensemble)	74.5	70.6	71.1	68.9
Iterative Attention (Ensemble)	76.9	72.0	74.1	71.0
AoA Reader (Ensemble)	78.9	74.5	74.7	70.8
AoA Reader (Ensemble + Re-ranking)	*80.3*	*75.7*	*77.0*	*74.1*
AoA Reader$^+$ (CRU) (Ensemble)	80.0	77.1	77.0	73.5
AoA Reader$^+$ (CRU) (Ensemble + Re-ranking)	**81.8**	**77.5**	**79.0**	**76.8**

Table 3: Results on the CBTest NE (named entity) and CN (common noun) datasets.

国内外阅读理解研究现状

- 阅读理解数据集的变化：HFL-RC PD&CFT

Document	1 ||| 人民网 1月 1日 讯 据 《纽约 时报》 报道， 美国 华尔街 股市 在 2013年 的 最后 一 天 继续 上涨， 和 全球 股市 一样， 都 以 最高 纪录 或 接近 最高 纪录 结束 本年 的 交易。 2 ||| 《纽约 时报》 报道 说， 标普 500 指数 今年 上升 29.6%， 为 1997年 以来 的 最 大 涨幅 ； 3 ||| 道琼斯 工业 平均 指数 上升 26.5%， 为 1996年 以来 的 最 大 涨幅 ； 4 ||| 纳斯达克 上涨 38.3%。 5 ||| 就 12月 31日 来说， 由于 就业 前景 看好 和 经济 增长 明年 可能 加速， 消费者 信心 上升。 6 ||| 工商 协进会 报告， 12月 消费者 信心 上升 到 78.1， 明显 高于 11月 的 72。 7 ||| 另 据 《华尔街 日报》 报道， 2013年 是 1995年 以来 美国 股市 表现 最 好 的 一 年。 8 ||| 这 一 年 里， 投资 美国 股市 的 明智 做法 是 追 着 " 傻钱 " 跑。 9 ||| 所谓 的 " 傻钱 " **X**， 其实 就 是 买 入 并 持有 美国 股票 这样 的 普通 组合。 10 ||| 这个 策略 要 比 对冲 基金 和 其它 专业 投资者 使用 的 更为 复杂 的 投资 方法 效果 好 得 多。
Query	所谓 的 " 傻钱 " **X**， 其实 就 是 买 入 并 持有 美国 股票 这样 的 普通 组合。
Answer	策略

Daily Mail

Microsoft MCTest

HFL-RC

- HFL-RC: People Daily & Children's Fairy Tale (COLING2016), 首个中文阅读理解数据集
  - 问题类型：完形填空类
  - 答案类型：篇章中的单个词（实体、普通名词）
  - 数据标注：机器自动
  - 数据规模：约87万个问题
  - 评价方法：准确率（与正确答案是否完全一致）

Consensus Attention-based Neural Networks for Chinese Reading Comprehension

Yiming Cui[†*], Ting Liu[‡], Zhipeng Chen[†], Shijin Wang[†] and Guoping Hu[†]
[†]iFLYTEK Research, Beijing, China
[‡]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China
[†]{ymcui,zpchen,sjwang3,gphu}@iflytek.com
[‡]tliu@ir.hit.edu.cn

HFL-RC

- 中文阅读理解数据集HFL-RC上的结果

	CNN News		CBTest NE		CBTest CN	
	Valid	Test	Valid	Test	Valid	Test
Deep LSTM Reader†	55.0	57.0	-	-	-	-
Attentive Reader†	61.6	63.0	-	-	-	-
Impatient Reader†	61.8	63.8	-	-	-	-
Human (context+query)‡	-	-	-	*81.6*	-	*81.6*
LSTMs (context+query)‡	-	-	51.2	41.8	62.6	56.0
MemNN (window + self-sup.)‡	63.4	66.8	70.4	66.6	64.2	63.0
Stanford AR♮	*72.4*	*72.4*	-	-	-	-
AS Reader♯	**68.6**	69.5	73.8	68.6	**68.8**	63.4
CAS Reader (mode: avg)	68.2	**70.0**	**74.2**	**69.2**	68.2	**65.7**

	People Daily		Children's Fairy Tale	
	Valid	Test	Test-auto	Test-human
AS Reader	64.1	67.2	40.9	33.1
CAS Reader (mode: avg)	**65.2**	**68.1**	41.3	**35.0**
CAS Reader (mode: sum)	64.7	66.8	**43.0**	34.7
CAS Reader (mode: max)	63.3	65.4	38.3	32.0

国内外阅读理解研究现状

- 阅读理解数据集的变化：Stanford SQuAD

Microsoft MCTest

DeepMind CNN & Daily Mail

Facebo CBT

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

SQuAD

- Stanford SQuAD (EMNLP2016)
  - 问题类型：真实用户提问
  - 答案类型：篇章中的一个文本片段
  - 数据标注：人工标注
  - 数据规模：约10万个问题
  - 评价方法：准确率（与正确答案是否完全一致），F1（与正确答案的模糊匹配度）

SQuAD: 100,000+ Questions for Machine Comprehension of Text

Pranav Rajpurkar and **Jian Zhang** and **Konstantin Lopyrev** and **Percy Liang**
{pranavsr,zjian,klopyrev,pliang}@cs.stanford.edu
Computer Science Department
Stanford University

SQuAD

- HFL目前在该数据集上，单模型、多模型分别排名第二名

Rank	Model (Single)	EM	F1	Avg
1	r-net (*MSRA*)	74.614	82.458	78.536
2	Interactive AoA (*HIT & iFLYTEK, HFL*)	72.507	80.656	76.582
3	jNet (*USTC & NRC & York Univ.*)	70.607	79.821	75.214
4	Ruminate Reader (*New York Univ.*)	70.639	79.456	75.048
5	Document Reader (*Facebook*)	70.733	79.353	75.043
6	ReasoNet (*MSR Redmond*)	70.555	79.364	74.960
7	FastQAExt (*GRC for AI*)	70.849	78.857	74.853
8	RaSoR (*Google NY, Tel-Aviv Univ.*)	70.849	78.741	74.795
9	Mnemonic Reader (*NUDT & Fudan*)	69.863	79.207	74.535
10	Multi-Perspective Matching (*IBM*)	68.877	77.771	73.324
11	T-gating (*Peking Univ.*)	68.132	77.569	72.851
12	BiDAF (*AI2 & Washington Univ.*)	67.974	77.323	72.649

综合对比

数据集	问题类型	答案类型	规模	数据标注	评价方法	优点	缺点
MCTest	真实提问	四选一	约500篇章	人工标注	准确率	真实问题	数量少
CNN/DailyMail	完形填空	篇章中某个词	380K/879K	机器生成	准确率	数据容易生成	非真实问题
CBTest	完形填空	篇章中某个词	700K	机器生成	准确率	数据容易生成	非真实问题
HFL-RC	完形填空	篇章中某个词	870K	机器+人工	准确率	数据容易生成	非真实问题
SQuAD	真实提问	篇章片段	100K	人工标注	准确率 F1	真实问题	答案须在篇章

首届中文阅读理解评测

- 由CIPS主办，HFL承办的CMRC2017阅读理解评测启动
- 已有超过30多家高校、企业报名参加本次评测
- 数据来源：儿童读物
- 评测方向：填空型问题、用户提问型问题

Winograd评测

- 以人工智能研究先驱 Terry Winograd 的名字命名
- 「Winograd 模式」是一种简单，但措辞含糊的自然语言问题。要想正确解答这个问题需要被测试者具备足够的「常识」，理解在现实世界当中人工智能、事物和文化规范是如何互相影响的。
- 举例：“市议员拒绝给示威者许可，因为他们提倡暴力。”
- 在 2016 年，有 4 个系统比赛回答一套 60 个 Winograd 模式问题的测试。胜出系统的准确率只达到了 **58%**

提纲

- 概述
- 句法语义分析评测
- 阅读理解评测
- **人机对话评测**
- 总结

人机对话的四大功能

	聊天	问答	指令	推荐
目的	闲聊	知识获取	完成任务/动作	信息推荐
领域	开放域	开放域	特定域	特定域
应用	娱乐、情感陪护	客服、教育	虚拟个人助理	个性化推荐
系统	小冰、度秘	Watson、Wolfram Alpha	Siri、Cortana、Allo、灵犀	Quartz
评价	主观+客观	客观	客观	客观

上苍为你关了一扇门，就会为你再开一扇窗，你会寻找到自己真正的幸福的！

我很难过

我们是朋友，有事不要憋在心里。

我很难受

不要放弃治疗！

我晕。。。。

那可不行，你要是晕了，我都说不清啦。

我生气了

张伟男呀，别难过啦~只要你的脚还在地面上，就别把自己看得太轻；只要你还生活在地球上，就别把自己看得太大。

最近感觉奶水不够

我也是母乳不足，奶水不够就吃奶粉，我也不知道怎么办，我也是母乳不足，奶水不够，

美赞臣奶粉怎么样

20:42

我的宝宝也是喝的这个奶粉，我觉得还不错，宝宝喝了不上火，大便也很好，而且宝宝也很健康，

孩子吃奶粉上火啊

给您推荐：清火宝 特福芬 豆汤 七星茶 果汁 蜂蜜水 凉茶 苹果 益生菌 金银花

好的 谢谢你

感激我的话就放心里吧，记得我的好就行。

聊天机器人“笨笨”

➢哈工大社会计算与信息检索研究中心研发

➢系统形式

- 微信服务号：2016年6月6日上线

➢主要功能

- 聊天（Chit-Chat）
- 指令执行（Language Grounding）
- 知识问答（QA）
- 新闻推荐（Recommendation）

首届中文人机对话技术评测—SMP-ECDT

- 评测委员会
  - **主席：张伟男（哈尔滨工业大学）**
  - 委员：陈志刚（科大讯飞股份有限公司）、车万翔（哈尔滨工业大学）、芮祥麟（华为公司）
  - 数据支持：科大讯飞股份有限公司
  - 奖金支持：华为公司
- 评测时间
  - 2017.4.24-2017.5.16：正式报名
  - 2017.5.17：发布训练及开发集
  - 2017.8.9：任务1评测
  - 2017.8.10-2017.8.12：任务2评测

任务1：用户意图领域分类

- 举例
  - 1） 你好啊，很高兴见到你！ （闲聊类）
  - 2） 我想订一张去北京的机票。 （任务型垂类：订机票）
  - 3） 我想找一家五道口附近便宜干净的快捷酒店（任务型垂类：订酒店）
- 闲聊和垂类两大类，其中垂类又细分为30个垂直领域
- 评价指标：准确率（p）、召回率（r）和F值

任务2：特定域任务型人机对话在线评测

请问您需要我查询出

我能查飞机票、火车
您要看哪个呢?

您要从哪里出发呢?

机票信息
5月17日
帮您查到了5月18日从哈尔滨
息查看

阅读全文

要去北京参加C

请问您需要我查询出行

20:10

我能查飞机票、火车票
您要看哪个呢?

您要从哪里出发呢?

您想要哪一天出发呢?

火车票信息
5月18日

到北京得订个酒店啊

您想要哪天入住呢?

酒店信息
5月18日

以下是我找到的5月19日北京地区的
幽谷神谭春立农家院　怀柔区怀北镇
村28号
标准间（公共卫生间）　[双床]
家庭间　[双床]　含早餐
五人间　[其他]　含早餐
8人间　[其他]　含早餐
套房　[其他]　含早餐

水长城西水湾欢乐度假村　怀柔区九
西水湾欢乐度假村
标准间A [双床]　含早餐
标准间B [双床]　含早餐
三人间A [三人间] 含早餐
三人间B [三人间] 含早餐
水长城游玩+特色套餐+床位房 [其他
水长城游玩+特色套餐+真人cs+床位
水长城游玩+篝火+烧烤+特色套餐+
早餐

挂甲峪老付农家院　平谷区挂甲峪
大床房　[大床]　含早餐

从五道口到国家会议中心坐什么公交车?

您在哪个城市呢?

北京啊

公交线路信息
5月18日

以下是我找到北京市从五道口到国家会议中心的公交线路信息:

步行1112米,从清华东路西口站(C口),乘地铁15号线(俸伯方向),经过3站,到奥林匹克公园站(H口)
步行171米,至终点 ¥1

从五道口站,乘86路(国家体育馆方向),经过7站,到国家体育馆站　¥1　详情:
86路　耗时:16分11秒　首班车:05:30　末班车:22:30

从展春园西路站,乘311路(来广营西桥东方向),经过8站,到洼里南口站　¥1　详情:
311路　耗时:19分37秒　首班车:06:30　末班车:23:00

从五道口站,乘地铁13号线(西直门方向),经过1站,到知春路站
站内换乘 步行328米,从知春路站,乘地铁10号线(西土城方向),经过4站,到北土城站

评测的难点：多轮分叉

- **用户完整意图：** 预定下周三（2017年4月12日）从北京去上海的上午或者中午出发，价格低于8折的便宜机票。

U：下周从北京去上海（相同的首轮对话输入）

R1：请问您需要订票吗？　　R2：请问您坐火车还是飞机？

U：恩，是的　　U：坐飞机吧

多轮分叉，导致无法事先给出标准答案
本来就没有标准路径，能够完成任务的就是正确的

人工评测方法

- 数据
  - 1） 完整的用户意图描述示例
  - 2） 静态航班、酒店、火车票数据资源
- 评测过程
  - 对不同的参赛系统均给定相同的首轮对话输入，启动在线评测，评测过程中评测员根据给定的完整意图描述与参赛系统进行交互，直至对话结束
- 对话结束的条件
  - 1）系统返回所有任务（单任务或多任务）的执行结果
  - 2）超过50个对话轮数之后仍未返回测试员所需的结果，测试员手动结束对话测试

人机对话评价指标

- 主观
  - 任务完成率
  - 用户满意度
  - 回复语言的自然度
  - 静态的数据库资源未覆盖情况的引导能力
- 客观
  - 对话轮数（自动统计），话轮越短越好

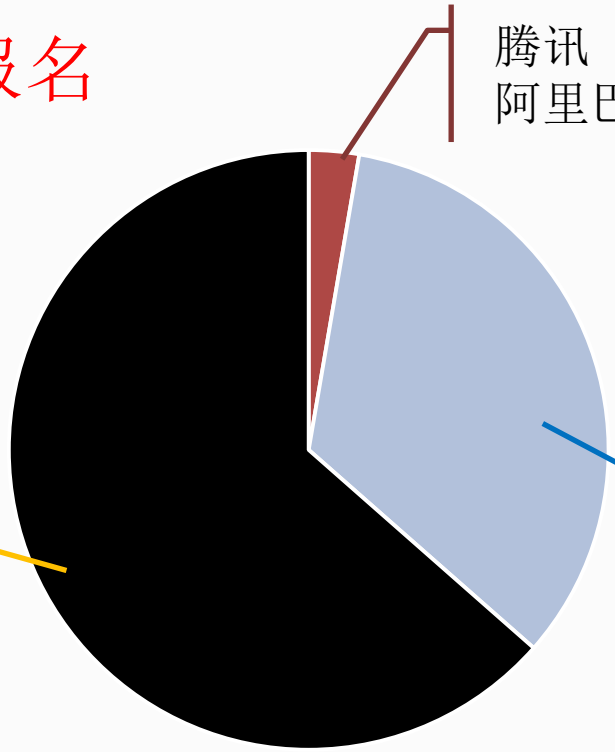
人机对话示例

- 用户完整意图：预定下周三（2017年4月12日）从北京去上海的上午或者中午出发，价格低于8折的便宜机票。
- 【示例】

U：下周从北京去上海
R：请问您需要订票吗？
U：恩，是的
R：请问您要下周哪天出发呢？
U：周三吧
R：帮您预订2017年04月12日从北京到上海的机票，请问有餐饮，价格等其他的要求吗？
U：尽量订上午的吧，价格找便宜一点的
R：帮您查询了2017年04月12日从北京到上海的机票，没有上午出发的航班，中午12：10出发是否可以？7.3折，910元
U：也行，就订这个吧
R：即将帮您预订2017年04月12日从北京到上海的机票，南方航空，CZ6412 空客321（中），12：10：00出发，7.3折，910元，不可退改签，确定要订票吗？
U：是的
R：已经帮您预订该航班机票，将跳转至付款页面

报名情况统计

共74支报名队伍



腾讯
阿里巴巴

复旦大学
中国科学院大学
香港理工大学
香港中文大学
武汉大学
北京邮电大学
山西大学
大连理工大学
…

深圳狗尾草智能科技有限公司
武汉飓拓科技有限公司
义语智能科技（上海）有限公司
智言科技(深圳)有限公司
…

图中所列为部分代表队伍所在的机构，且排名不分先后

NIPS 2017 人机对话技术评测

- 组织者
  - Mikhail Burtsev, Valentin Malykh, 莫斯科物理技术学院（ MIPT ）
  - Ryan Lowe, 麦吉尔大学（McGill University）
  - Iulian Serban, Yoshua Bengio, 蒙特利尔大学（University of Montreal）
  - Alexander Rudnicky, Alan W. Black, 卡内基梅隆大学（CMU）
- 评测任务
  - 给定特定文章（新闻或维基百科文章）的人机对话评测
  - 评测方式为人工评测

NIPS 2017 人机对话技术评测

- 评价指标（人工打分）
  - 单轮回复的质量：1-10分
  - 整个对话片段的质量：0-10分
  - 最终分别取平均分
- 评测时间
  - 2017.6.18：资格认证
  - 2017.7.24-2017.7.30：发布数据
  - 2017.11.20-2017.12.3：赛前人类评测（产生的对话用于调试系统）
  - 2017.12.4-2017.12.9：正式评测

提纲

- 概述
- 句法语义分析评测
- 阅读理解评测
- 人机对话评测
- **总结**

总结

- 句法语义分析是人为定义的问题
- 问题的定义形式多样、多变，总体上从在从句法向语义发展
- 评测数据的规模很有限，每一种语言有1-2万句子作为评测数据，小语种更少
- 评测方法简单

小结

- 阅读理解
  - 近期阅读理解成为国际研究热点，反应出NLP的研究向篇章级理解、推理发展的趋势
  - 评测的难点在于：如何设计题目才能有效地迫使机器采用更深入的推理方法，而不是简单的搜索匹配。
- 人机对话
  - 人的计划的评测非常有挑战性，如何自动地评测人机对话技术，本身就是一个值得深入研究的课题

总结

- 技术评测在不断地推动自然语言处理的技术的发展
- 新的问题和新的评测方式不断提出
- 难度略高于当前技术水平的评测，对技术发展最有利
- 评测不能只给出一个指标，更应该帮助诊断当前技术的缺陷
- 应该在原理创新的基础上参与评测，而不是一味追求更高的排名

致谢

- 哈工大社会计算与信息检索研究中心（HIT-SCIR）
  - **车万翔**副教授（句法语义分析）
  - **张伟男**博士（人机对话）
- 哈工大讯飞联合实验室（HFL）
  - **崔一鸣**研究员（阅读理解）

谢　谢！