



知识与自然语言处理

吴 华

(百度自然语言处理部首席科学家)

2019年5月4日

自然语言处理

自然语言处理研究实现人与计算机之间用自然语言进行有效通信的各种理论和方法。



电影《Her》：计算机拥有类似人类的语言能力

自然语言处理

自然语言处理研究实现人与计算机之间用自然语言进行有效通信的各种理论和方法。



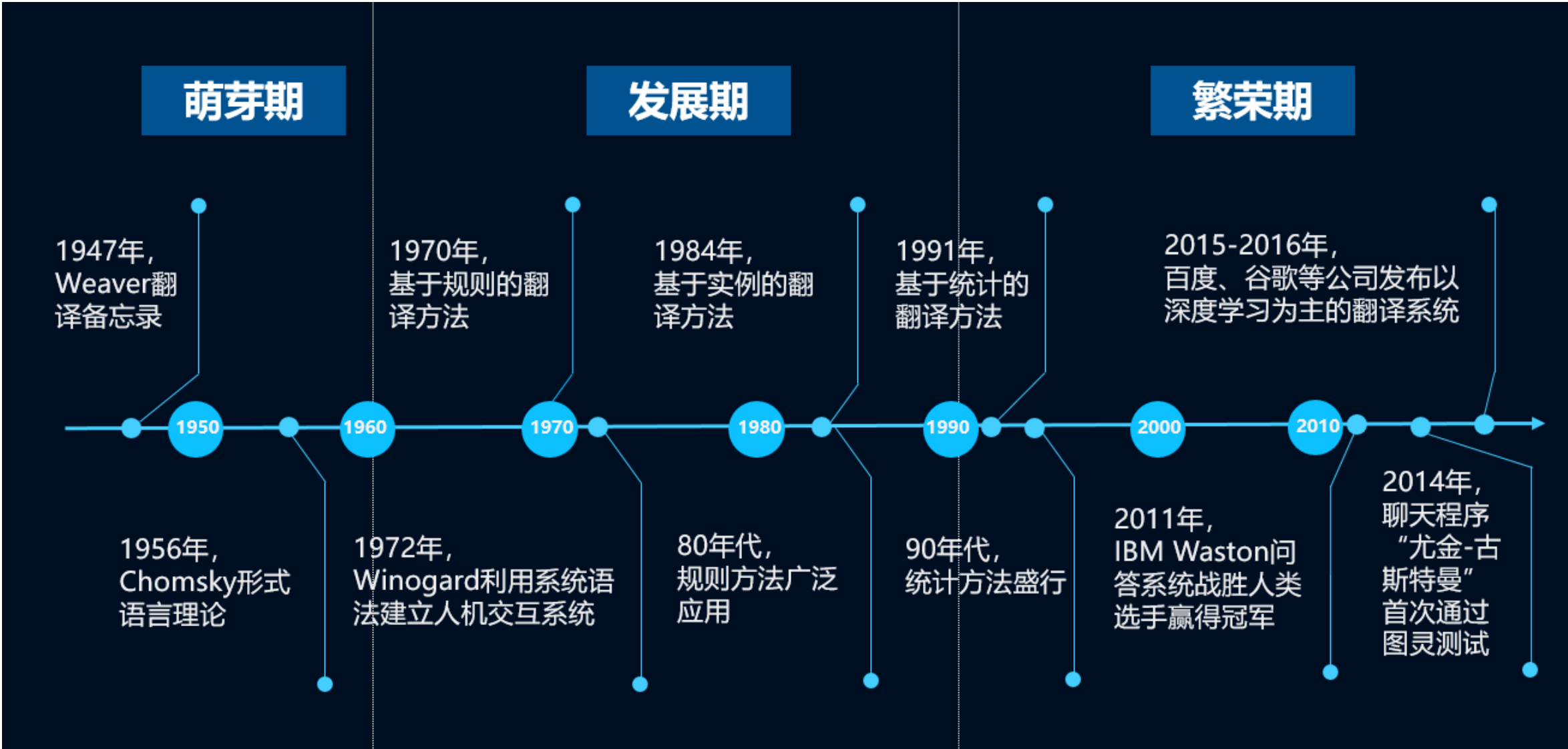
人类梦想：自由的跨语言沟通

自然语言处理目标

使计算机拥有人类的语言能力

实现跨语言沟通

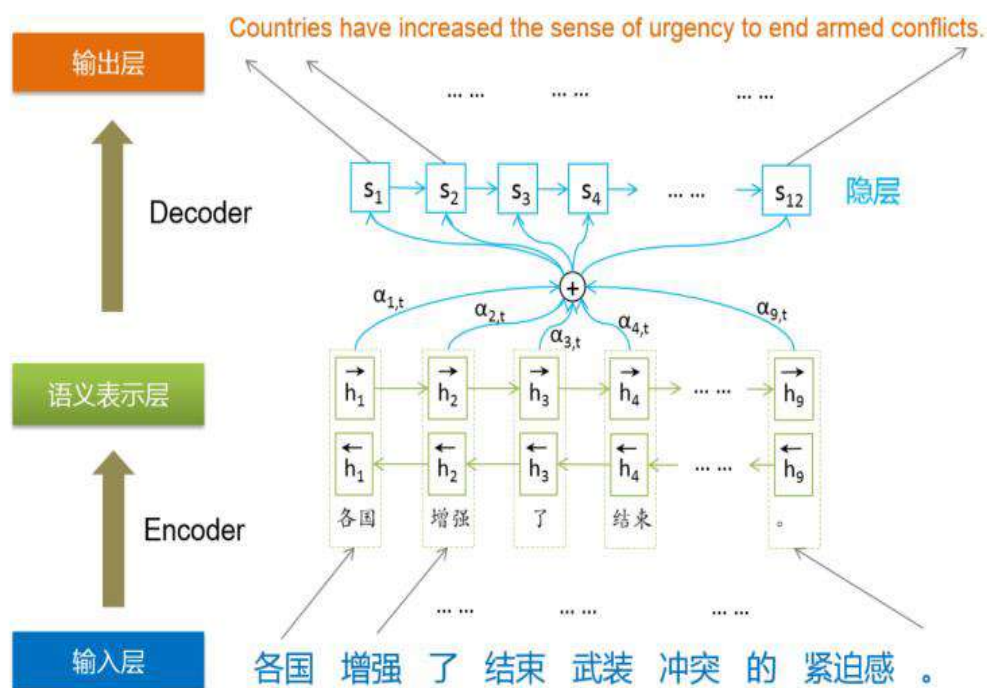
自然语言发展历程



深度学习+大数据促进了NLP的发展

机器翻译的水平达到CET6的水平

在SQuAD测试集上，机器的阅读理解准确率基本与人类持平



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, **grau-pel** 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?

grau-pel

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

within a cloud

挑战：没有理解能力的自然语言处理技术

常见句子翻译不对



聊天没有逻辑



加上“对抗句子”，准确率 **75%->36%**

Article: Super Bowl 50

Paragraph: “Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.”

Question: “What is the name of the quarterback who was 38 in Super Bowl XXXIII?”

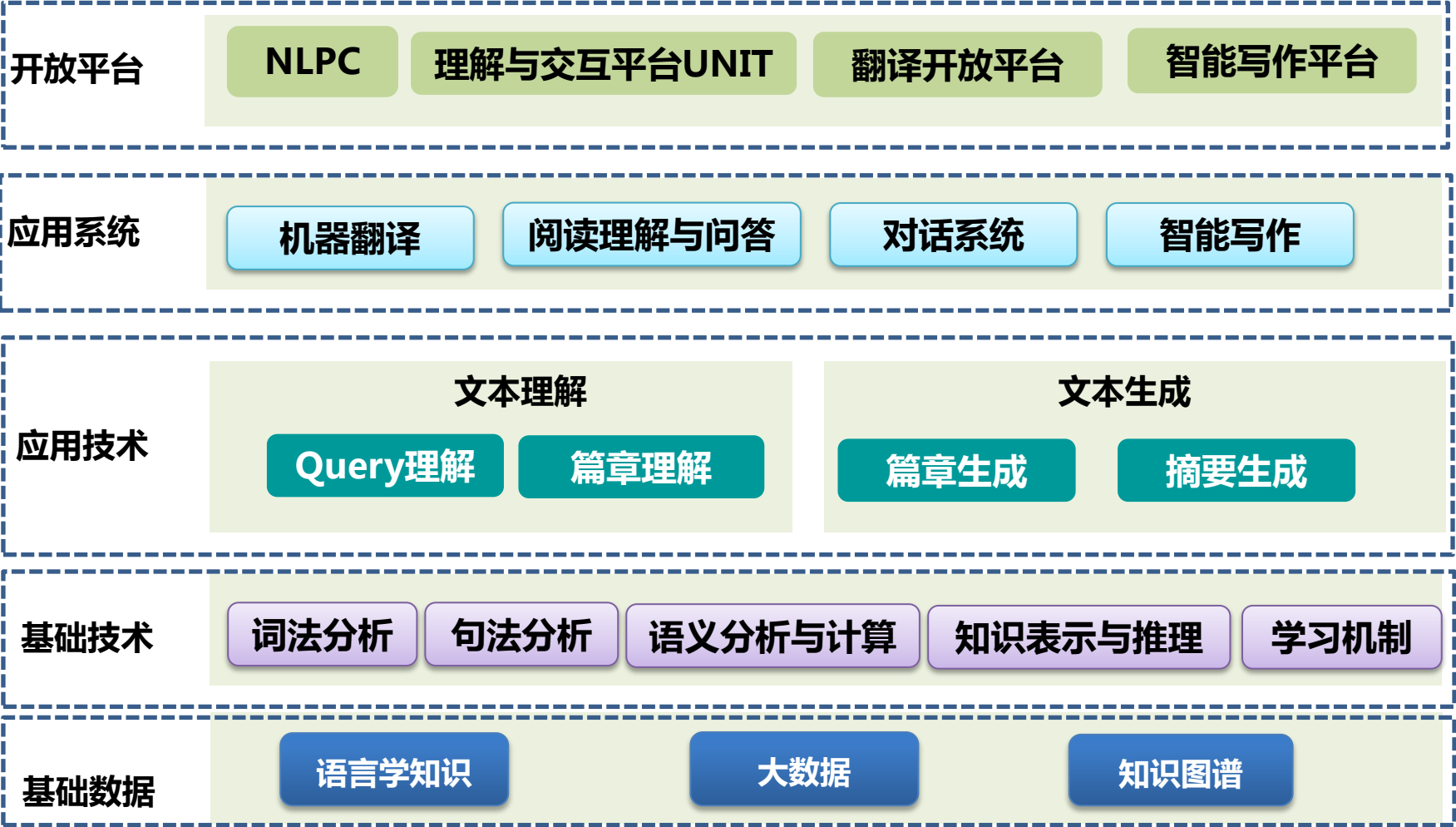
Original Prediction: John Elway

Prediction under adversary: Jeff Dean

NLP发展方向



百度语言与知识技术



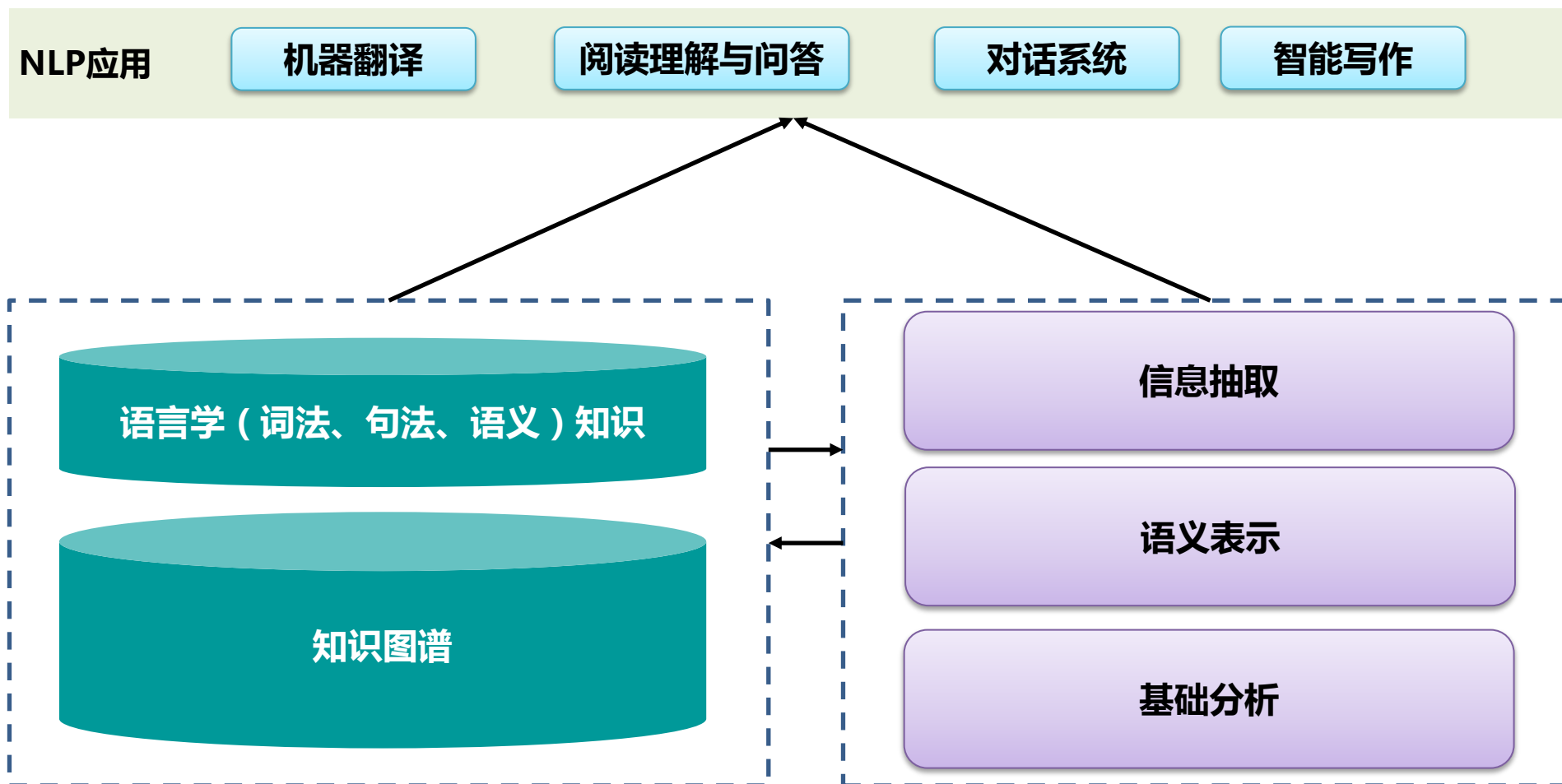
什么是知识

Knowledge is a familiarity, awareness, or understanding of someone or something, such as facts, information, descriptions, or skills, which is acquired through experience or education by perceiving, discovering, or learning.

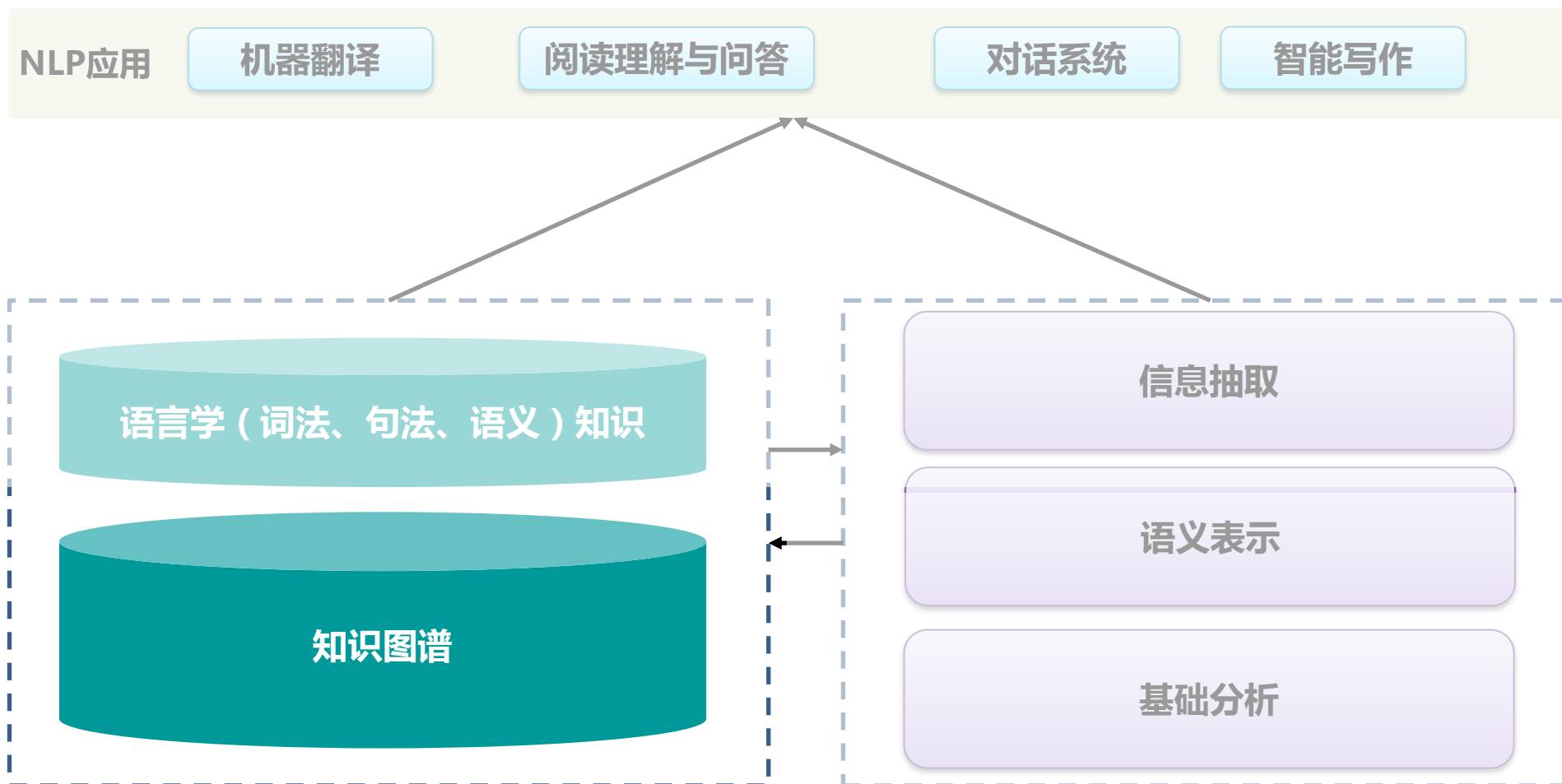
自然语言处理中的知识：

- 知识图谱
 - 实体图谱
 - 事件图谱
 - 关注点图谱
 - POI图谱
 -
- 语言学知识
 - 词法信息（分词、词性、专名）
 - 句法信息
 - 语义信息
 -

Outline



Outline



多元异构知识图谱

关注点图谱

- 对用户的关注点及其之间关联关系进行建模
- 节点：关注点（实体/事件/话题等）
- 边：关系（包含/关联/推演等）
- 百万级节点、千万级边

用户图谱

- 与10亿级用户画像、Passport等进行链接

实体图谱

- 对事物或抽象概念及其相互间关系进行建模
- 节点：基本事物 / 抽象概念
- 属性：各种客观存在的基础属性
- 边：关系（关联/包含/并列/.....）
- 十亿级实体、千亿级事实

事件图谱

- 对客观世界的动态变化进行建模
- 节点：事件 / 事件集合
- 边：关系（因果、从属、时序等）
- 百万级节点、千万级边

POI图谱

- 对客观世界的地点（经纬度）进行建模
- 节点：POI
- 边：关系（临近、包含、方位等）

大数据

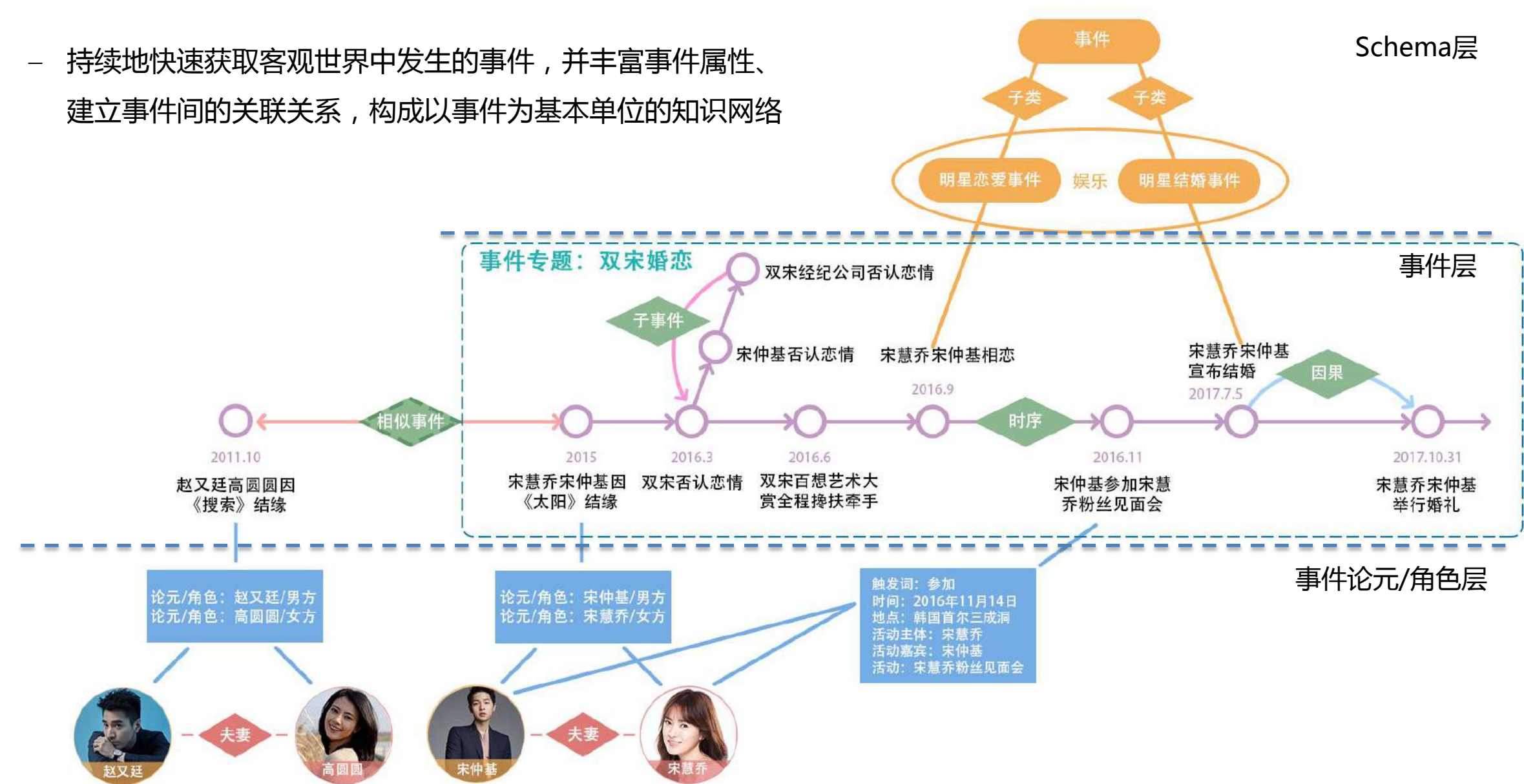
互联网数据

行业数据

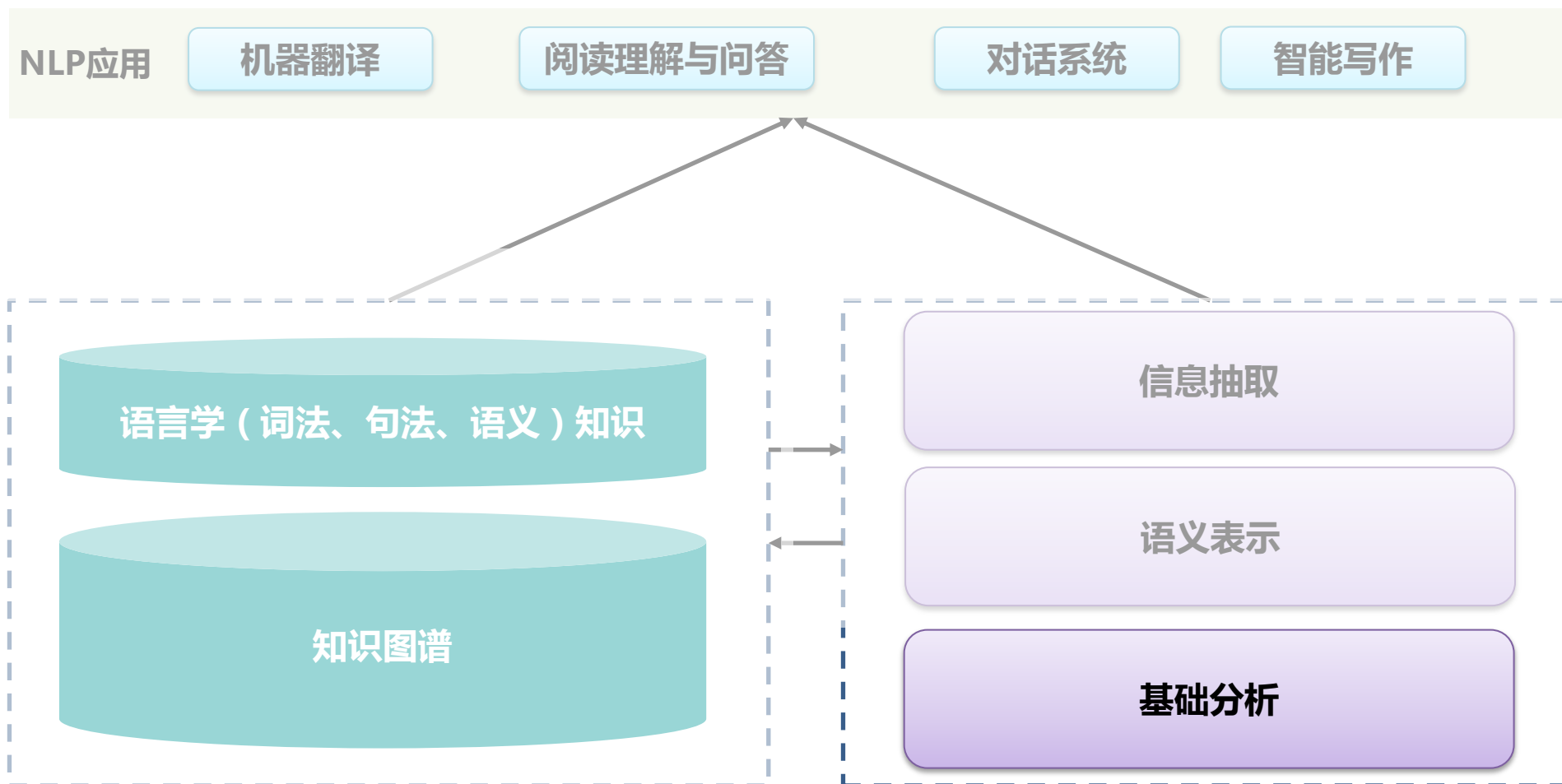
用户数据

事件图谱

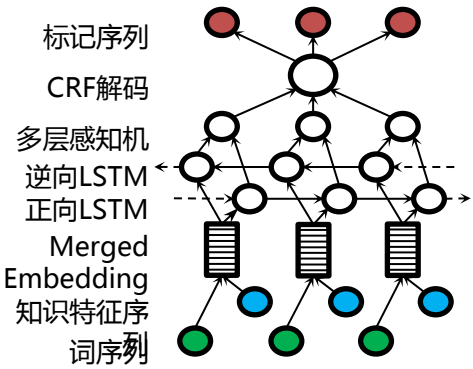
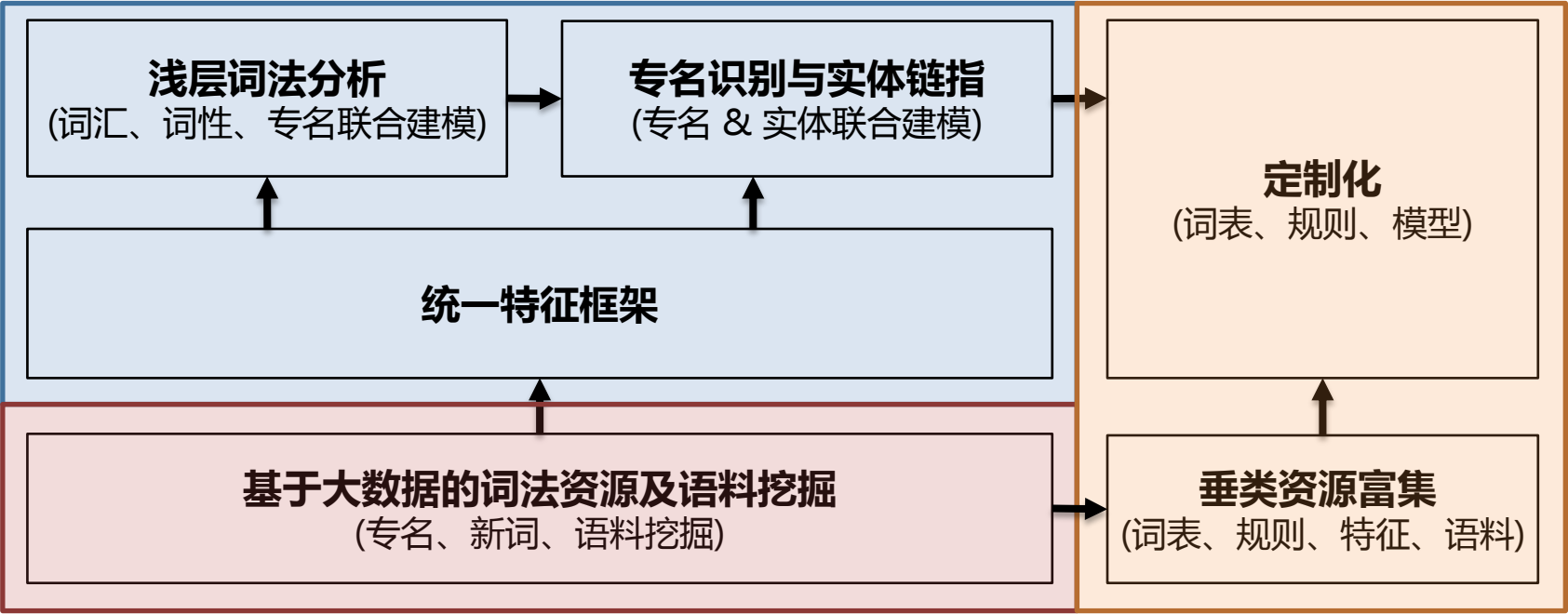
- 持续地快速获取客观世界中发生的事件，并丰富事件属性、建立事件间的关联关系，构成以事件为基本单位的知识网络



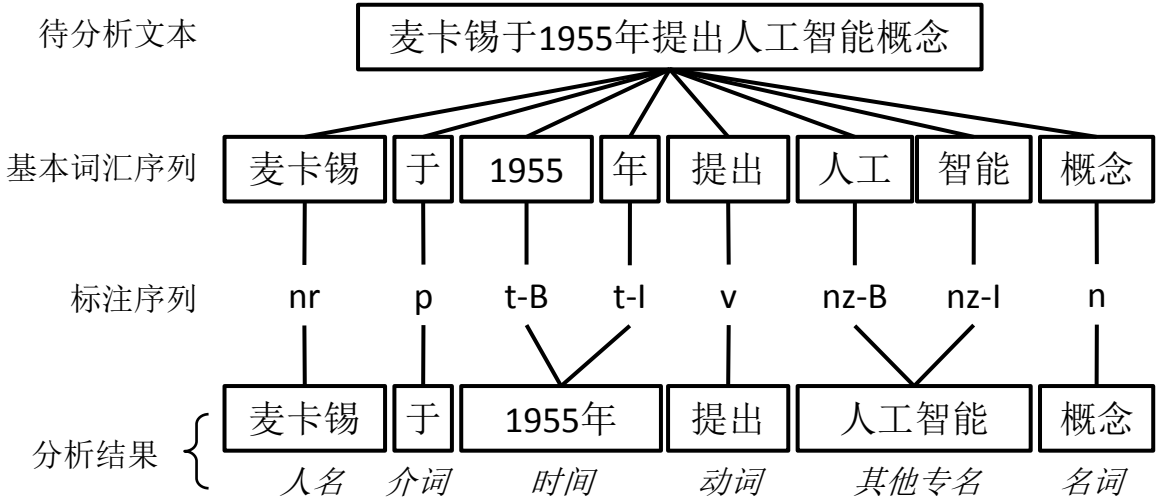
Outline



语言学知识构建：词法分析



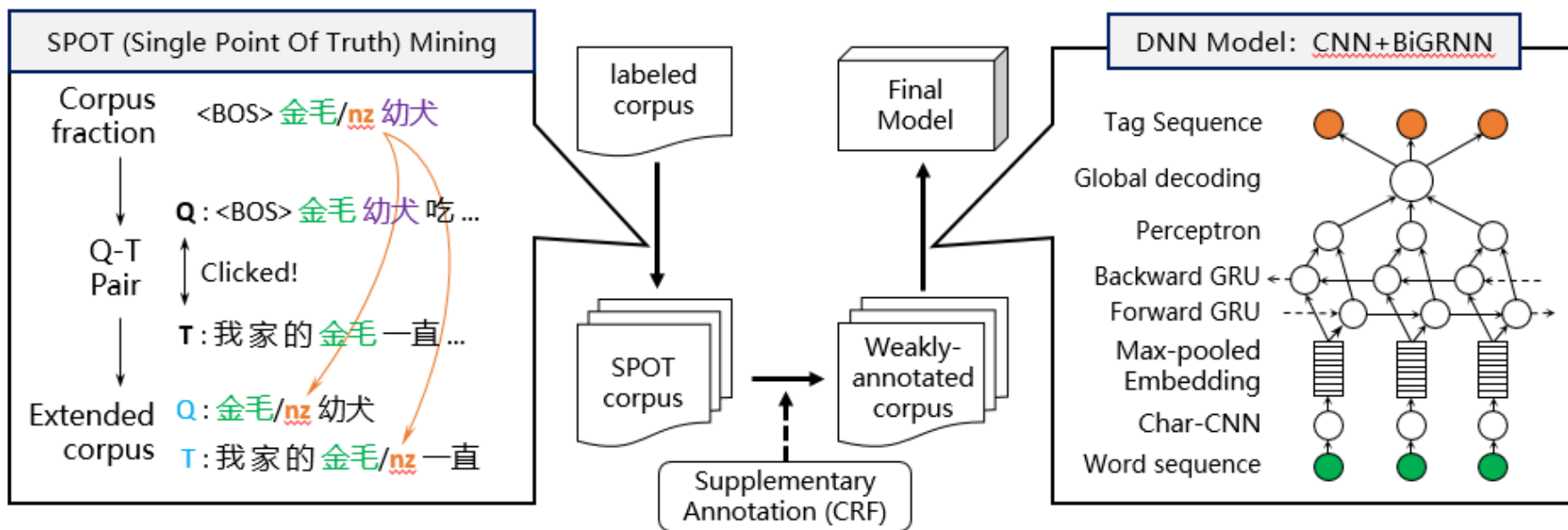
- 新词词表：百万量级
- 先验词性表：万量级
- 专名库：千万量级



词性	含义	词性	含义	词性	含义	词性	含义
n	普通名词	f	方位名词	s	处所名词	t	时间名词
nr	人名	ns	地名	nt	机构团体名	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词
a	形容词	ad	副形词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号

LAC: 联合分词、词性和专名识别

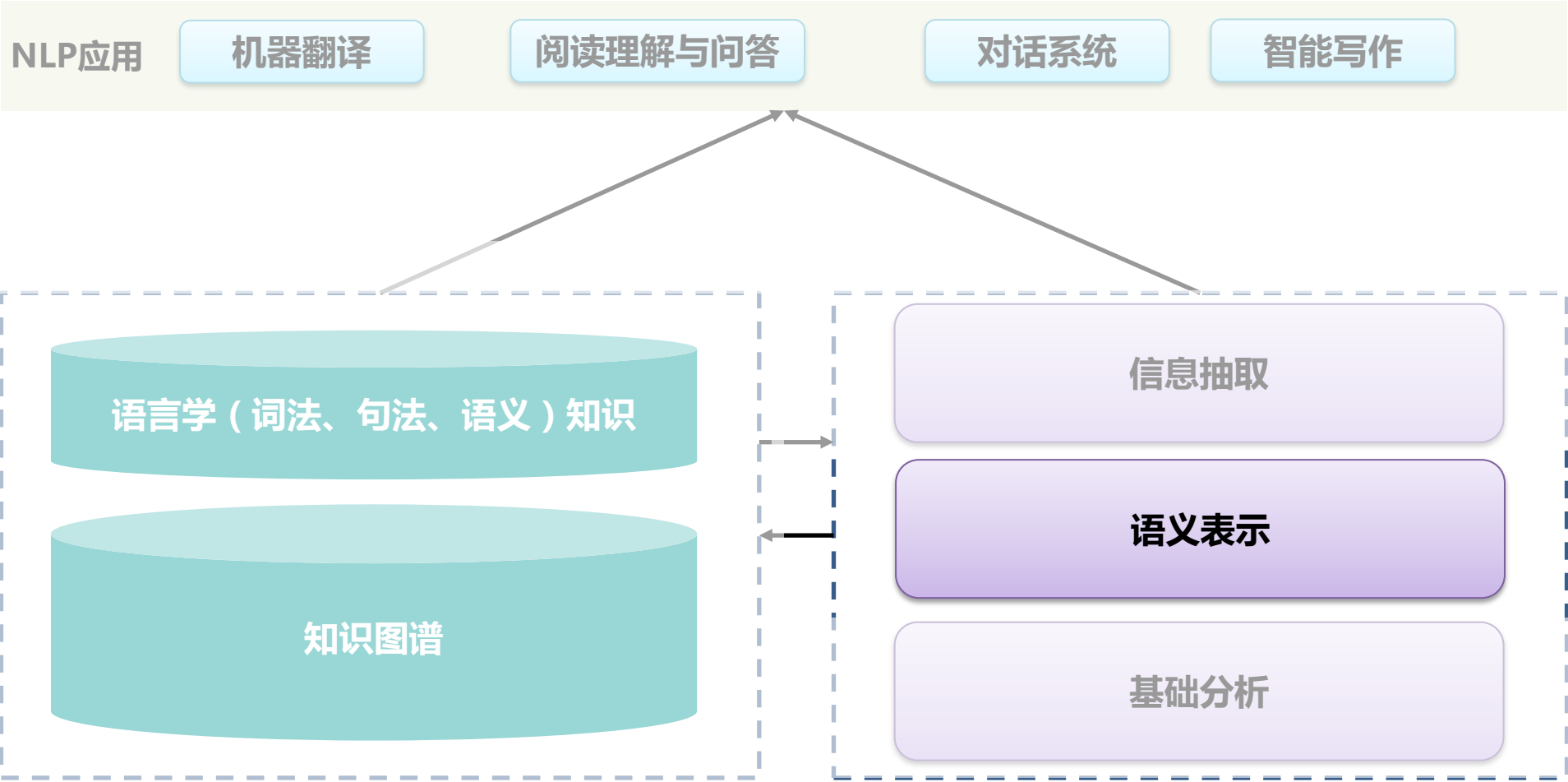
用弱标记数据（**click feedback**）扩大训练语料



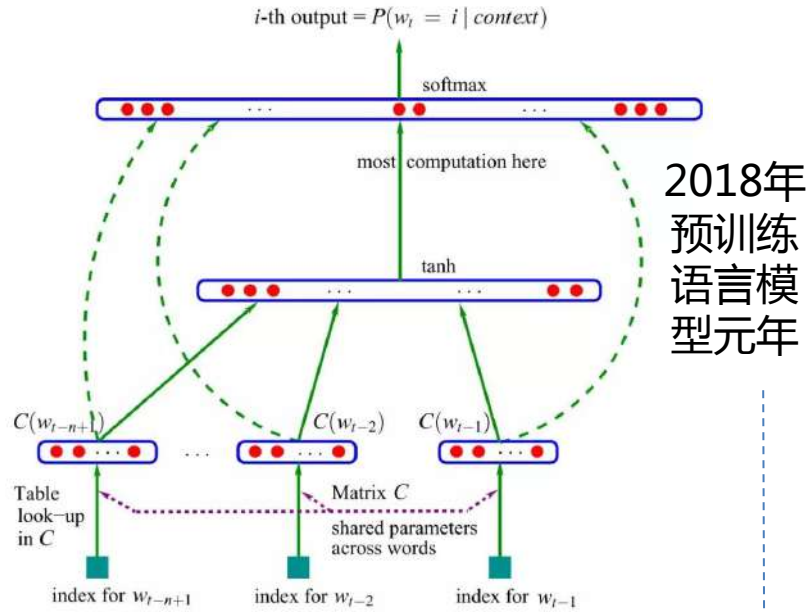
Corpus size: 200 thousand → 130 million sentences
Precision of the labeled data: >95%
Benefits: Error (segmentation + tagging) rate -1.6%
Long words recognition P +6.3%, R +4.7%

https://github.com/PaddlePaddle/models/tree/develop/PaddleNLP/lexical_analysis

Outline

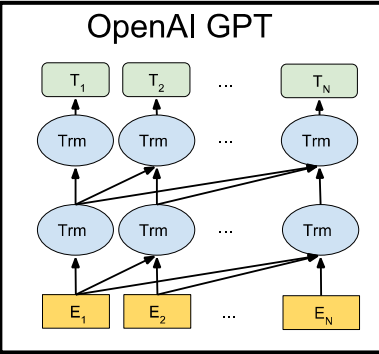
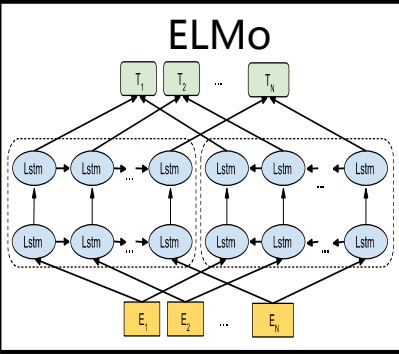


语义表示演进：基于大规模无标记语料的语言模型

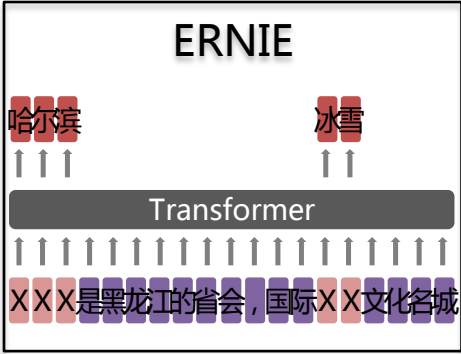
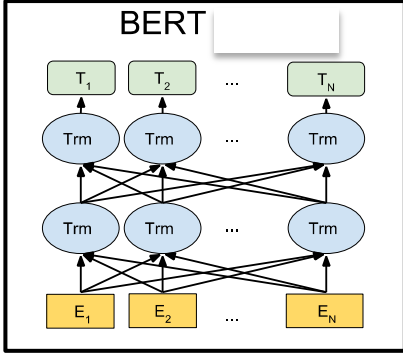


2018年
预训练
语言模
型元年

通过LR和/或RL语言模型任务训练



通过Masked语言模型任务训练



2001年，
Bengio等提出
前馈神经网络
语言模型

2013年，大规模词
嵌入训练工具出现

上下文相关的词
向量模型

深层单向
Transformer

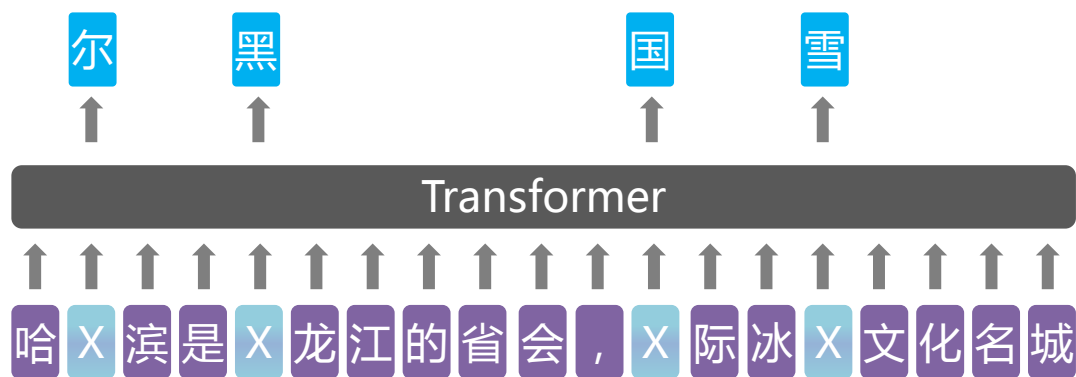
深层双向
Transformer

基于知识增强的
语义建模

ERNIE: Enhanced Representation through kNowledge IntEgration

- BERT：基于字单元的语义建模

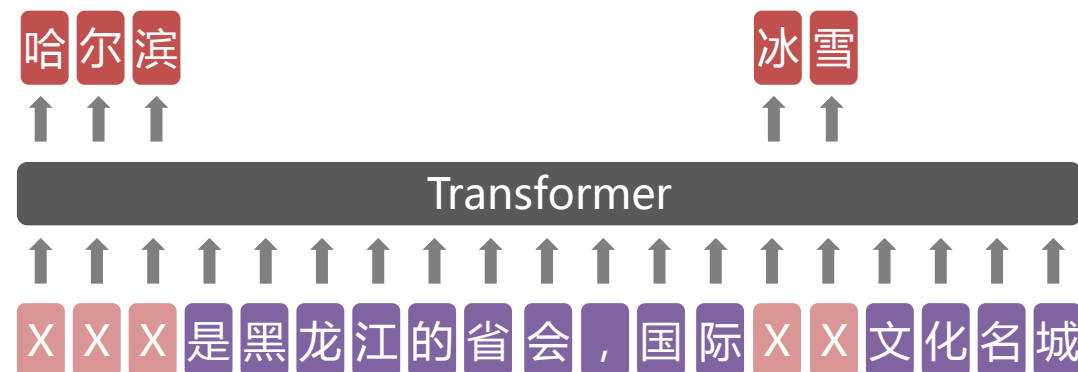
- 基于字单元的双向Masked语言模型
- 缺乏明确的语义概念及对应的语义关系



哈 _ 滨 是 _ 龙 江 的 省 会 ， _ 际 冰 _ 文 化 名 城

- ERNIE：融合知识的语义建模

- 基于词、实体概念等先验语义知识单元建模
- 通过海量文本数据学习真实世界的语义关系



_ _ _ 是 黑 龙 江 的 省 会 ， 国 际 _ _ 文 化 名 城

ERNIE的能力

- ✓ 海量训练语料：基于百科、新闻、贴吧多源数据训练，80亿字
- ✓ 设置和BERT-base一样：L=12, H=768, A=12

ERNIE中文效果全面领先BERT

<https://github.com/PaddlePaddle/LARK/tree/develop/ERNIE>

NLP-TASK(中文)	ERNIE	BERT
命名实体识别（MSRA-NER）	93.8%(+1.2)	92.6%
自然语言推理（XNLI）	78.4%(+1.2)	77.2%
自动问答（NLPCC-DBQA）	82.7%(+1.9)	80.8%
情感分析（ChnSentiCorp）	95.4%(+1.1)	94.3%
语义相似度计算（LCQMC）	87.4%(+0.4)	87.0%

ERNIE知识推理能力

文本	ERNIE	BERT	答案
2006年9月，__与张柏芝结婚，两人婚后育有两儿子——大儿子Lucas谢振轩，小儿子Quintus谢振南；2012年5月，二人离婚。	谢霆锋	谢振轩	谢霆锋
戊戌变法，又称百日维新，是__、梁启超等维新派人士通过光绪帝进行的一场资产阶级改良。	康有为	孙世昌	康有为
__是中国神魔小说的经典之作，达到了古代长篇浪漫主义小说的巅峰，与《三国演义》《水浒传》《红楼梦》并称为中国古典四大名著。	西游记	《小》	西游记
相对论是关于时空和引力的理论，主要由__创立。	爱因斯坦	卡尔斯所	爱因斯坦

语义表示在实际产品中的应用

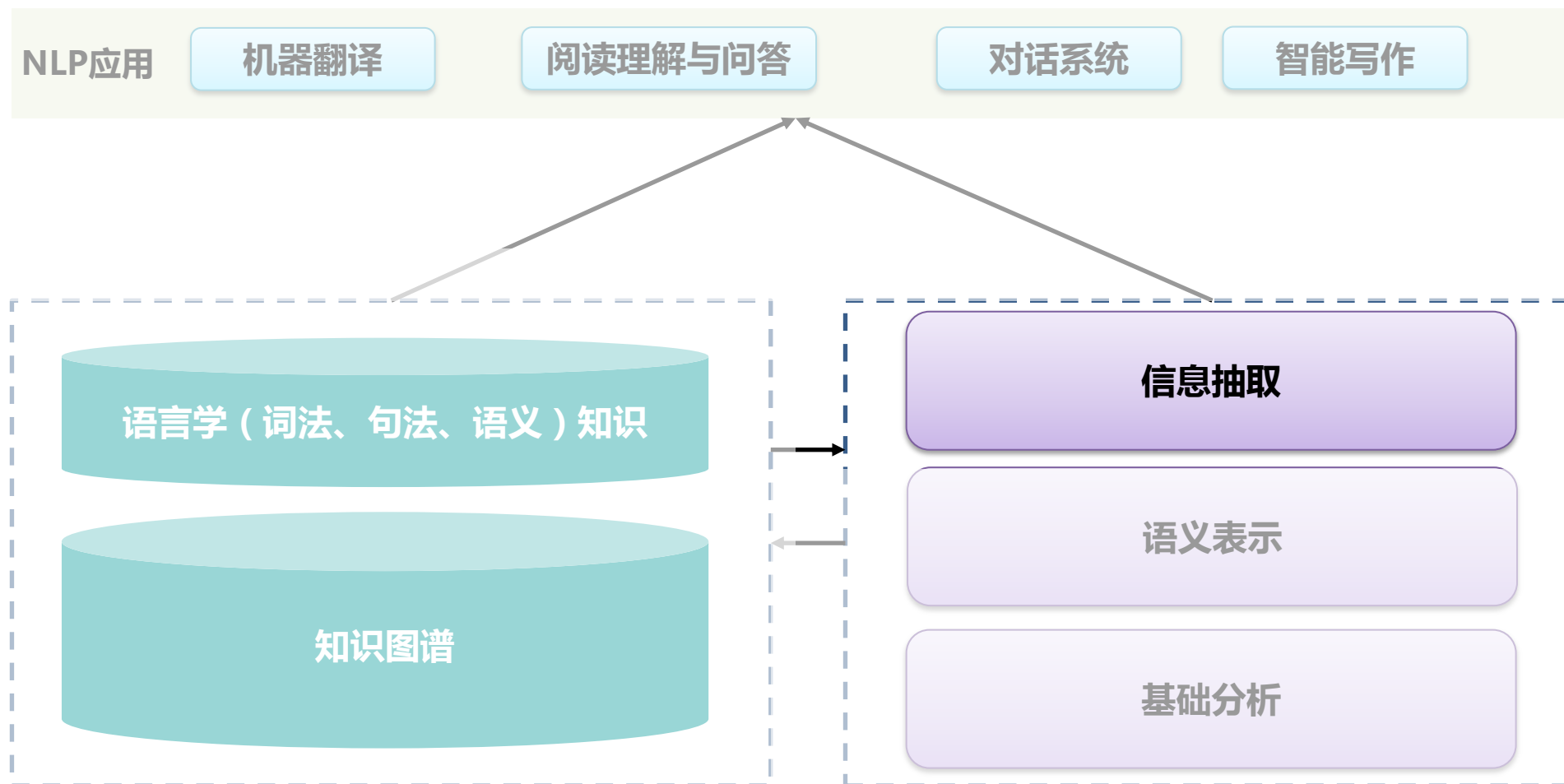
- ERNIE在广告、智能助手、FEED、NLP基础任务效果

项目	评估指标	业务基线	ERNIE	BERT
ERNIE 在 Feed 去重的应用	AUC	74.1%	84.2% (+10.1)	82.9% (+8.8)
ERNIE 在凤巢相关性的应用	AUC	88.0%	92.3% (+4.3)	91.4% (+3.4)
ERNIE 在度秘意图理解的应用	f1-score	79.1%	89.2% (+10.1)	87.9% (+8.8)
ERNIE 在 NLP 的应用	对话情绪识别	ACC	94.4% (+3.3)	93.6% (+2.5)
	LAC 词法分析	f1-score	90.3% (+2.3)	89.3% (+1.3)
	评论情感分类	ACC	96.0% (+1.6)	95.6% (+1.2)

- 上线：广告相关性，ERNIE Online化与迁移学习项目

- ✓ 无线网页：CPM1 +0.77% ,
- ✓ 落地页维度0分改善2.76%（PA评估），0+1分改善 2.00%（PA评估）

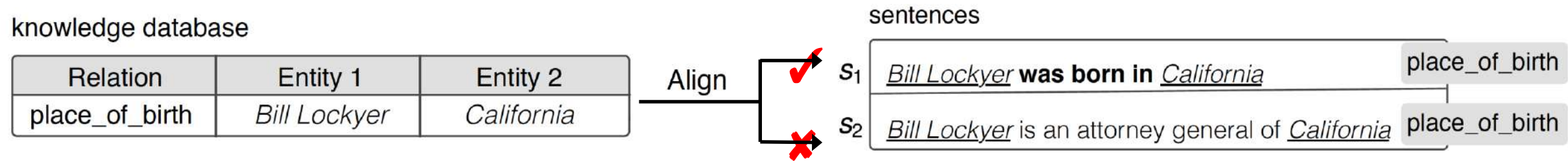
Outline



信息抽取：利用知识进行远监督学习和降噪

Motivation :

- 远监督学习通过对齐知识库和未标注语料自动获取大量的训练语料
- 远监督仅要求实体对齐，因而经常引入大量未实际表达关系的**噪声**样例



Our Assumption:

- 可靠的远监督标注数据应该是模型可解释的**，即基于此数据训练出的模型能指出表达该关系的指示词
- 我们通过Attention Regularization强制要求模型关注关系指示词，进而更好的选择正确的标注语料

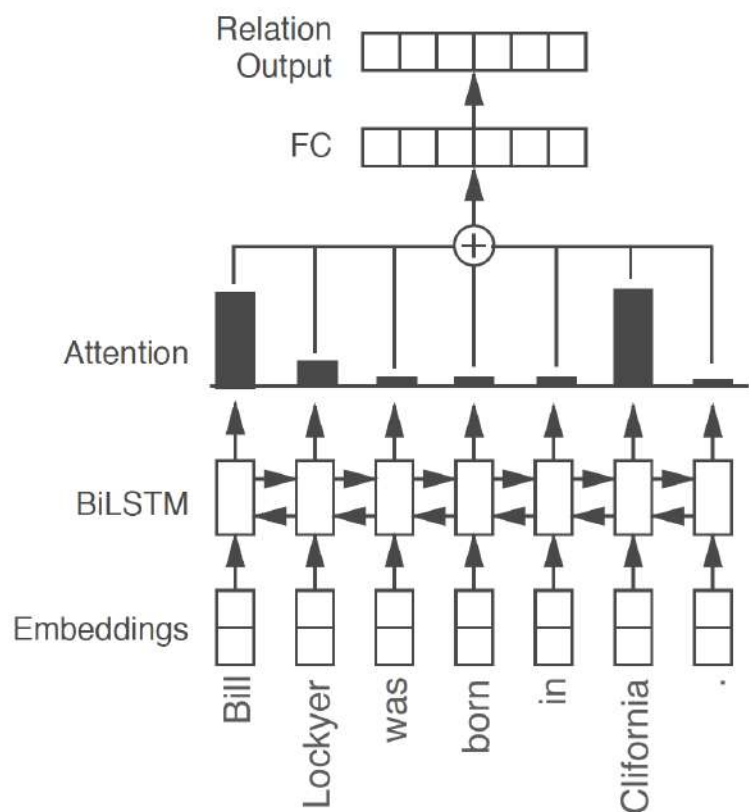
	Entity 1: AOL Entity 2: Jim Kimsey Relation: /business/company/founders	Entity 1: Kent Snyder Entity 2: Senomyx Relation: /business/person/company
传统远监督	Jim Kimsey , a founder of AOL ; Jack Valenti , former head ... 0.36 0.19	... said Senomyx 's chief executive , Kent Snyder . 0.30 0.03 0.31
我们的方法	Jim Kimsey , a founder of AOL ; Jack Valenti , former head ... 0.12 0.13 0.14 0.12	... said Senomyx 's chief executive , Kent Snyder . 0.15 0.15 0.13 0.13 0.11

不同训练方法下Attention分布对比

远监督降噪：基于Attention Regularization的迭代训练

- 模型：BiLSTM + Attention

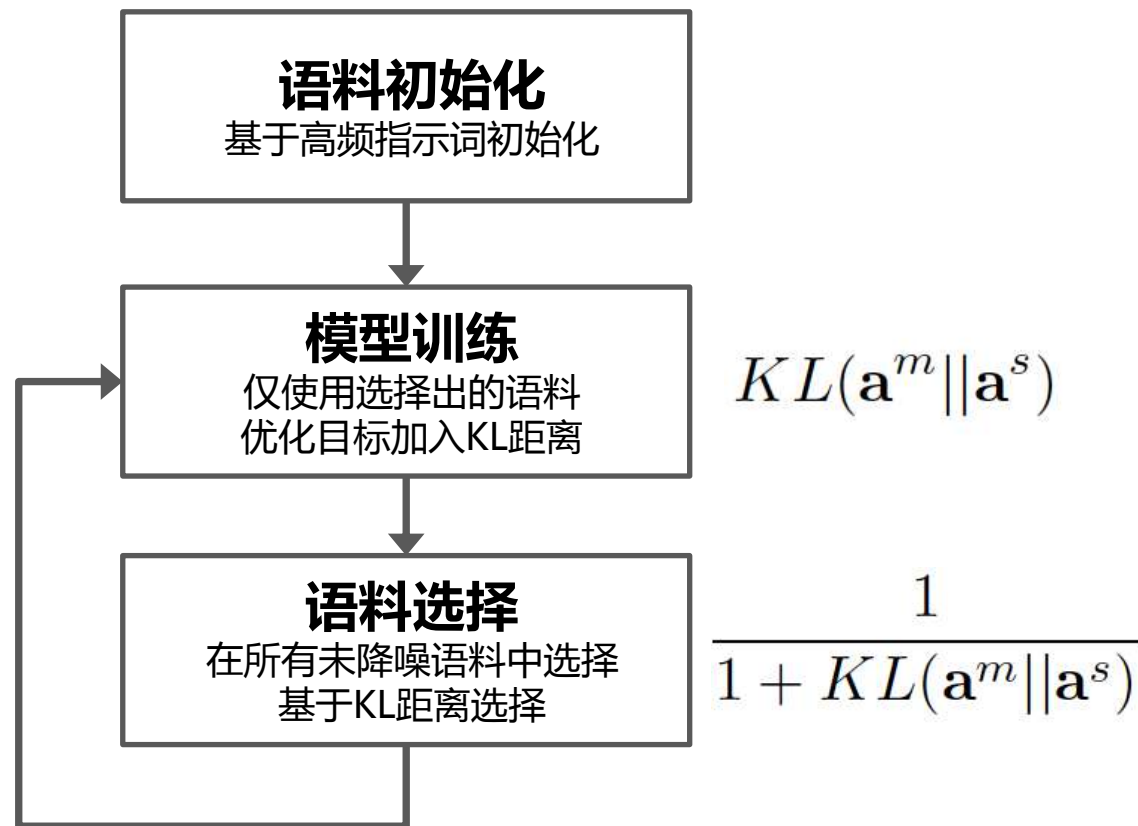
$$\mathbf{a} = \text{softmax}(\mathbf{w}^T \mathbf{M})$$



- 训练：基于Attention正则化，迭代选择可靠语料训练模型
- 我们希望模型attention分布 \mathbf{a}^s ，与期望分布 \mathbf{a}^m 尽量接近

$$KL(\mathbf{a}^m || \mathbf{a}^s) = \sum \mathbf{a}^m \log \frac{\mathbf{a}^m}{\mathbf{a}^s}$$

- \mathbf{a}^m 基于指示词计算得到，指示词集合随训练过程逐步挖掘扩展



小样本学习：利用知识进行远监督学习

- 实验数据为NYT数据：训练数据为远监督自动标注的数据，测试集合为人工标注数据
- 测试集扩展：允许每个句子标注多个关系。测试集合共551个句子，671个标注关系

方法对比

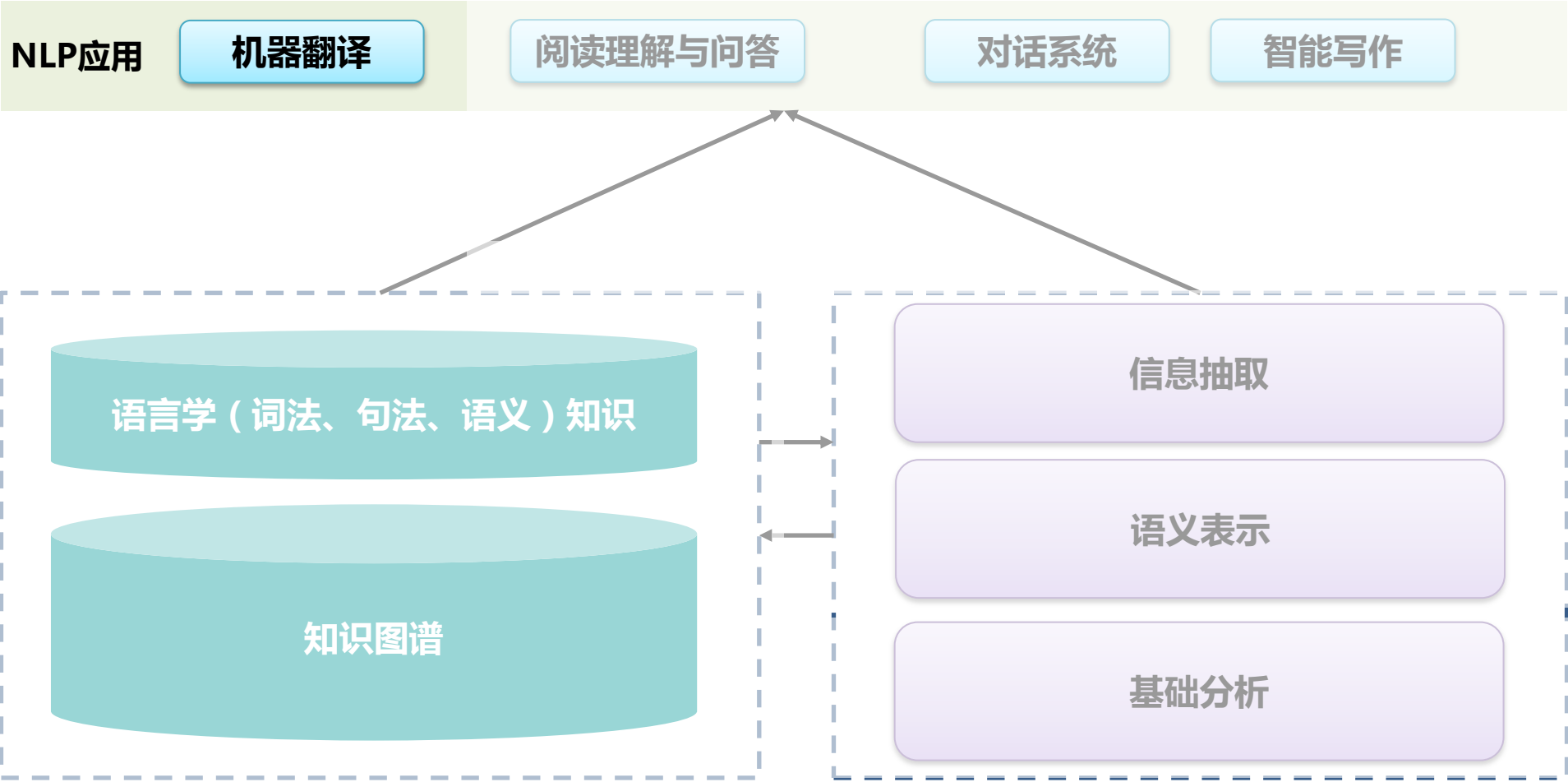
方法		P	R	F1
远监督学习	CNN (Zeng et al., 2014)	35.75	64.54	46.01
	PCNN (Zeng et al., 2015)	36.06	64.86	46.35
	Bi-LSTM	35.52	67.41	46.53
	BILSTM+ATT	34.93	65.18	45.38
远监督降噪	PCNN+SelATT(Lin et al., 2016)	45.41	30.03	36.15
	CNN+RL1(Qin et al., 2018)	39.41	61.61	48.07
	CNN+RL2(Feng et al., 2018)	40.23	63.78	49.34
	Ours	65.23	56.79	60.90

训练中各个策略的效果

方法	P	R	F1
BiLSTM + ATT	34.93	65.18	45.38
+模型初始化	70.95	40.57	51.63
+训练时使用Att. Reg.	68.70	50.99	58.53
+迭代语料选择	65.23	56.79	60.90

Att. Reg. 为Attention Regularization缩写

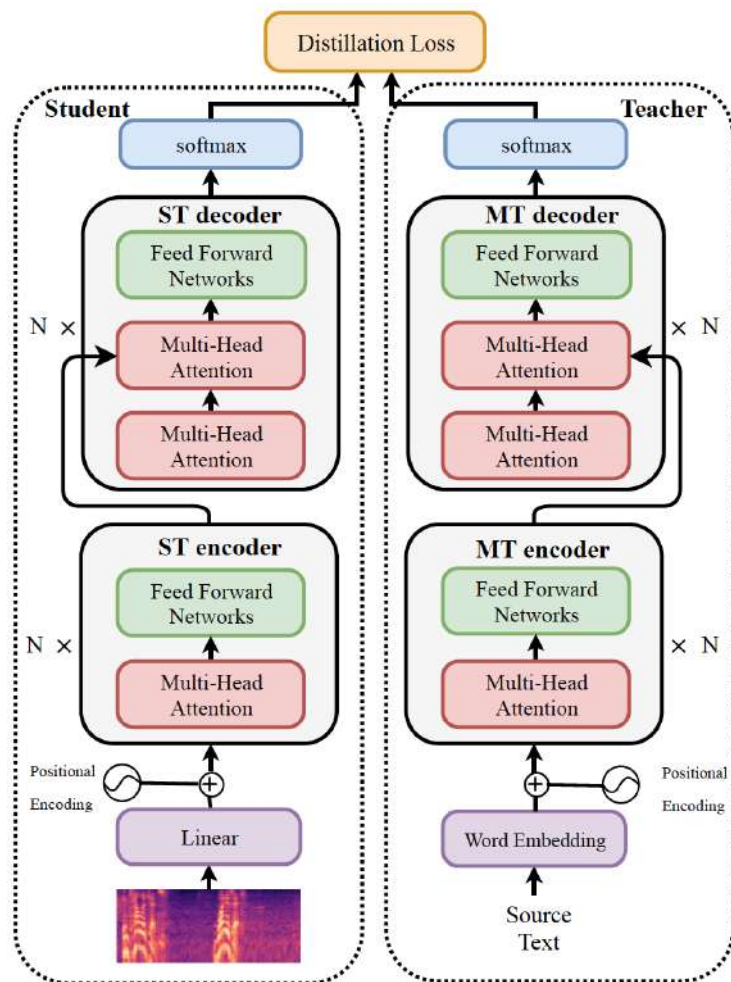
Outline



同传技术: Knowledge Distillation for End-to-End Speech Translation

Approach

Model: Good NMT model teaches bad Speech Translation model



$$L_{ST}(D; \theta) = - \sum_{(s,y) \in D} \log P(y|s; \theta)$$

$$\log P(y|s; \theta) = \sum_{t=1}^N \sum_{k=1}^{|V|} \mathbb{1}(y_t = k) \log P(y_t = k | y_{<t}, s; \theta)$$

$$L_{KD}(D; \theta, \theta_T) = - \sum_{(x,y) \in D} \sum_{t=1}^N \sum_{k=1}^{|V|} Q(y_t = k | y_{<t}, x; \theta_T) \log P(y_t = k | y_{<t}, x; \theta)$$

$$L_{ALL}(D; \theta; \theta_T) = (1 - \lambda) L_{ST}(D; \theta) + \lambda L_{KD}(D; \theta, \theta_T)$$

Results

LibriSpeech: English speech to French text

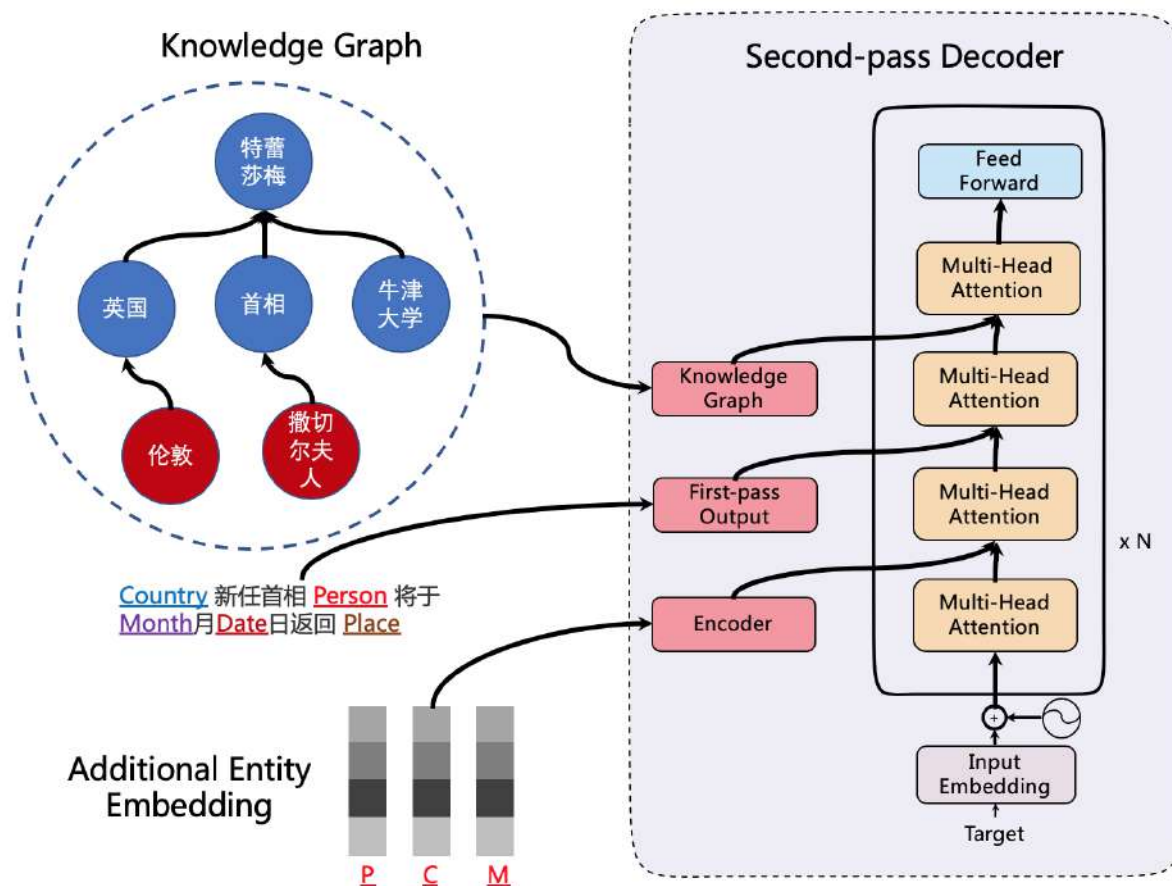
Training set: 100 hours, 47.3 million parallel sentences

Test set: 4 hours, 2048 sentences

LibriSpeech	Method	greedy	beam	ensemble
Bérard [10]	Pipeline	14.6	14.6	15.8
	End-to-end	12.3	12.9	15.5
	Pre-trained	12.6	13.3	15.5
Ours	Pipeline	15.75	17.85	18.4
	End-to-end	10.19	13.15	17.8
	Pre-trained KD	13.89	14.30	17.8
		14.96	17.02	

知识用于机器翻译: Two-pass Round Translation with Entity Reasoning

Leveraging Knowledge as enhanced attention



Preliminary Results

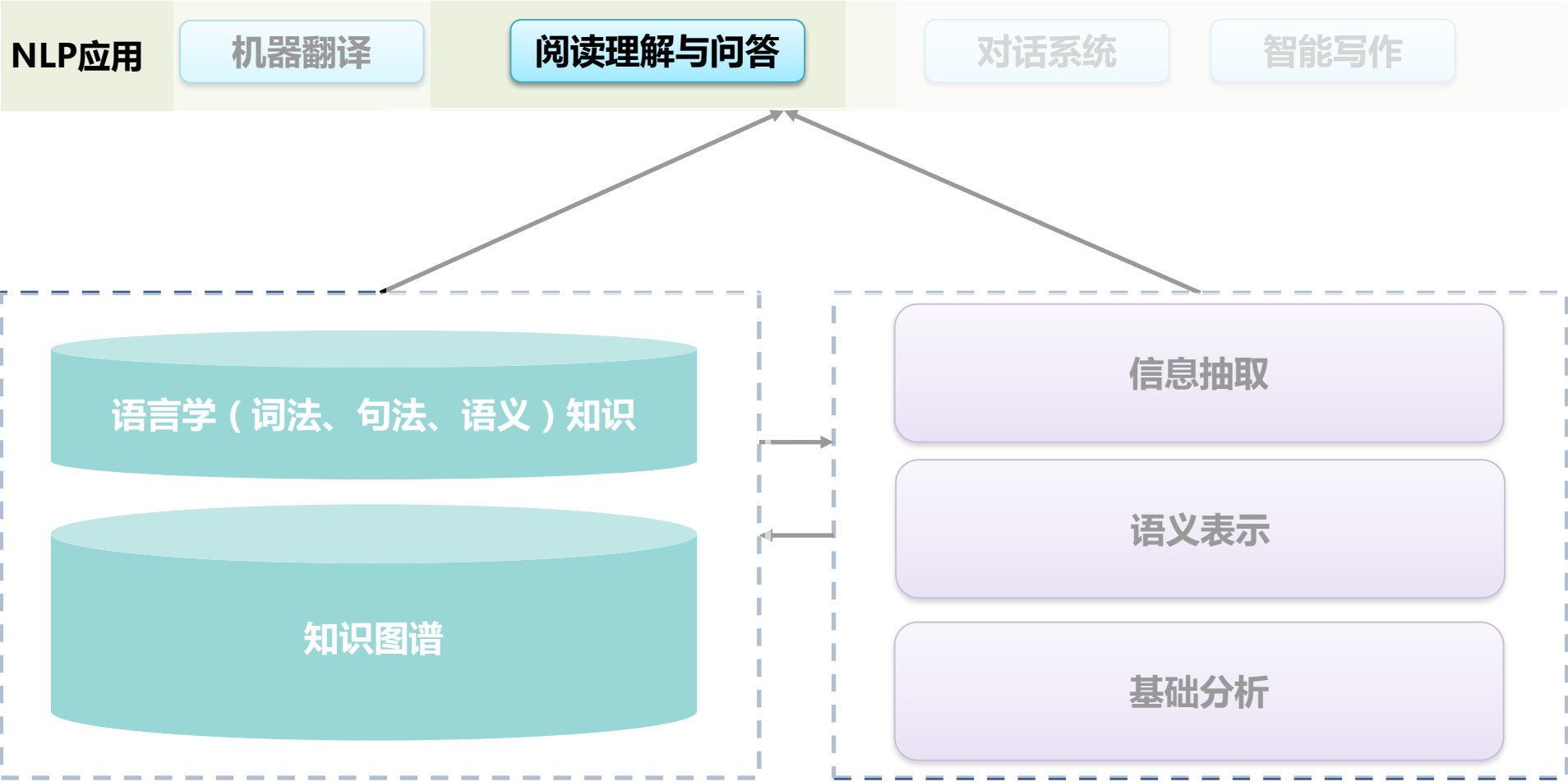
	BLEU	Entity Accuracy
Baseline	47.63	98.01
+ Entity Embedding	47.75	98.80

英国新任首相梅艳芳将于8月24日返回伦敦

英国新任首相特蕾莎梅将于8月24日返回伦敦

Input: The new British Prime Minister Teresa Mei will return to London on August 24

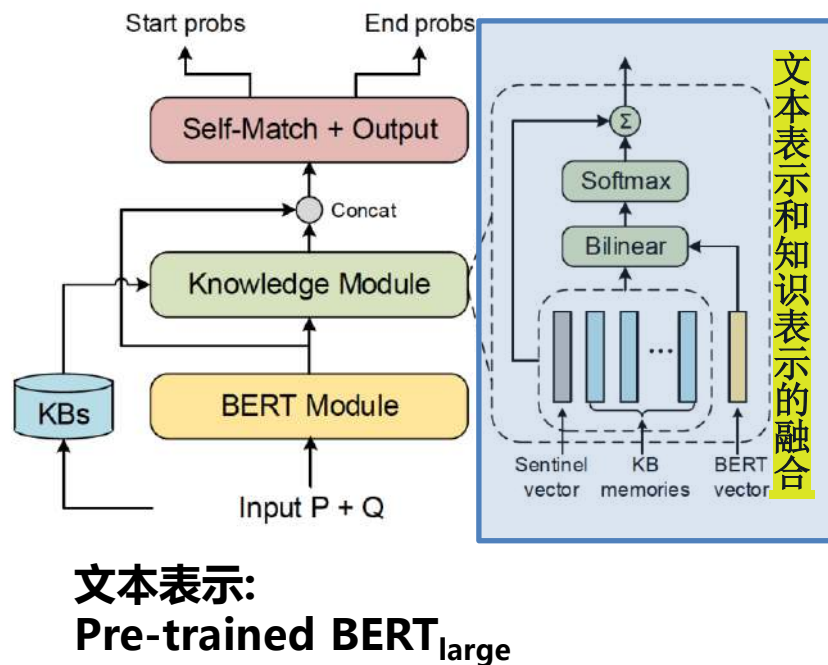
Outline



知识用于阅读理解

文本表示和知识表示的融合模型应用于阅读理解

知识表示:
Pre-trained BILINEAR
on NELL and WordNet



$$\gamma_{ij} \propto \exp(\mathbf{u}_j^\top \mathbf{H} \mathbf{s}_i^L)$$

\mathbf{u}_j 代表知识, \mathbf{s}_i^L 上一层输入

$$\tau_i \propto \exp(\mathbf{v}^\top \mathbf{H} \mathbf{s}_i^L)$$

- 知识的表示

$$\mathbf{t}_i = \sum_j \gamma_{ij} \mathbf{u}_j + \tau_i \mathbf{v}$$

知识用于阅读理解：实验结果

ReCoRD上的结果

Model	Dev		Test	
	F1	EM	F1	EM
(Leaderboard)				
Human	91.64	91.28	91.69	91.31
#1 DCReader+BERT	—	—	71.98	70.49
#2 BERT _{BASE}	—	—	57.99	55.99
#3 DocQA w/ ELMo	45.39	44.13	46.65	45.44
#4 SAN	39.09	38.14	40.72	39.77
#5 DocQA	37.89	36.59	39.76	38.52
(Our Models)				
BERT	72.16	70.22	—	—
BERT-KB _{WordNet}	72.75	70.56	—	—
BERT-KB _{NELL}	72.52	70.54	—	—
BERT-KB _{BOTH}	73.61	71.60	74.76	73.01

SQuAD 1.1上效果

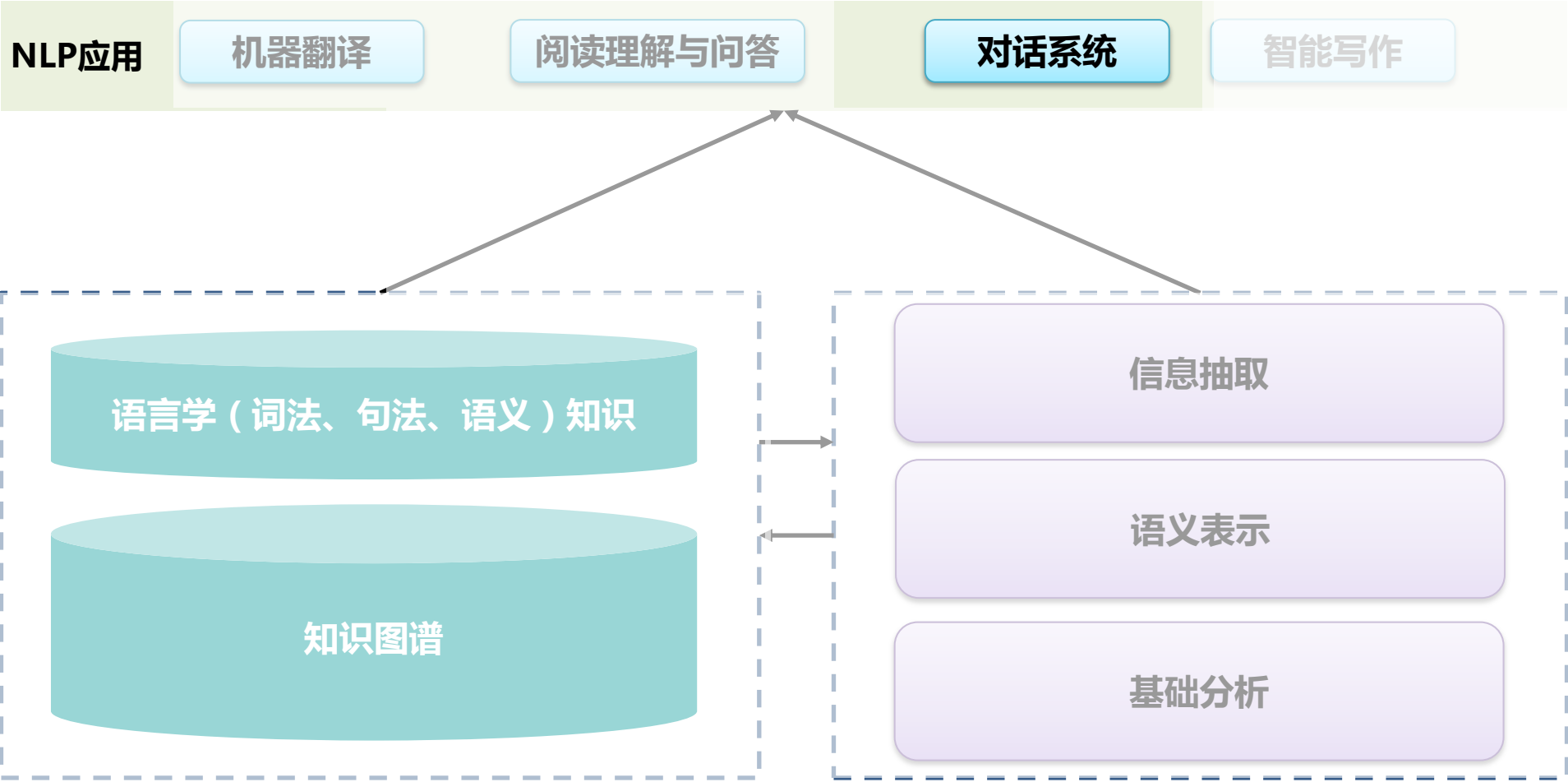
Model	Dev		Test	
	F1	EM	F1	EM
(Leaderboard)				
Human	90.5	80.3	91.22	82.30
#1 BERT+TriviaQA	91.1	84.2	91.83	85.08
#2 WD	—	—	90.56	84.40
#3 nlnet	—	—	90.13	83.47
#4 MARS	—	—	89.55	83.19
#5 QANet	—	—	89.31	82.47
(Our Models)				
BERT	91.24	84.41	—	—
BERT-KB _{WordNet}	91.70	85.15	92.43	85.94
BERT-KB _{NELL}	91.69	85.02	—	—
BERT-KB _{BOTH}	91.64	84.96	—	—

阅读理解应用于搜索问答

- 搜索问答中短答案的抽取（图1）
- 搜索问答中长答案的主干定位（图2、3）
 - 应用阅读理解技术前，长答案展示时**50%左右bad case**均为答案主干未露出，影响体验



Outline



开放域对话的问题

- 基于上下文的单轮对话系统为主

- **机器回答是被动的**

- 首先解决的是相关性问题、其次是信息丰富度的问题

- 技术的趋势是：在检索或者生成模型中引入知识（比如：CCM模型）

- **机器回答是主动的：以ask follow-up questions 为主**

在今年ACL上有两篇有代表性的文章：

- Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information
- Learning to Ask Questions in Open-domain Conversational Systems with Typed Decoders

How to configure path or set environment variables for installation?

i'm aiming to install ape, a simple code for pseudopotential generation. i'm having this error message while running ./configure
<error message>
So I have the library but the program installation isn't finding it.
Any help? Thanks in advance!

Initial Post

Question
comment

What version of ubuntu do you have?

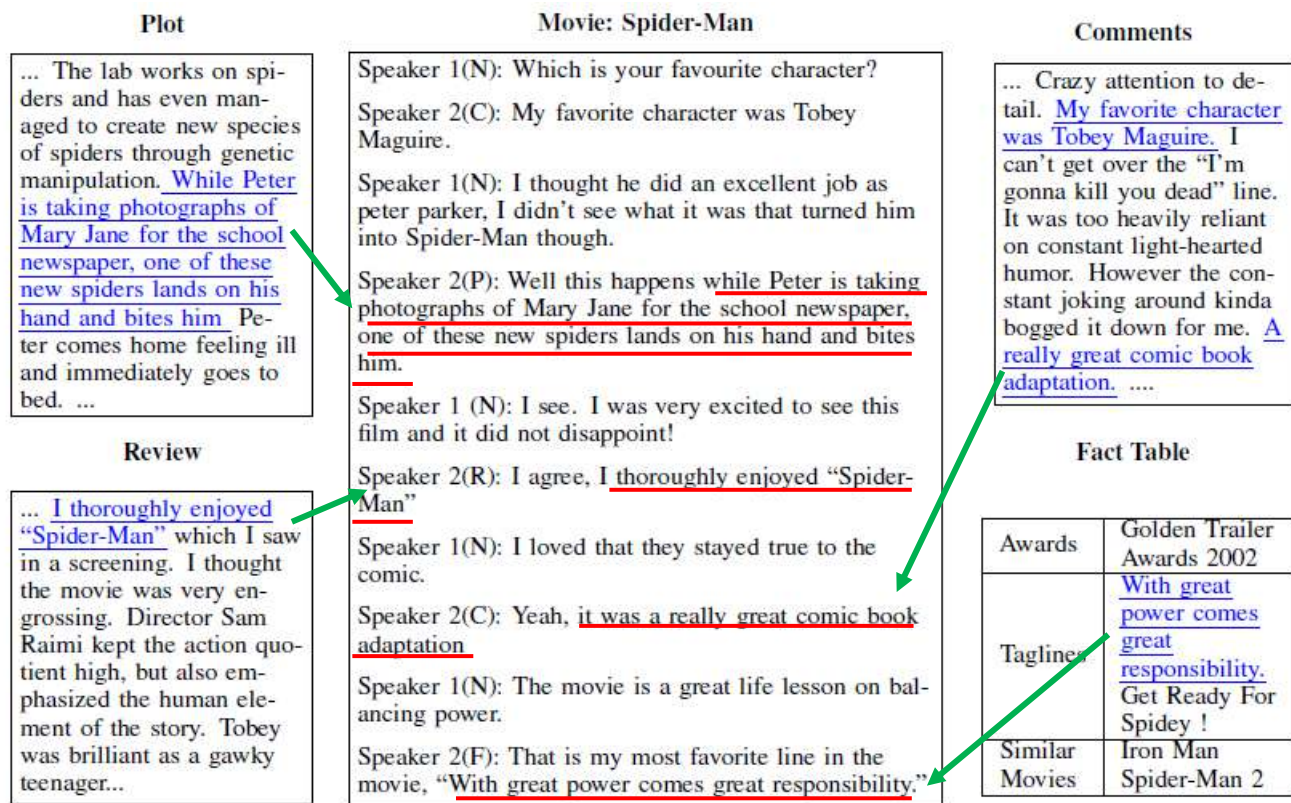
i'm aiming to install ape in Ubuntu 14.04 LTS, a simple code for pseudopotential generation. i'm having this error message while running ./configure
<error message>
So I have the library but the program installation isn't finding it.
Any help? Thanks in advance!

Edit as an answer
to the question

Updated Post

开放域对话的问题

- 有明确多轮的对话系统：主要是限定知识下的多轮系统
 - 基于Background information的多轮对话数据集



➤ 优点

- 引入了异构知识：结构化和非结构化
- 结合了对话和阅读理解

➤ 缺点

- 任务目标不好评估
- 没有真正建模对话逻辑（与目标不完全清晰有关）

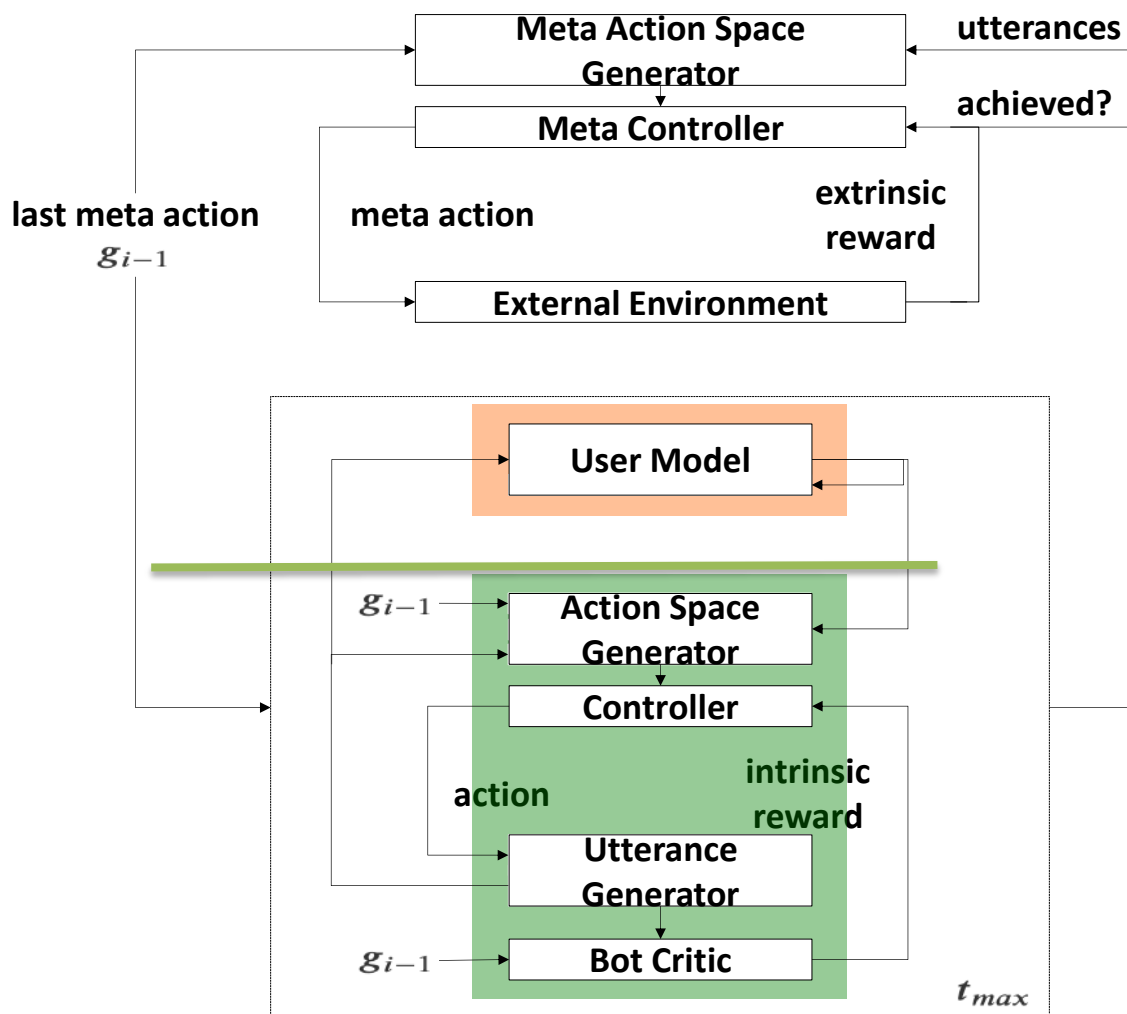
我们的目标：知识驱动的可控对话系统

- 可控：自动规划对话路径，做到topic之间自由切换
- 有内容：知识驱动的多轮对话

	任务对话	闲聊系统	知识驱动 固定Topic聊天	ours
明确的对话目标	Y	N	N	可以自动规划目标
有内容	Y	N	Y	Y
可控	Y	N	N	Y
系统主动	Y	N	Y	Y

知识驱动的对话系统：层次化RL方法

- 基于知识图谱的可控聊天
- 层次化RL方案
 - 聊什么：上层基于知识图谱规划聊天目标或者路径 (topic)
 - 怎么聊：下层基于对话语料建模完成目标
- 优势
 - 直接建模知识对话的控制机制
 - 框架灵活，通过 reward 可以定制对话逻辑
 - RL+SL 优于 SL，可以降低对训练数据的依赖



知识驱动的对话系统：效果

Bot: 看过桥口亮辅导演的电影吗？

User: 没看过，是什么电影？

Bot: 《周围的事》，这部电影是木村多江主演的，推荐你去看下。

User: 哈哈，我看过，这部电影的电影确实不错。

Bot: 嗯嗯，我还知道一部剧情片，是一部剧情片，我觉得还可以。

User: 你说的是哪部？

Bot: 《一个陌生女人的来信》，不过口碑一般。

User: 我看过了，确实不错。

Bot: 嗯嗯，我还知道一部日本的电影，是佐藤祐市导演的。

User: 不是很清楚，叫什么名字？

Bot: 《我在一家黑公司上班，已经快撑不下去了》，网友评论，简直是回味了当初那半年恐怖的实习啊！。

User: 这电影名字听起来很有意思。

Bot: 嗯嗯，2008年6月的电影，还有一部也是2008年6月的电影，也推荐给你吧。

User: 是什么电影啊？

Bot: 《糊涂侦探》，虽然口碑一般，但是我觉得还不错，你可以看看。

User: 好的，谢谢推荐。

Bot: 嗯嗯，还有一部口碑不错的电影也推荐给你吧。

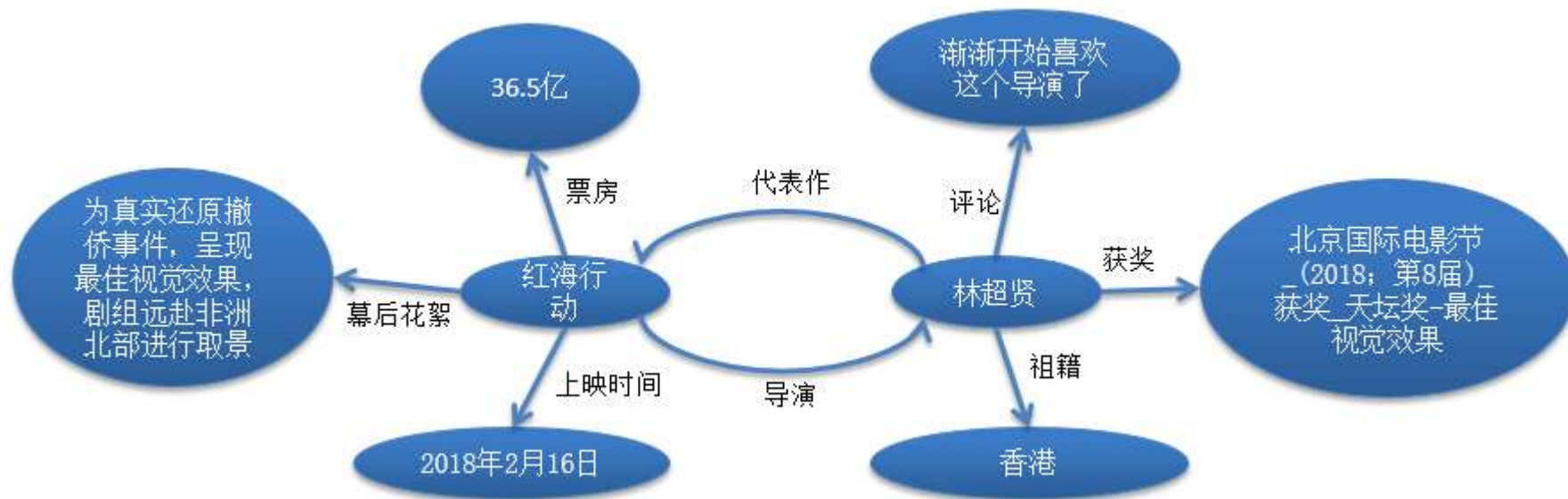
User: 说来听听。

Bot: 《夺宝奇兵3》，有评论说，惊险刺激，扣人心弦，大快人心，happyending。

User: 那我得去看看了。

知识驱动的对话系统：组织竞赛

- 基于知识图谱主动引导对话
- **有明确目标**：start -> topic a -> topic b
- 数据：3万个对话，约12万轮对话，其中10万训练集，1万开发集，1万测试集
- 比赛：在“2019语言与智能技术竞赛”设立了知识驱动对话比赛：<http://lic2019.ccf.org.cn/talk>
报名团队数量，1536，提交团队139，现处于打榜阶段。



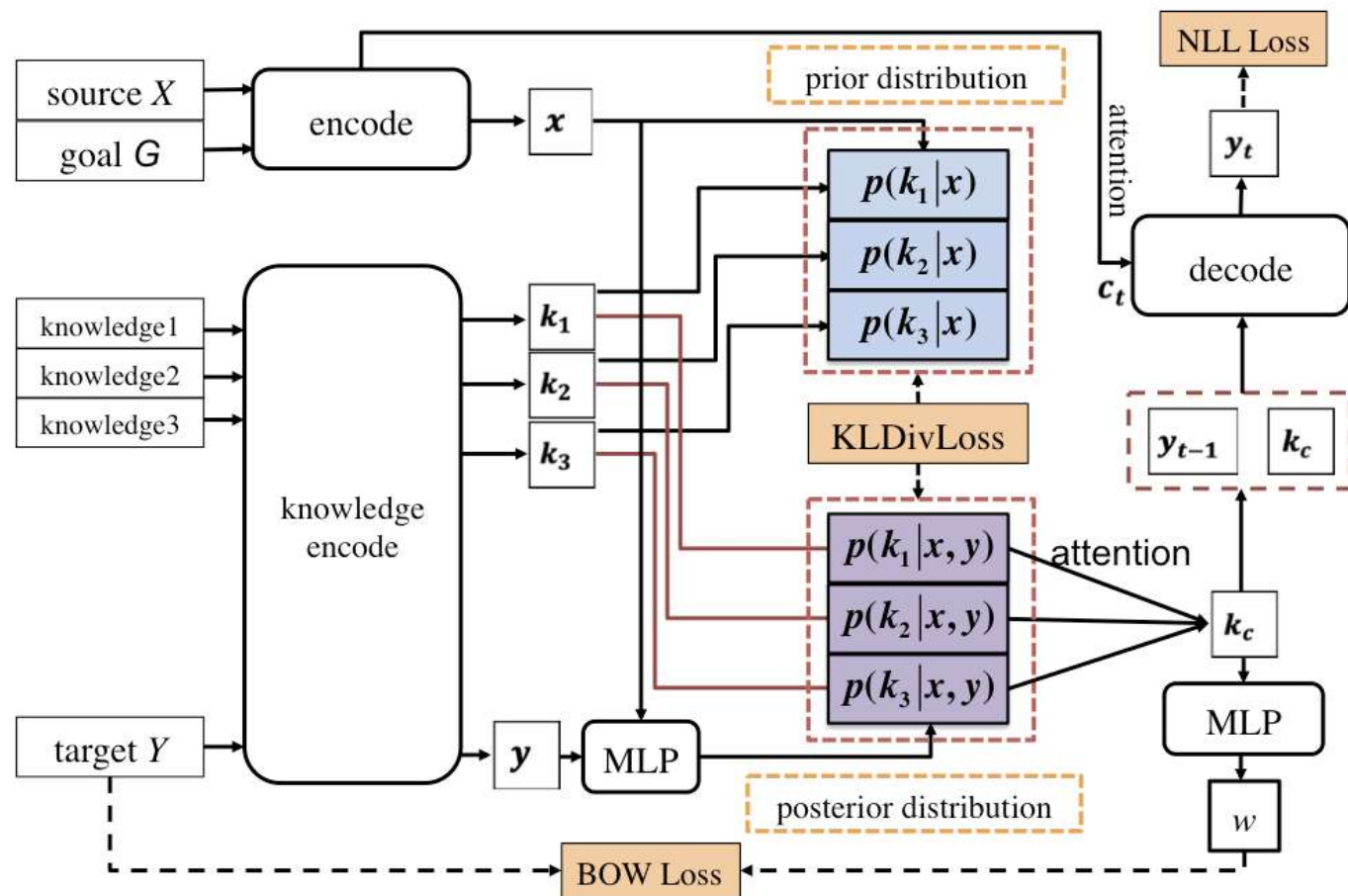
知识驱动的对话系统：组织竞赛

- 开源基线：在github上开源了主动对话模型：

<https://github.com/baidu/knowledge-driven-dialogue/>

- Paper : *Learning to Select Knowledge for Response Generation in Dialog Systems.*

<http://cn.arxiv.org/abs/1902.04911>



知识驱动的对话系统

Goal	[START]->麦兜·饭宝奇兵->彭博 [START]->McDull: Rise of the Rice Cooker->Bo Peng				
Knowledge	麦兜:饭宝奇兵 McDull: Rise of the Rice Cooker	评论 comment	不错，很搞笑 nice, very hilarious	norm generation	Bot: 你喜欢看什么小怪兽的电影啊？(Do you like <u>little monster</u> movie ?)
		评分 rating	6.9		User: 小孩子看的玩意儿了(I believe that is for babies)
		上映时间 date published	2016年9月15日 2016.9.15		Bot: 我最近看了一部叫麦兜·饭宝奇兵的电影，我觉得挺好看的。 (I recently watched one. <u>McDull: Rise of the Rice Cooker</u> . It's <u>very hilarious</u>)
		类型 genre	动画 cartoon		User: 看来你是个宝宝啊 (You must be a baby)
		主演 Hero	彭博 Bo Peng		Bot: 嗯，这是一部很搞笑的电影，不过我觉得还挺搞笑的。(That is a <u>very hilarious</u> anyway.)
		评论 comment	这个小怪兽也太恶心了 this little monster is disgusting		User: 真的吗？谁演的？(Really? Who is the hero?)
	彭博 Bo Peng	评论 comment	缺少爆发力 lack of explosive performance	norm Seq2seq	Bot: 这部电影的 <u>主演是彭博</u> ，我很欣赏他的。(The <u>starring is Bo Peng</u> . I admire him very much)
		血型 blood type	O型 type O		Bot: 你喜欢看动画片吗？(Do you like <u>cartoon</u> ?)
		性别 gender	男 male		User: 小时候还挺喜欢了，大了就不怎么喜欢看了 (I used to when I was a little baby. But I quit cartoon after growing up)
		职业 occupation	演员 actor		Bot: 那你看过麦兜·饭宝奇兵吗？(So have you ever watched <u>MuDull: Rise of the Rice Cooker</u> ?)
		代表作 representative work	麦兜:饭宝奇兵 McDull: Rise of the Rice Cooker		User: 没看过，讲什么的 (Not yet, what is in the movie?)
					Bot: 这部电影的 <u>主演是彭博</u> 。(The <u>starring is Bo Peng</u> .)
norm Retrieval				Bot: 我很喜欢看导演安格斯·麦克莱恩的电影。(I like <u>Angus Mclean's</u> Movie)	
				User: 什么电影呢 (for example ?)	
				Bot: 《 <u>麦兜·饭宝奇兵</u> 》这部电影是2016年9月15日上映的，你可以看一下。 (" <u>McDull: Rise of the Rice Cooker</u> ", it was <u>released on September 15th 2016</u> . You may have a try.)	
				User: 回头看一下(Ok, I will watch it later)	
				Bot: 这部电影是2016年9月15日上映的呢，这部电影的主演我很喜欢。 (It is <u>released on September 15th, 2016</u> , I really like the starring of the show)	
			User: 谁呀，你这么喜欢(Who is that?)		
			Bot: <u>彭博</u> ， <u>1957年3月20日出生的</u> ，经验很丰富的演员，你可以看看他的作品。 (<u>Bo Peng</u> , <u>born at March 20th, 1957</u> , a very experienced actor. You may take a look at his movies)		

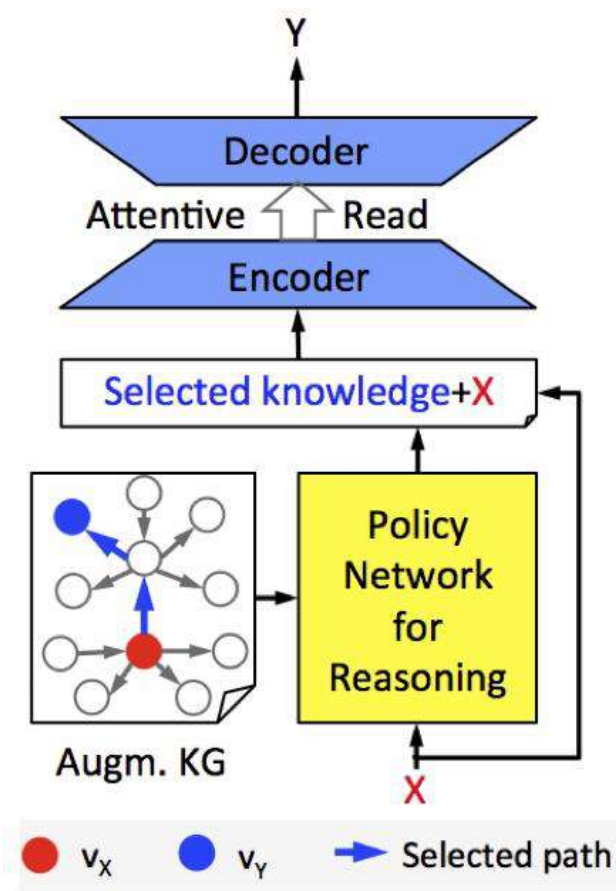
知识驱动的对话系统：基于图推理技术

- 基于图的两阶段模型
 - 基于上下文的知识选择
 - 基于知识的回复生成
- 图表示：融合结构化知识和非结构化知识
 - TransE表示节点和边的结构信息
 - Bert对长文本进行文本语义表示
- 知识选择：采用 Multi-hop 知识推理技术进行对话知识选择
 - 将输入 X 映射到知识图中节点 v_X
 - 将 X 具体将 X 过MLP后得到抽象relation 的表示
 - 基于 $v_X + \text{relation}(X)$ 寻找回复所需的 v_{gt}

论文已提交ArXiv

<https://arxiv.org/abs/1903.10245>

Knowledge Aware Conversation Generation with Explainable Reasoning on Augmented Graphs



知识驱动的对话系统：基于图推理技术

- 知识选取形式化为 finite-horizon 序列决策过程
 - 具体视为 POMDP 问题求解：因为节点在跳转过程中，无法获取 environment 的完整状态
 - 学习一个 RL based agent：因为缺失正确路径的标注信息，所以采用 RL 来学习 Policy Network

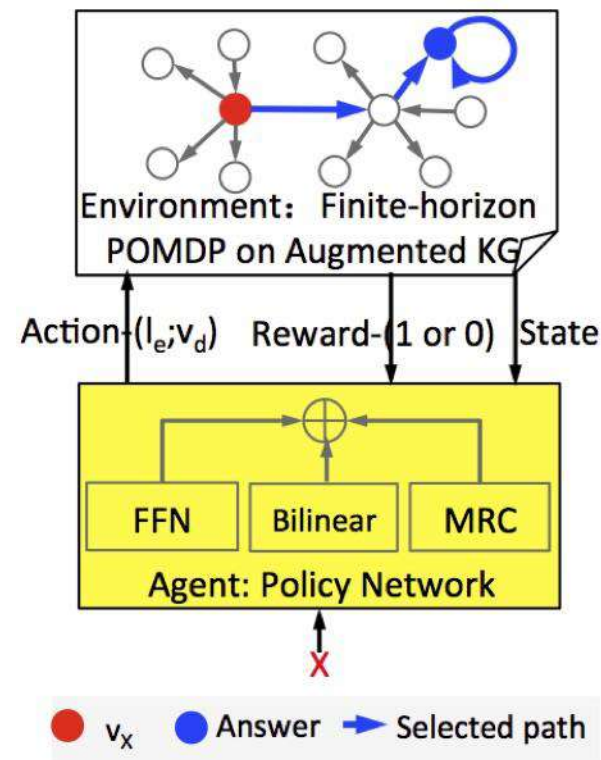
Rewards :

$$- \mathcal{R}(S_T) = I\{v_T = v_{gt}\} \in \{+1, 0\}$$

Policy Network :

$$- P(\mathcal{A}_{S_t}) = \text{softmax}(\alpha P_{FFN}(\mathcal{A}_{S_t}) + \beta P_{Bilinear}(\mathcal{A}_{S_t}) + (1 - \alpha - \beta) P_{MRC}(\mathcal{A}_{S_t}))$$

v_t 为 t 时刻节点， v_{gt} 为目标节点， P_{FFN} 、 $P_{Bilinear}$ 、 P_{MRC} 为打分策略



知识驱动的对话系统：实验结果

- 数据集：
 - 多轮电影知识聊天数据
 - 数据特点：
 - 领域集中在电影领域
 - 所有回复句所依赖的背景知识集合都明确给出
 - 所有回复句所使用的具体知识内容也明确标注
- 评估指标：
 - 知识选取：
 - BLUE-4、ROUGE-2、ROUGE-L
 - 回复生成：
 - BLUE-4、ROUGE-2、ROUGE-L；
 - 人工评估：采取 win : tie : loss 方式对比评估合适度和丰富度

Conversational Pairs			Augmented KG		
#Train. pairs	34486		#Vertices	117373	
#Valid. pairs	4388		#Relations	11	
#Test pairs	4318		#Triples	251138	
Factoid knowledge			Non-factoid knowledge		
#Total	ver-	21028	#Total	ver-	96345
tices			tices		
#Vertices	in	2740	#Vertices	in	31746
utterances			utterances		

Table 1: The upper table shows statistics of the Reddit dataset and associated knowledge base. The lower table provides the detailed statistics of the two types of knowledge mentioned in utterances.

知识驱动的对话系统：实验结果

	BLEU-4	ROUGE-2	ROUGE-L
Model	(%)	(%)	(%)
Seq2seq	1.59	5.73	14.49
HRED	2.80	13.45	23.45
MemNet	1.26	9.95	17.50
GTTP	11.05	17.70	25.13
BiDAF+G	32.45	31.28	36.95
CCM	2.54	4.40	15.25
AKGCM	31.93	31.87	37.09

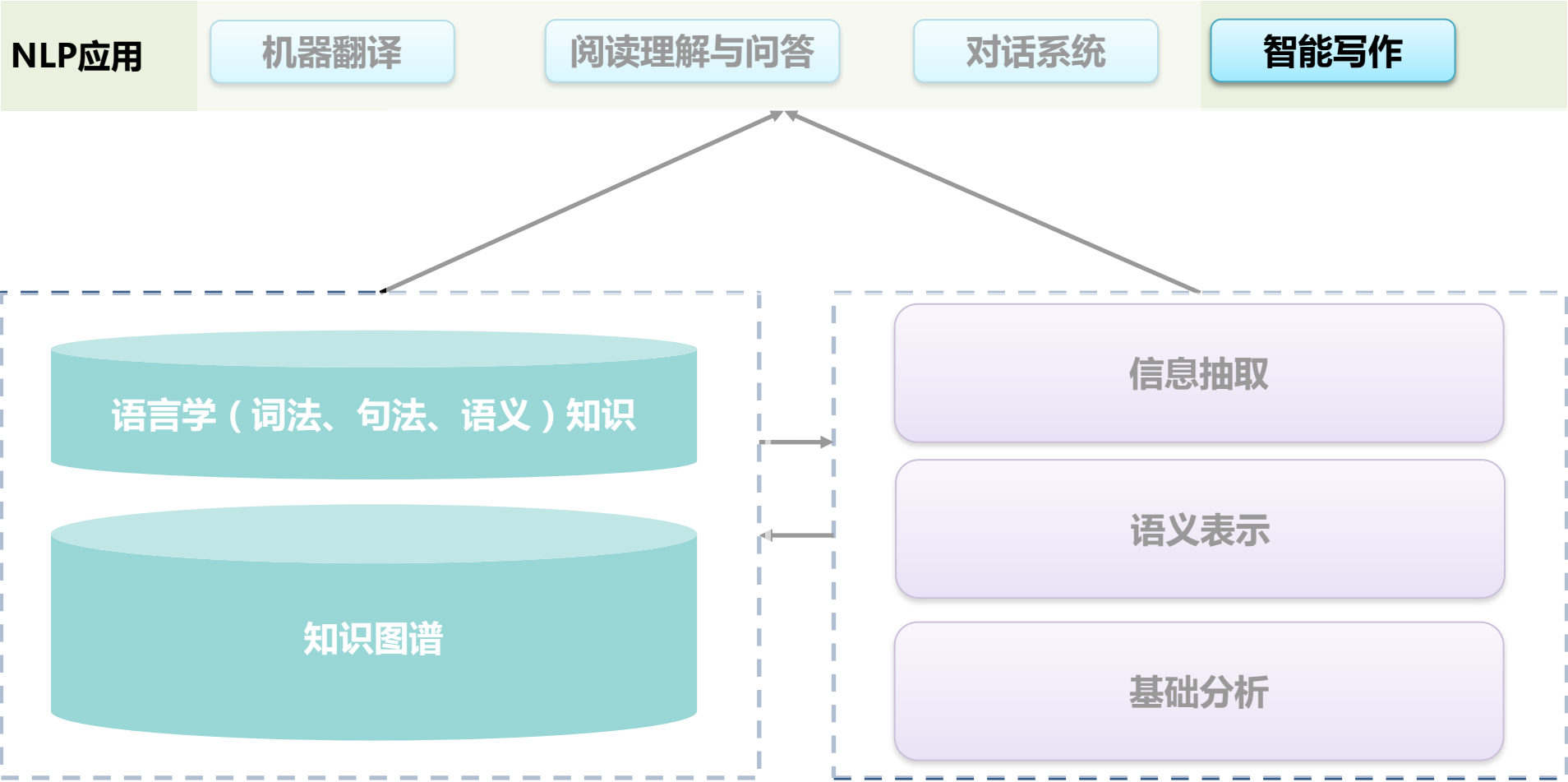
Table 2: Results of automatic evaluations. AKGCM outperforms all the baselines, except BiDAF+G, significantly (sign test, p -value $< 2.2e-16$) in terms of all the metrics.

	Appr.	Infor.
	AKGCM vs. *	AKGCM vs. *
Model	Win/Tie/Lose	Win/Tie/Lose
Seq2seq	0.63/0.28/0.09	0.76/0.16/0.08
HRED	0.78/0.19/0.03	0.81/0.17/0.02
MemNet	0.67/0.28/0.05	0.78/0.17/0.05
GTTP	0.49/0.36/0.15	0.49/0.34/0.17
BiDAF+G	0.20/0.68/0.12	0.20/0.67/0.13
CCM	0.83/0.16/0.01	0.84/0.16/0.00

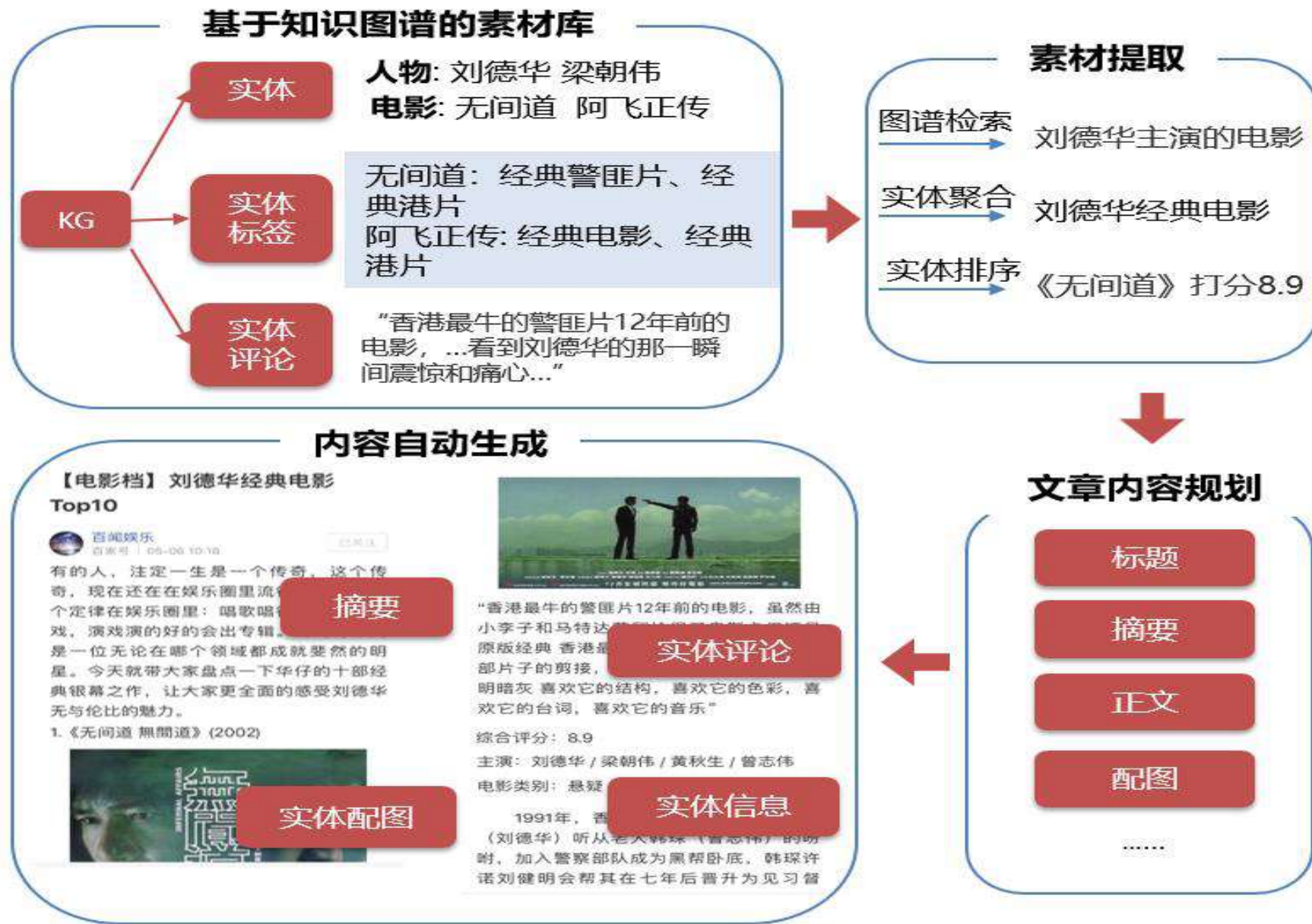
Table 3: Results of human evaluations. AKGCM outperforms all the baselines significantly (sign test, p -value < 0.005) in terms of the two metrics.

Movie	Affliction (1997)
Message	I thought she was interesting to watch, but I found Wade more intriguing. I was mortified with what happened to Glen.
Ground-truth response	Yeah , that was crazy when Wade douses Glen in gasoline and sets him and the garage on fire.
Seq2seq	It was an absurdist screwball comedy.
HRED	I liked Clerks , I loved Clerks 2.
MemNet	I think it was the best of the movie that is a great movie.
GTTP	I think it was a very slow moving novel like story but the character were played well.
BiDAF+G	I also liked the scene where Wade fight , and Wade clubs Glen , killing him.
Selected knowledge	Wade fight, and Wade clubs Glen, killing him. Numbly, Wade douses Glen in gasoline and sets him and the garage on fire.
CCM	A empathy organized Auto named Martin _UNK (his more) made him a audience talk of _NONE.
AKGCM	It was crazy when Numbly, Wade douses Glen in gasoline and sets him and the garage on fire.
Selected knowledge	Numbly, Wade douses Glen in gasoline and sets him and the garage on fire.

Outline

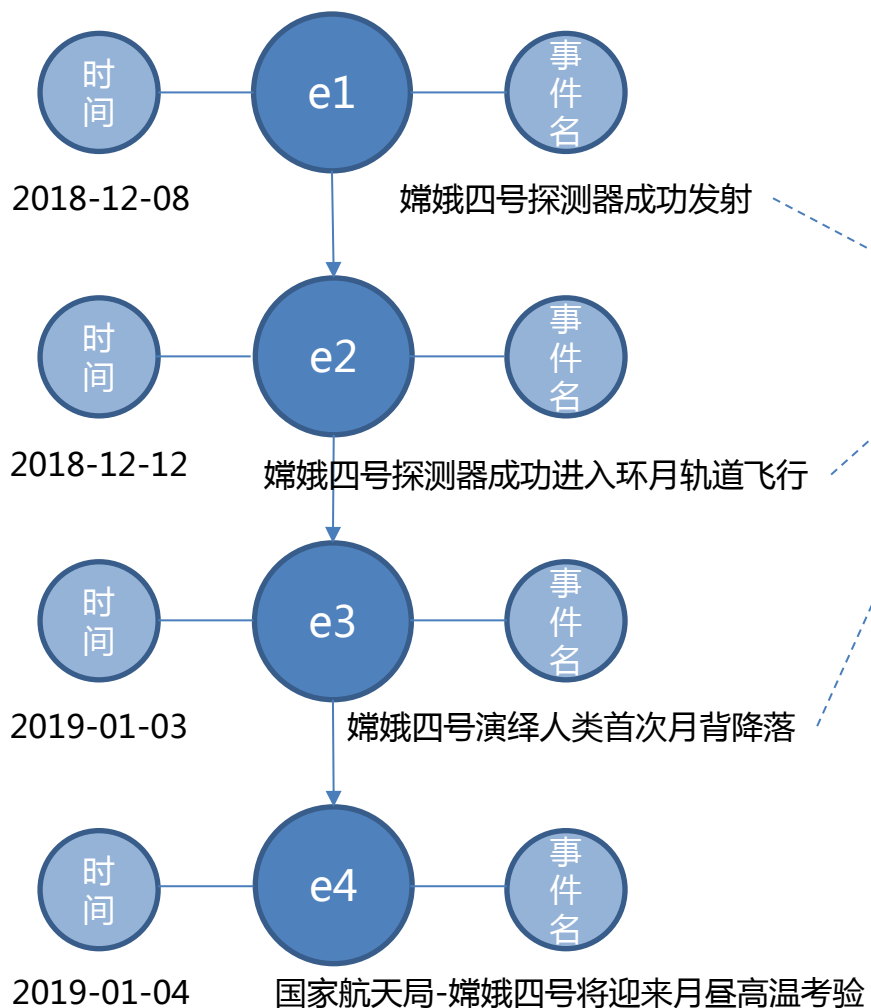


基于图谱的智能写作



智能写作平台

文本到文本生成 - 基于事件脉络生成聚合类文章



获取事件资讯内容

篇章规划

自动文摘

生成

嫦娥四号成功发射：一文看懂事件始末

百闻事件
百度号 01-15 21:20 已关注

截止到01月16日，嫦娥四号成功发射的最新进展如下：

国家航天局副局长、探月工程副总指挥吴艳华在14日国新办新闻发布会上表示，以嫦娥四号任务圆满成功为标志，我国探月工程四期和深空探测工程全面拉开序幕。嫦娥五号月面采样返回任务将于今年年底左右实施，我国首次火星探测任务将于2020年前后实施。国家航天局正在组织国内专家对后续规划进行论证，基本明确还有三次任务。中国实施探月工程对商业航天具有深远意义。国家航天局秘书长、新闻发言人李国平表示，中国的空间站目前正在建设过程当中，预计到2022年前后建成。在中国载人航天工程和空间站建设过程中，中国与多个国家和组织开展了载人航天探测方面的合作。



根据全网新闻报道，小编为大家整理了这一事件的来龙去脉如下：

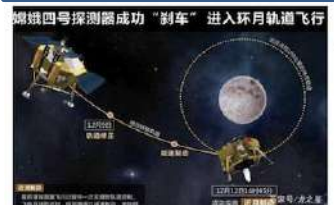
2018年12月8日：嫦娥四号探测器成功发射

12月8日2时23分，我国在西昌卫星发射中心用长征三号乙运载火箭成功发射嫦娥四号探测器，开启了月球探测的新旅程。嫦娥四号探测器后续将经历地月转移、近月制动、环月飞行，最终实现人类首次月球背面软着陆，开展月球背面就位探测及巡视探测，并通过已在使命轨道运行的“鹊桥”中继星，实现月球背面与地球之间的中继通信。



2018年12月12日：嫦娥四号探测器成功进入环月轨道飞行

中国国家国防科技工业局、国家航天局12日晚发布消息说，北京时间当天16时45分，嫦娥四号探测器经过约110小时奔月飞行，到达月球附近，成功实施近月制动，顺利完成“太空刹车”，被月球捕获，已进入近月点约100公里的环月轨道。据介绍，近月制动是月球探测器飞行过程中一次关键的轨道控制。根据实时遥测数据监视判断，嫦娥四号探测器顺利进入环月轨道，近月制动获得圆满成功。



2019年01月03日：嫦娥四号演绎人类首次月背降落

根据新华社消息，2019年1月3日10时26分，嫦娥四号探测器自主着陆在月球背面南极-艾特肯盆地内的冯·卡门撞击坑内，实现人类探测器首次在月球背面软着陆。12月30日，嫦娥四号探测器成功在环月轨道实施变轨控制，顺利进入着陆轨道，择机降落。1月3日，嫦娥四号探测器成功降落月球背面，这是人类首次登陆月球背面，也是中国探月工程的一个新的里程碑。



2019年01月04日：国家航天局-嫦娥四号将迎来月昼高温考验

此后，嫦娥四号将迎来月昼高温考

智能写作平台




百度智能写作平台v1.1已在AI开放平台上线，欢迎体验 ai.baidu.com/tech/intelligentwriting



应用场景



特色优势

-  领先的语言理解与生成能力
-  一站式提供创作能力
-  简单易用

自动写作效果

秒级生成文章
中长尾内容占比提升**2倍以上**
累计发文量**百万量级**，累计阅读量**十亿量级**



阅读量/天



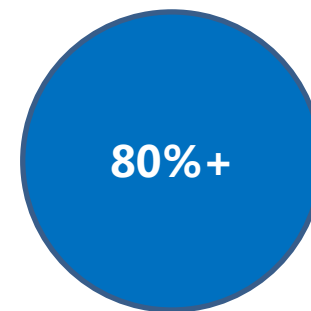
上线文章数/天



覆盖领域



覆盖类型



阅读完成率

开放和开源技术

NLP CLOUD SERVICE

Lexical Analysis

Joint word segmentation, POS tagging and NER

Text Correction

Identifying and correcting text errors

Word Embedding

Available for TB-size data training

Language Model

Available for large vocabulary output

Dependency Parsing

Improved with large-scale weakly-labeled data

Semantic Similarity

Embedding trained with large weakly-labeled data

Sentiment Analysis

Sentiment classification and opinion mining

Document Tagging

Document classification and document tagging

Find more at: ai.baidu.com/tech/nlp

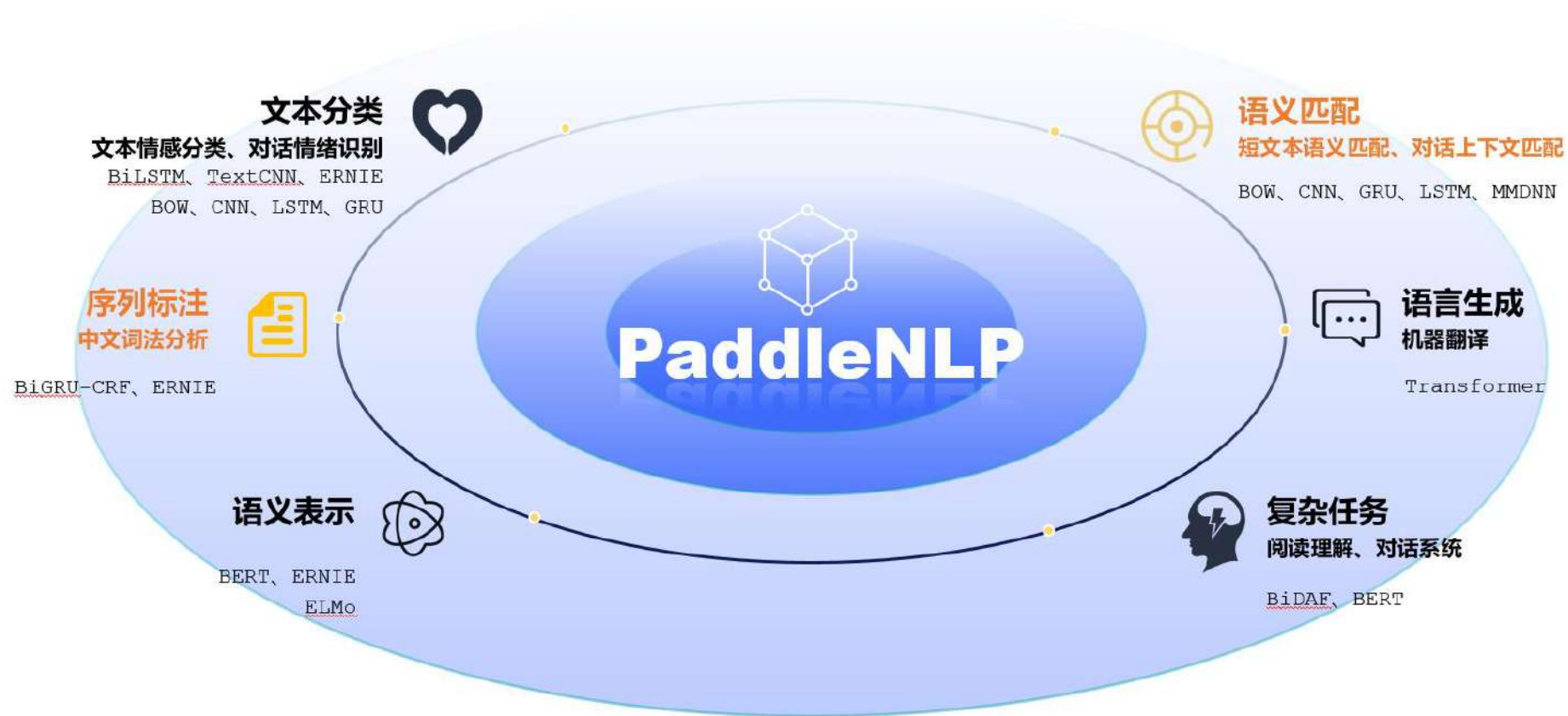
UNIT: 理解与交互平台

百度翻译开放平台

智能写作平台

<http://ai.baidu.com/tech/nlp>

开放和开源技术



<https://github.com/PaddlePaddle/models/tree/develop/PaddleNLP>

发布数据，和业界同行一起举办竞赛

2019语言与智能技术竞赛

提供面向真实应用场景的阅读理解、信息抽取及基于知识图谱的对话数据集，
促进学术研究，提升语言理解的人机交互的智能水平，推动语言与智能领域技术发展和应用。

CCMT 2019 - BSTC

BSTC (Baidu Speech Translation Corpus) is a large-scale dataset for automatic simultaneous interpretation. BSTC version 1.0 contains 50 hours of real speeches, including three parts, the audio files, the transcripts, and the translations. The corpus can be used to build automatic simultaneous interpretation system.



机器阅读理解

挑战当前优秀的阅读理解系统尚不能正确回答的问题，全面评测机器进行深度语言理解以回答复杂问题的能力。



知识驱动对话

探索机器模拟人类基于知识主动聊天的能力，开放12万轮中文对话数据。



信息抽取

提供业界最大规模的中文信息抽取数据集，让机器具备从海量自然语言文本中自动获取知识的能力。

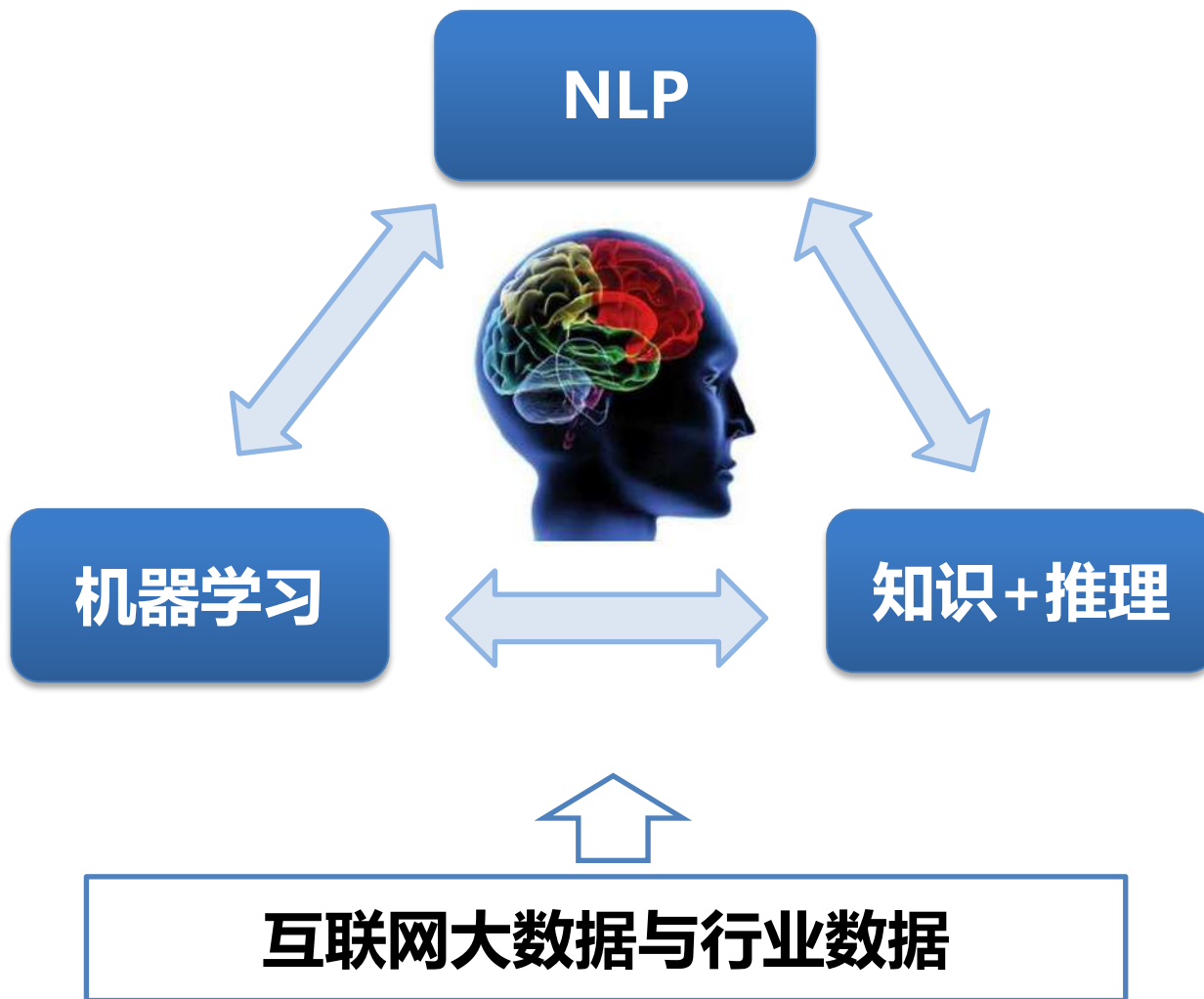
BSTC



BSTC version 1.0 contains 50 hours of real speeches, including three parts, the audio files, the transcripts, and the translations. The corpus can be used to build automatic simultaneous interpretation system.

- 总报名队伍数量：2,376
- 总提交次数：共420个队伍参与打榜，提交2,357次。

自然语言处理：We are still on the way





Thanks!