中国科学院大学
University of Chinese Academy of Sciences

自然语言处理
Natural language Processing

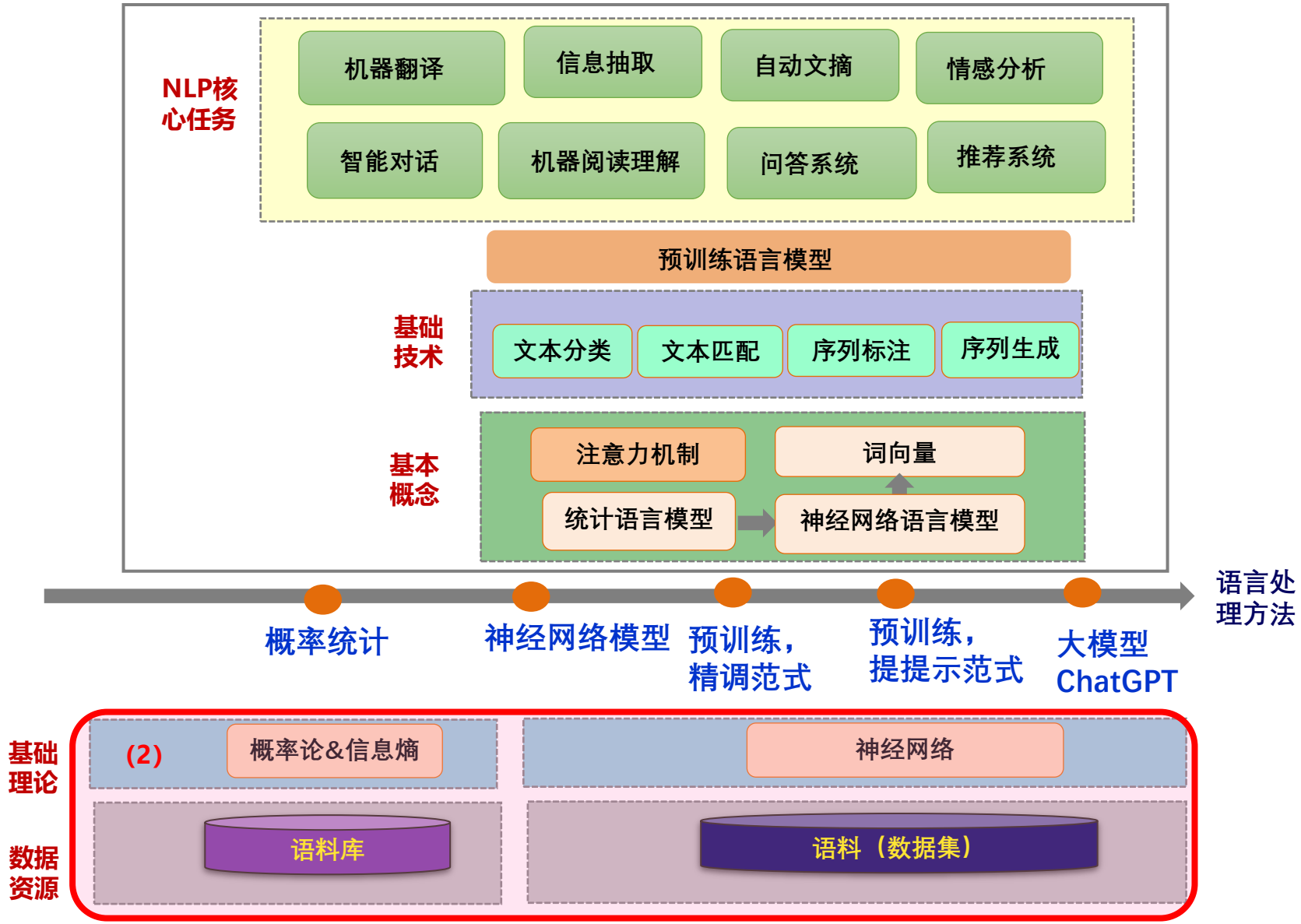
授课教师：胡玥

2025.09

课程编码： 1800860812O3P2002H 课程名称：自然语言处理 授课团队 胡玥 、陈钰枫

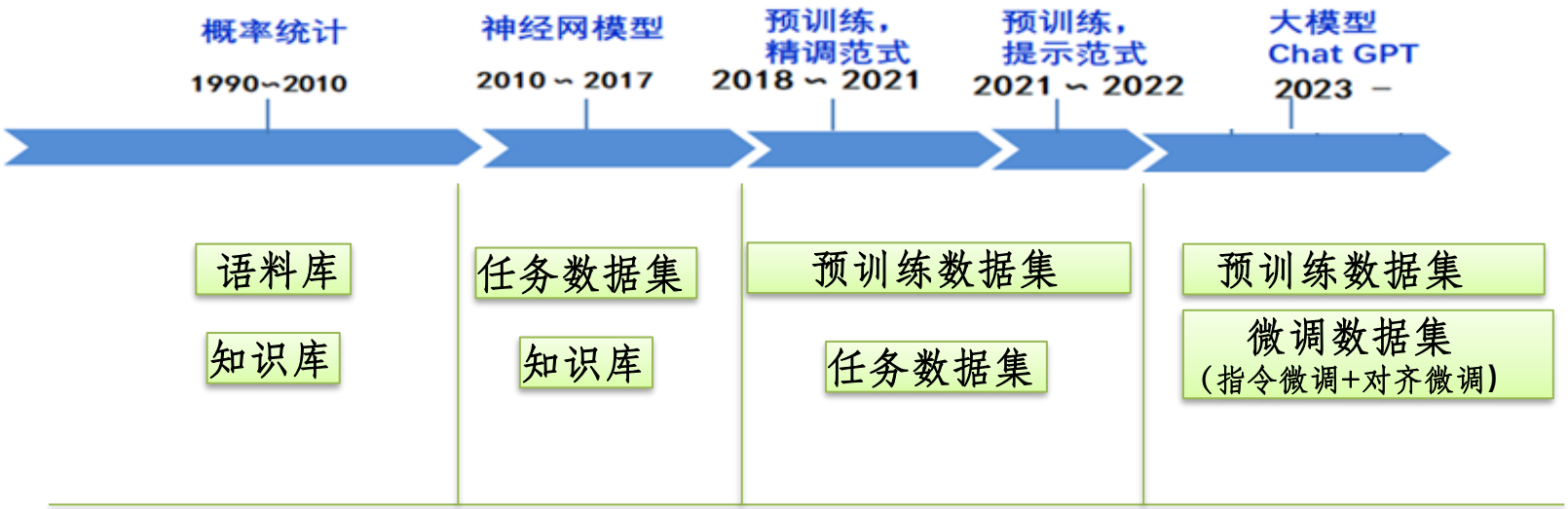
第 2 章　数据资源

基于深度学习的自然语言处理授课体系



内 容 提 要

2.1 数据资源概述



数据资源

语料库　任务数据集　预训练数据集　微调数据集

知识库　知识图谱　常识图谱　指令微调数据　对齐微调数据

2.1 数据资源概述

数据资源发展历程

数据资源按机器学习方法分为

- 概率统计时期数据资源：包括第一代~第三代语料库 ，主要是构建大型的通用语料库
- 深度学习数据资源：第四代语料，主要来源于互联网的海量数据

第一代（1970 - 80年代)

百万词级， 以语言研究为导向 。 如，Brown语料库 ， LLC语料库 等

第二代（1980 - 90年代)

千万词级， 词典编撰 --应用导向 。

如，COBUILD语料库 （2000万词级）Longman语料库

第三代（1990年代- 2010年代 ）

超大规模（上亿词级），标准编码体系，深度标注/多语种，NLP应用，
如， ACL/DCI语料库 ，UPenn树库 ，LDC 等

第四代（2010 深度学习)

互联网信息作为语料库

内 容 提 要

2.1　数据资源概述

2.2　统计时代语料资源

1. 语料库

2. 知识库

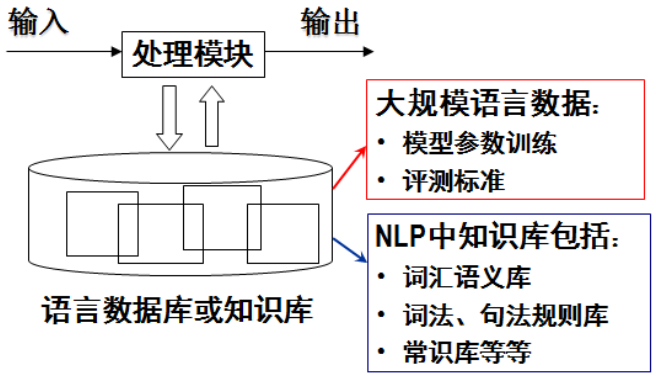
2.3　深度学习时代数据资源

1. 语料库

■ 语料库(corpus)基本概念

存放语言材料的仓库，现代的语料库是指存放在计算机里的原始语料文本或 经过加工后带有语言学信息标注的语料文本。

以语言的真实材料为基础来呈现语言知识，反映语言单位的用法和意义，基本以知识的原始形态表现——语言的原貌



1. 语料库

■ 语料库基本概念

语料库有三点特征

- 语料库中存放的是在实际使用中真实出现过的语言材料；
- 语料库是以计算机为载体承载语言知识的基础资源，但并不等于语言知识；；
- 真实语料需要经过分析、处理和加工，才能成为有用的资源。

语料库的作用

- 支持语言学研究和语言教学研究
- 支持NLP系统的开发

1. 语料库

■ 语料库的类型及相关术语

□ 按内容构成和目的划分 (4种类型)

◆ 异质的 (heterogeneous) - [黄昌宁，2002]

最简单的语料收集方法，没有事先规定和选材原则。

◆ 同质的（homogeneous）

与“异质”正好相反，比如美国的 TIPSTER 项目只收集军事方面的文本。

◆ 系统的（systematic）

充分考虑语料动态和静态问题、代表性和平衡问题以及语料库规模等问题。

◆ 专用的（specialized）

如：北美的人文科学语料库。

1. 语料库

■ 语料库的类型及相关术语

□ 按语言种类划分

- ◆ 单语的
- ◆ 双语的或多语的

平行语料库： 篇章对齐 / 句子对齐 / 结构对齐

例如，机器翻译中的双语对齐语料库

C: 早晨好！
E: Good morning.
C: 您能给我一杯咖啡吗？
E: Could you give me a cup of coffee?
… …

C: 早晨$_1$ 好$_2$ ！$_3$
E: Good$_2$ morning$_1$.$_3$

1. 语料库

■ 语料库的类型及相关术语

□ 按是否加工处理过（标注）划分

- ◆ 生语料库：未经加工的，没有任何切分、标注标记的原始语料库
- ◆ 熟语料库：经过加工，带有切分、标注标记的语料库
  - 具有词性标注
  - 句法结构信息标注(树库)
  - 语义信息标注

1. 语料库

■ 语料库的类型及相关术语

□ 共时语料库与历时语料库

◆ 共时语料库 是为了对语言进行共时(同一时段)研究而建立的语料库。研究大树的横断面所见的细胞和细胞关系，即研究一个共时平面中的元素与元素的关系。

◆ 历时语料库 是为了对语言进行历时研究而建立的语料库。研究大树的纵剖面所见的每个细胞和细胞关系的演变，即研究一个历时切面中元素与元素关系的演化。

1. 语料库

■ 典型的语料库资源

★布朗语料库 (Brown Corpus)

- 20世纪60s，Francis 和 Kucera 在布朗(Brown)大学建立，是世界上第一个根据系统性原则采集样本的标准语料库，100万词规模；
- 15种题材，共500个样本，每个样本不少于2000词；
- 1970s Greene 和 Rubin 设计了TAGGIT词性标注系统（词类标记81种，上下文约束规则3300条），自动标注正确率77%。

★ LLC口语语料库(London-Lund Corpus of Spoken English)

- 1960s 伦敦大学著名语言学家Quirk 组织，瑞典隆德(Lund)大学教授 Svartvik 主持录入计算机
- 87个文本，每个文本约5000词，最终规模 50万词
- 5大类：面对面交谈；电话交谈；讨论；采访；辩论，未经准备的当众评论、论证、演讲，经准备的当众演讲
- 标注：语调、节律、关键词(语段)，词类、出现次数、搭配关系等

1. 语料库

■ 典型的语料库资源

★ 朗文语料库 (Longman Corpus)

- 朗文语料库委员会 (Longman Corpus Committee)
- 选自1900～的20世纪英语：知识性(informative)文60%，想象性
- (imaginative)文本40%
- 2800 万词，10个分布广泛的领域：自然和纯科学、应用科学、社会科学、世界事务等

★ 宾州(Pennsylvania)大学树库 (UPenn Tree Bank)

(http://www.ldc.upenn.edu/)

- 美国宾州大学计算机系 M. Marcus 教授主持
- 1993年完成约300万词次英语句子的语法结构标注
- 2000年完成第一版汉语树库，约10万词次，4185个句子
- Chinese Tree Bank (CTB) 中汉语词性(part-of-speech)被划分为33类，23类句法标记(Syntactic tags)

1. 语料库

■ 典型的语料库资源

★ 布拉格依存树库 (Prague Dependency Treebank, PDT)

（http://www.elsnet.org/nps/0040.html）

由捷克布拉格查尔斯大学(Charles University in Prague) 组织开发，目前已经建成三个语料库：捷克语依存树库、捷克语－英语依存树库和阿拉伯语依存树库。有形态和句法分析层的标注工作和树库的深层语法层(tectogrammatical layer) 的信息标注

★ Europarl(欧洲议会)

http://www.statmt.org/europarl/

- 包括21种欧洲语言的版本，现在每种语言达到6000万字
- 来源：欧洲议会的会议记录，时间跨度从1996年至2011年
- 目前这个语料库还在继续扩建中

1. 语料库

■ 典型的语料库资源

★ UMBC WebBase Corpus（马里兰大学）

http://ebiquity.umbc.edu/resource/html/id/351

- 语言为英文
- 处理后，规模30亿字
- 来源：斯坦福WebBase项目2007年抓取的1亿网页

★ One Billion Word Benchmark(Google)

http://www.statmt.org/lm-benchmark/

- 数据库含有大约 10 亿英语单词，词汇有 80 万
- 来源：沃尔玛(WMT)2011年及以前的新闻数据

详细介绍：

https://github.com/tensorflow/models/tree/master/lm_1b

1. 语料库

■ 典型的语料库资源

★ 北京大学开发的CLKB

- 现代汉语语法信息词典，含8万词的360万项语法属性描述；
- 汉语短语结构规则库，含600多条语法规则；
- 标注语料库1.5亿字，其中精加工的有5200万字，标注义项2800万字；
- 平行语料库，含对译的英汉句对100万；
- 多领域术语库，有35万汉英对照术语。

★ 台湾中研院平衡语料库（Sinica Corpus)

(http://rocling.iis.sinica.edu.tw/ROCLING/corpus98/)

- 520万词次(789万汉字)汉语平衡语料库
- 语料选自1990年至1996年期间出版的哲学、艺术、科学、生活、社会和文学领域的文本
- 2003年增加了汉英平行语料库，含 2373 个汉英平行对照文本；北大现代汉语语料库，规模约为8500万汉字

1. 语料库

■ 典型的语料库资源

★ 中国中文语言资源联盟(Chinese LDC)

http://www.chineseldc.org

- 会员单位70多个
- 各类语言资源80余种
- 正式对外转让时间从2005年3月起
- 已共享资源超过133套，销售总额已经达到108万元人民币
- 授权评测单位使用超过40套

1. 语料库

■ 语料库示例

例句：北京大学计算语言所富士通人民日报标注分词语料库样例：

- 历史/n 将/d 铭记/v 这个/r 坐标/n：/w 北纬/b ４１．１/m 度/q、/w 东经/b １１４．３/m 度/q；/w
- 人们/n 将/d 铭记/v 这/r 一/m 时刻/n：/w １９９８年/t １月/t １０日/t １１时/t ５０分/t。/w
- [中国/ns 政府/n]nt 顺利/ad 恢复/v 对/p 香港/ns 行使/v 主权/n，/w 并/c 按照/p “/w 一国两制/j”/w、/w “/w 港人治港/l”/w、/w 高度/d 自治/v 的/u 方针/n 保持/v 香港/ns 的/u 繁荣/an 稳定/an。/w

1. 语料库

北京大学计算语言所语料库标记：

代码	名称	帮助记忆的诠释
Ag	形语素	形容词性语素。形容词代码为a，语素代码 g 前面置以A。
a	形容词	取英语形容词adjective的第1个字母。
ad	副形词	直接作状语的形容词。形容词代码a和副词代码d并
an	名形词	具有名词功能的形容词。形容词代码a和名词代码
b	区别词	取汉字“别”的声母。
c	连词	取英语连词conjunction的第1个字母。
Dg	副语素	副词性语素。副词代码为d，语素代码 g 前面置以
d	副词	取adverb的第2个字母，因其第1个字母已用于形容
e	叹词	取英语叹词exclamation的第1个字母。
f	方位词	取汉字“方” 的声母。
g	语素	绝大多数语素都能作为合成词的“词根”，取汉字“
h	前接成分	取英语head的第1个字母。
i	成语	取英语成语idiom的第1个字母。
j	简称略语	取汉字“简”的声母。
k	后接成分	
l	习用语	习用语尚未成为成语，有点“临时性”，取“临”的声
m	数词	取英语numeral的第3个字母，n，u已有他用。
Ng	名语素	名词性语素。名词代码为n，语素代码 g 前面置以
n	名词	取英语名词noun的第1个字母。
nr	人名	名词代码n和“人(ren)”的声母并在一起。
ns	地名	名词代码n和处所词代码s并在一起。
nt	机构团体	“团”的声母为t，名词代码n和t并在一起。
nz	其他专名	“专”的声母的第1个字母为z，名词代码n和z并在一起。
o	拟声词	取英语拟声词onomatopoeia的第1个字母。
p	介词	取英语介词prepositional的第1个字母。
q	量词	取英语quantity的第1个字母。
r	代词	取英语代词pronoun的第2个字母,因p已用于介词。
s	处所词	取英语space的第1个字母。
Tg	时语素	时间词性语素。时间词代码为t,在语素的代码g前面置以T。
t	时间词	取英语time的第1个字母。
u	助词	取英语助词auxiliary 的第2个字母,因a已用于形容词。
Vg	动语素	动词性语素。动词代码为v。在语素的代码g前面置以V。
v	动词	取英语动词verb的第一个字母。
vd	副动词	直接作状语的动词。动词和副词的代码并在一起。
vn	名动词	指具有名词功能的动词。动词和名词的代码并在一起。
w	标点符号	
x	非语素字	非语素字只是一个符号，字母x通常用于代表未知数、符号。
y	语气词	取汉字“语”的声母。
z	状态词	取汉字“状”的声母的前一个字母。

1. 语料库

■ 语料库示例

★ 宾州大学树库 (UPenn Tree Bank) (http://www.ldc.upenn.edu/)

- 约300万词次英语句子的语法结构标注
- 2000年完成第一版汉语树库，约10万词次，4185个句子
- Chinese Tree Bank (CTB) 中汉语词性(part-of-speech)被划分为33类，23类句法标记(Syntactic tags)
- 宾州树库扩展 PropBank 对原树库中的句法节点标注上特定的论元标记(argument label)，使其保持语义角色的相似性。

1. 语料库

◆ 汉语词性(part-of-speech) 标注集

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WPS	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	"	Left quote	(' or ")
POS	Possessive ending	*'s*	"	Right quote	(' or ")
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PPS	Possessive pronoun	*your, one's*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

- 33 类
- NN 名词、NR 专业名词、NT 时间名词、VA可做谓语的形容词、VC “是”、VE “有”作为主要动词、VV 其他动词、AD 副词、M 量词，等等。

1. 语料库

◆ 汉语词性及句法标注数据

例句：他还提出一系列具体措施的政策要点。

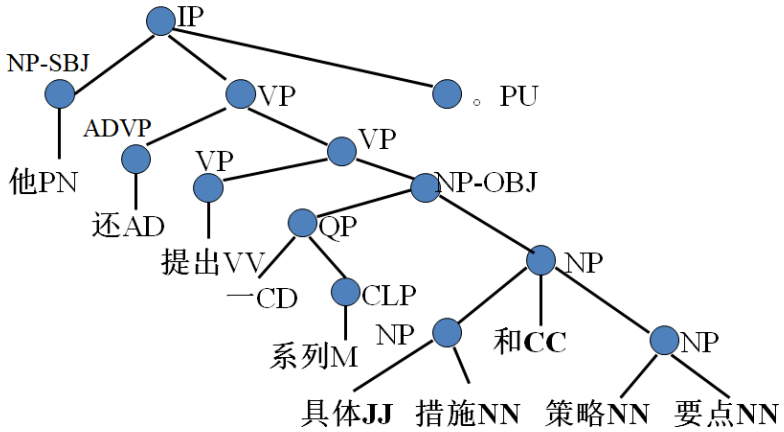
分词标注：

他/PN　还/AD　提出/VV　一/CD　系列/M　具体/JJ　措施/NN　和/CC
政策/NN　要点/NN　。/PU

句法标注：

```
( IP ( NP-SBJ (PN 他 ))
    ( VP (ADVP ( AD 还 ))
        ( VP ( VV 提出 ))
            ( NP-OBJ (QP ( CD 一)
                ( CLP ( M 系列 )))
            (NP ( NP ( ADJP ( JJ 具体)
                    ( NP (NN 措施)))
                ( CC 和)
                ( NP ( NN 政策 )
                    ( NN 要点 ))))))
    ( PU 。 ))
```

句法树：



1. 语料库

◆ 语义角色标注数据

例如1，John broke the window.

- 事件是 “打碎 (breaking event)”
- John 为事件的 制造者 (instigator)
- window为 受事者 (patient)
- 窗户被打碎 (broken window)为事件的结果

内 容 提 要

2.1　数据资源概述

2.2　统计时代语料资源

1. 语料库

2. 知识库

2.3　深度学习时代数据资源

2. 语言知识库

语言知识库：从大量的实例语料中提炼、抽象、概括出来的系统的语言知识，如电子词典、句法规则库、词法分析规则库等。

典型的知识库：

- **WordNet（http://wordnet.princeton.edu/）**
- **知识图谱**
- **常识知识库**
- **……**

2. 语言知识库

◆ **典型的知识库：**

WordNet（http://wordnet.princeton.edu/）

- 普林斯顿大学(Princeton University) 认知科学实验室 George A. Miller教授领导开发。
- 开发目的：解决词典中同义信息的组织问题
- 目前规模：95600 英语词条，其中，51500个简单词，44100 个搭配词。70100个词义(同义词集合)。
- 五大类词汇：名词、动词、形容词、副词、虚词。(实际上 WordNet 中仅包含前4类)

2. 语言知识库

- **名词的25个独立起始概念：**

{动作，行为，行动}、{自然物}、{动物，动物系}、{自然现象}、{人工物}、{人，人类}、{属性，特征}、{植物，植物系}、{身体，躯体}、{所有物}、{认知，知识}、{作用，方法}、{信息，通信}、{量，数量}、{事件}、{关系}、{直觉，情感}、{形状}、{食物}、{状态，情形}、{团体，组织}、{物质}、{场所，位置}、{时间}、{目的}

- **21000个动词词形、约8400个词义，14个文件：**

照顾动词，功能动词，变化动词，认知动词，通信动词，竞争动词，消费动词，接触动词，创作动词，感情动词，运动动词，感觉动词，占用动词，社会交往动词，天气变化动词。

- **19500个形容词词形，近10000个词义**

描述性形容词，参照修饰形容词，颜色形容词，关系形容词。

2. 语言知识库

- 特色：根据词义组织词汇信息，从某种意义上讲，它是一部语义词典。

- 4 种语义关系：
  - 同义关系 (synonymy)
  - 反义关系 (antonymy)
  - 上下位关系(hypernymy)或称从属/上属关系：如：{枫树}是{树}的下位，{树}是{植物}的下位。
  - 部分关系(meronymy)或称部分/整体关系。

2. 语言知识库

WordNet 的应用

词汇消歧，语义推理，理解等。

例如： 食堂 没 地方，我 在 饭馆 吃 了 蛋 炒饭。

“地方”的三种意思：

指地理位置　　如：在祖国各个地方

指空间　　如：没地方

指部分　　如：他说的有些地方不对

三个含义在两棵不同的名词集成语义树上：

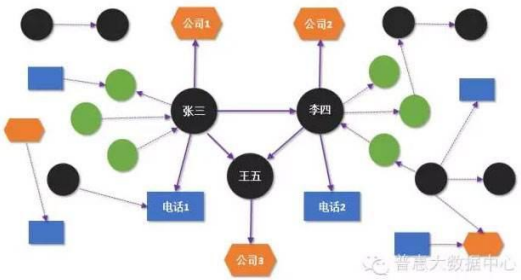
2. 语言知识库

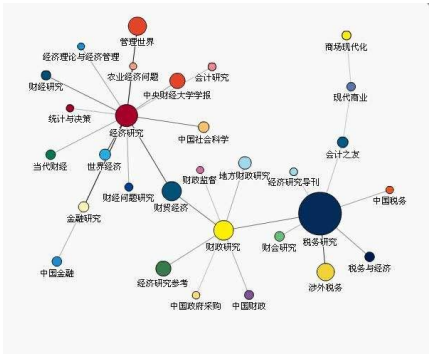
◆ 知识图谱

也称为科学知识图谱，它通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论与方法与计量学引文分析、共现分析等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构达到多学科融合目的的现代理论。为学科研究提供切实的、有价值的参考。

--- 百度百科

知识图谱本质上是一种语义网络。其结点代表实体(entity)或者概念(concept)，边代表实体/概念之间的各种语义关系；用来描述真实世界中存在的各种实体和概念，及实体、概念之间的关联关系。





2. 语言知识库

★ Freebase™

http://freebase.com

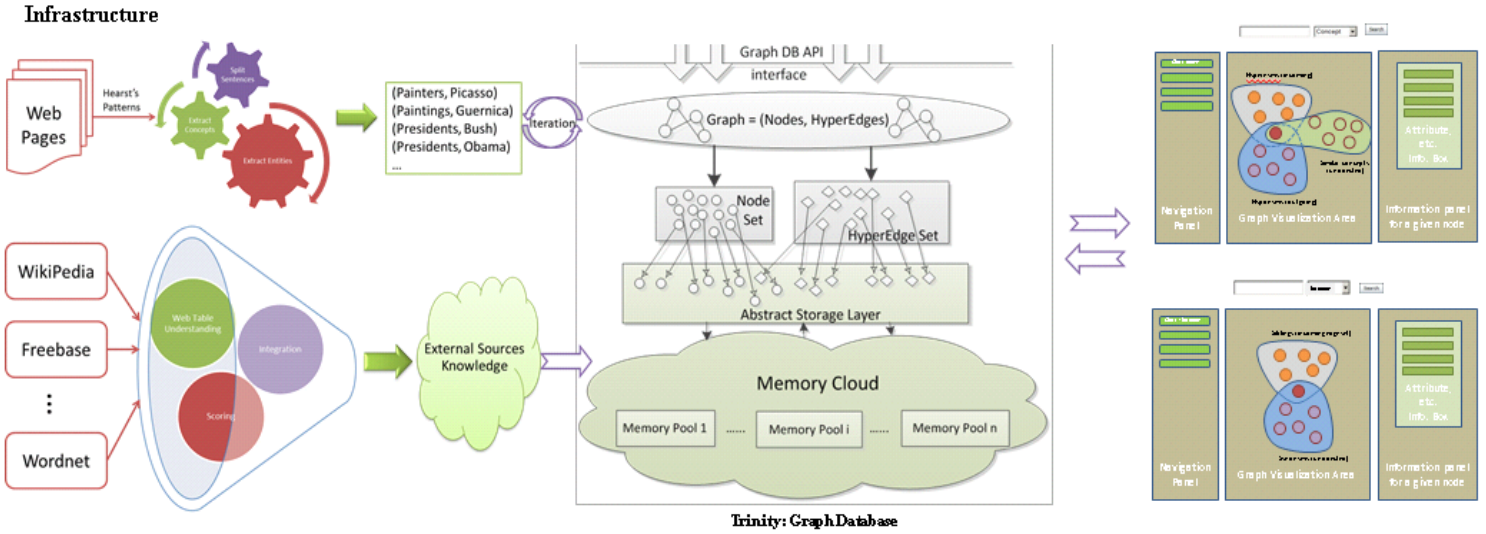
- 包含3900万个实体和18亿条实体关系
- 允许任何人创建、修改、查询的知识库，即众包模式。
- 存储的是结构化良好、机器也可读的数据格式。
- 2010年7月被Google收购。2015年，Google宣布将逐步关停Freebase，Freebase原有的数据迁移至WikiData。

2. 语言知识库

★ ProBase

http://research.microsoft.com/enus/projects/probase/

- 微软构建的知识图谱。
- 目标是使机器“意识到”人类的精神世界，使机器能更好地了解人类的沟通。
- 包含5,376,526个唯一概念，12.501.527个唯一实例和85,101,174个 IsA关系。



2. 语言知识库

★ *Microsoft Concept Graph*

https://concept.research.microsoft.com/Home/Download

- 微软亚洲研究院2016年10月27日正式发布，用于帮助机器更好地理解人类交流并且进行语义计算。
- 知识图谱包含了超过540万条概念。
- 包含的知识来自于数以亿计的网页和数年积累的搜索日志，可以为机器提供文本理解的常识性知识。

2. 语言知识库

http://dbpedia.org

- 由德国莱比锡大学等机构发起的项目，从维基百科中抽取实体关系，包括1千万个实体和14亿条实体关系。
- 数据集以多达125种不同语言表示。
- DBpedia项目使用资源描述框架（RDF）来表示提取的信息，包括30亿个RDF三元组：从维基百科的英文版提取的5.8亿和从其他语言版本提取的24.6亿。

2. 语言知识库

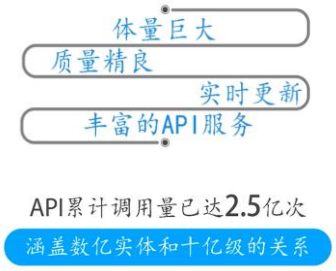
★ 中文通用百科知识图谱（CN-DBpedia）

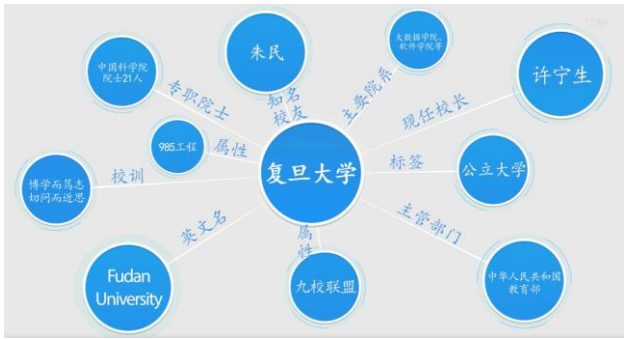
http://openkg.cn/dataset/cndbpedia

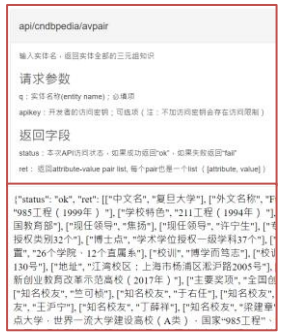
CN-DBpedia是由复旦大学知识工场实验室研发并维护的大规模通用领域结构化百科，其前身是复旦GDM中文知识图谱。主要从中文百科类网站（如百度百科、互动百科、中文维基百科等）的纯文本页面中提取信息，经过滤、融合、推断等操作后，最终形成高质量的结构化数据。

包含900万+的百科实体以及6700万+的三元组关系

提供API服务，可以直接调用接口







2. 语言知识库

★ 最大的多语言百科全书式的字典和语义网络（BabelNet)

https://babelnet.org/download

BabelNet是一个多语词汇语义网络，由罗马大学计算机科学系的计算语言学实验室所创建。和WordNet类似，BabelNet将不同语言中的词语以同义词集的形式进行组织，称之为Babel synset。

BabelNet（v3.7）包含了大约1400万个同义词集和7.46亿个词义，覆盖了271种语言，包括全部的欧洲语言、大多数亚洲语言及拉丁语。

2. 语言知识库

ConceptNet5

http://openkg.cn/dataset/conceptnet5-chinese（中文）

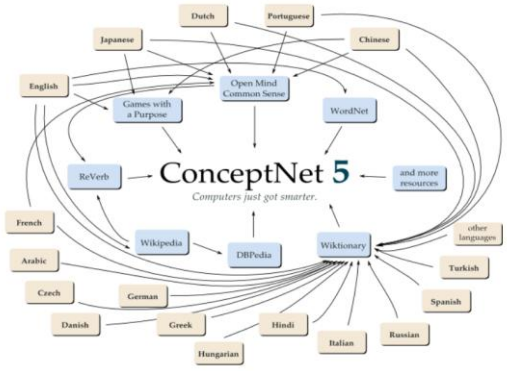
https://github.com/commonsense/conceptnet5/wiki/Downloads（英文）

ConceptNet是常识知识库，主要依靠互联网众包、专家创建和游戏三种方法来构建。

以三元组形式的关系型知识构成，包含有2800万关系描述，

可以看成是多个三元组组成的一张语义图网络。

	uri	relation	start	end	json
0	/a/[/r/Antonym/,/c/zh/上/,/c/zh/下/]	/r/Antonym	/c/zh/上	/c/zh/下	{"dataset": "/d/wiktionary/fr", "license": "cc...
1	/a/[/r/Antonym/,/c/zh/上/a/,/c/fr/suivant/]	/r/Antonym	/c/zh/上/a	/c/fr/suivant	{"dataset": "/d/wiktionary/fr", "license": "cc...
2	/a/[/r/Antonym/,/c/zh/上/a/,/c/zh/下/]	/r/Antonym	/c/zh/上/a	/c/zh/下	{"dataset": "/d/wiktionary/fr", "license": "cc...
3	/a/[/r/Antonym/,/c/zh/上/r/,/c/fr/dessous/]	/r/Antonym	/c/zh/上/r	/c/fr/dessous	{"dataset": "/d/wiktionary/fr", "license": "cc...
4	/a/[/r/Antonym/,/c/zh/上/r/,/c/zh/下/]	/r/Antonym	/c/zh/上/r	/c/zh/下	{"dataset": "/d/wiktionary/fr", "license": "cc...
...	...	...	...	...	...
624799	/a/[/r/UsedFor/,/c/zh/鼻毛夹/,/c/zh/拔獅子鼻毛/]	/r/UsedFor	/c/zh/鼻毛夹	/c/zh/拔獅子鼻毛	{"dataset": "/d/conceptnet/4/zh", "license": "...
624800	/a/[/r/UsedFor/,/c/zh/鼻涕/,/c/zh/哭/]	/r/UsedFor	/c/zh/鼻涕	/c/zh/哭	{"dataset": "/d/conceptnet/4/zh", "license": "...
624801	/a/[/r/UsedFor/,/c/zh/鼻涕/,/c/zh/痛哭/]	/r/UsedFor	/c/zh/鼻涕	/c/zh/痛哭	{"dataset": "/d/conceptnet/4/zh", "license": "...
624802	/a/[/r/UsedFor/,/c/zh/１_筆/,/c/zh/工作/]	/r/UsedFor	/c/zh/１_筆	/c/zh/工作	{"dataset": "/d/conceptnet/4/zh", "license": "...
624803	/a/[/r/UsedFor/,/c/zh/ｍｐ_３/,/c/zh/聽歌/]	/r/UsedFor	/c/zh/ｍｐ_３	/c/zh/聽歌	{"dataset": "/d/conceptnet/4/zh", "license": "...



2. 语言知识库

★ ATOMIC

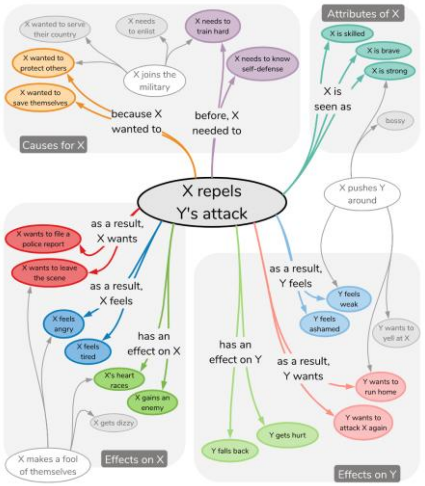
https://mosaickg.apps.allenai.org/model_comet2020_people_events

https://storage.googleapis.com/ai2-mosaic/public/atomic/v1.0/atomic_data.tgz

ATOMIC是一个日常常识推理知识图谱，关注的是以类型化的if-then关系和变量组织的推理知识。

包含超过30万个事件，涉及87.7万个推理关系。

Event	Type of relations	Inference examples	Inference dim.
"PersonX pays PersonY a compliment"	If-Event-Then-Mental-State	PersonX wanted to be nice PersonX will feel good PersonY will feel flattered	xIntent xReact oReact
	If-Event-Then-Event	PersonX will want to chat with PersonY PersonY will smile PersonY will compliment PersonX back	xWant oEffect oWant
	If-Event-Then-Persona	PersonX is flattering PersonX is caring	xAttr xAttr
"PersonX makes PersonY's coffee"	If-Event-Then-Mental-State	PersonX wanted to be helpful PersonY will be appreciative PersonY will be grateful	xIntent oReact oReact
	If-Event-Then-Event	PersonX needs to put the coffee in the filter PersonX gets thanked PersonX adds cream and sugar	xNeed xEffect xWant
	If-Event-Then-Persona	PersonX is helpful PersonX is deferential	xAttr xAttr

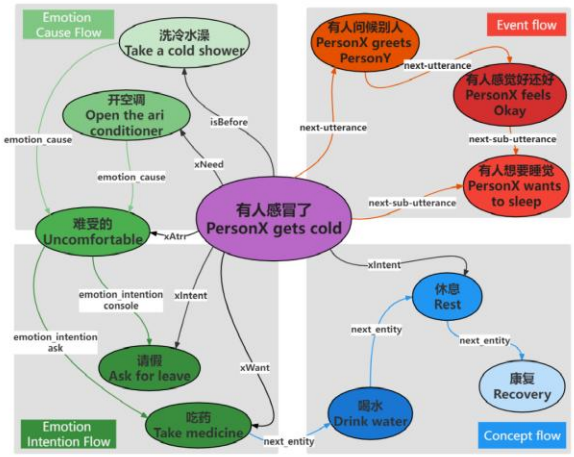


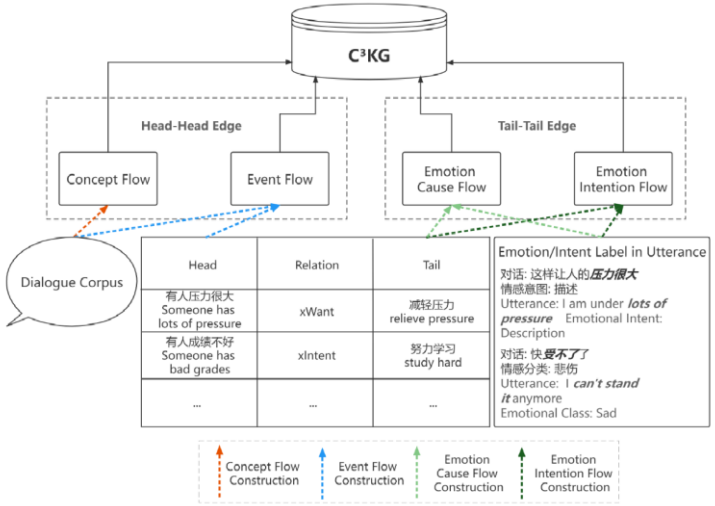
2. 语言知识库

★ C³KG

https://github.com/XiaoMi/C3KG

C³KG是一个融合常识知识和对话流信息的中文常识对话知识图谱，作者定义了4个对话流关系：事件流、概念流、情感-原因流和情感-意图流。包含超过128万个三元组。





2. 语言知识库

★ CMeKG

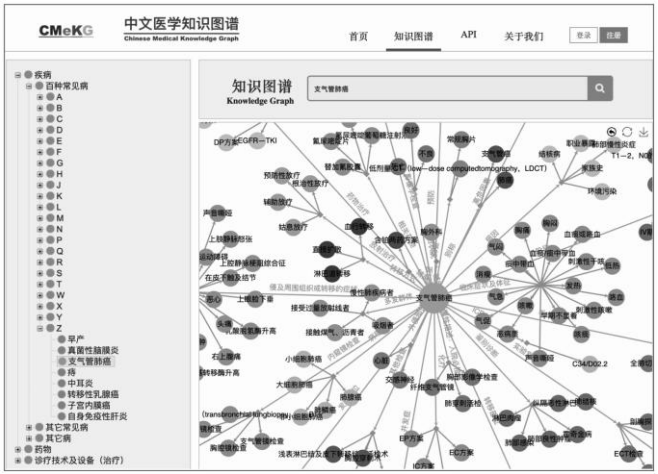
http://jcip.cipsc.org.cn/CN/abstract/abstract2840.shtml

http://www5.zzu.edu.cn/nlp/info/1018/1785.htm

CMeKG是基于大规模医学文本数据构建的中文医学知识图谱，具有广泛的医学知识覆盖面以及丰富的信息描述形式。

CMeKG2.0包含有一万多种疾病、一万多个症状、近两万多种药物和三千种诊疗手段。

关系类型	关系子类型	关系取值	样例
语义	ICD编码	字符串	C34、D02.2
	同义词、别称	字符串	肺癌，支气管肺癌，支气管癌
疾病—部位	转移部位	解剖学类	淋巴结，锁骨上淋巴结，纵隔淋巴结
	外侵部位		心脏，交感神经
	发病部位		左肺，右肺
疾病—症状	临床症状	症状类	刺激性干咳，血痰/痰中带血，咯血
	侵及周围组织转移的症状		声音嘶哑，面部水肿，颈部水肿
疾病—检查	辅助检查	诊疗技术及设备类	胸部影像学检查，纤维支气管镜，肺穿刺活检
	影像学检查		X线胸片，CT，磁共振成像
	内窥镜检查		支气管镜检查，经支气管针吸活检术
	筛查		低剂量CT，常规胸片
	实验室检查		细胞学检查，剖胸探查术，ECT检查
	其他检查		胸腔穿刺术，胸膜活检术，浅表淋巴结及皮下转移结节活检术



2. 语言知识库

知识图谱列表：

类别	名称	网址
基于维基百科	DBPedia	http://dbpedia.org
	YAGO	http://yago-knowledge.org
	Freebase	http://freebase.com
	WikiTaxonomy	http://www.hits.org/english/research/nlp/download/wikitaxono my.php
	BabelNet	http://babelnet.org
开放知识抽取	KnowItAll	http://openie.cs.washington.edu
	NELL	http://rtw.ml.cmu.edu
	Probase	http://research.microsoft.com/enus/projects/probase/
	ATOMIC	https://storage.googleapis.com/ai2-mosaic/public/atomic/v1.0/atomic_data.tgz
	CMeKG	http://jcip.cipsc.org.cn/CN/abstract/abstract2840.shtml

内 容 提 要

2.1 数据资源概述

2.2 统计时代语料资源

2.3 深度学习时代数据资源

1. 任务数据资源
2. 预训练数据资源
3. 预训练微调数据资源

1. 任务数据资源

任务数据：

在神经网络方法（第二范式）时期，各类语言处理任务主要依靠构建专门的任务神经网络模型，并通过有监督学习对模型进行训练以完成对应任务。在这一阶段，所需的数据资源主要用于训练任务模型，因此使用的数据集主要是针对具体任务进行标注的任务数据集。但由于任务参数知识有限，在需要世界知识的任务上往往表现不佳，因此在第二范式时期常常会引入知识图谱、常识图谱等外部资源。

1. 任务数据资源

◆ 文本分类任务数据集

★ 20 Newsgroups

- 约 18000 篇来自 20 个不同新闻组的新闻文章
- 每个新闻组代表一个特定的主题类型，如体育、计算机技术、医学、政治等

★ AG News

- 由 AG 语料库中四个最大类型（世界、体育、商业、科技）的文章标题和描述字段构建而成
- 每个类包含 30000 个训练样本和 1900 个测试样本

Text:

A fair number of brave souls who upgraded their SI clock oscillator have shared their experiences for this poll. Please send a brief message detailing your experiences with the procedure. Top speed attained, CPU rated speed, add on cards and adapters, heat sinks, hour of usage per day, floppy disk functionality with 800 and 1.4 m floppies are especially requested. I will be summarizing in the next two days, so please add to the network knowledge base if you have done the clock upgrade and haven't answered this poll. Thanks.

Label:

comp.sys.mac.hardware

1. 任务数据资源

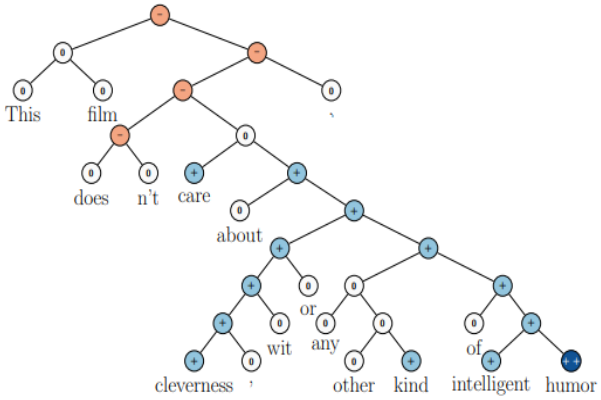
◆ 情感分析任务数据集

★ 亚马逊评论数据集 (Amazon Review Data)

- 来自亚马逊平台的用户商品评价及其对应的情感评分
- 包含数百万条商品评论，评论对象涵盖了从电子产品到日用商品的各类产品
- 评分系统通常采用 1～5 星的评分方式

★ 斯坦福情绪树库 (Stanford Sentiment Treebank)

- 包含 11,855 条电影评论单句
- 每条句子通过斯坦福解析器生成 215154 个短语，并对每个短语标注五类情感：负面（--）、略微负面（-）、中性（0）、略微正面（+）和正面（++）
- SST-5 是包含五种标签的细粒度分类数据集
- SST-2 是只包含正负两种标签的二分类数据集



整个句子的情感评分是通过自底向上、逐步聚合的方式计算得出

1. 任务数据资源

◆ 机器翻译任务数据集

★ 开放字幕数据集 (OpenSubtitles Corpus)

- 收录了来自电影和电视节目的字幕，涵盖丰富的日常对话和口语表达
- 共包含 1689 个双语文本，覆盖 60 种语言，总计 26 亿个句子

★ 联合国平行语料库 (United Nations Parallel Corpus)

- 包含联合国会议的官方文件及其多语言翻译，覆盖了包括英语、法语、西班牙语、中文、俄语和阿拉伯语在内的多种语言
- 涵盖政治、法律、国际事务等领域的专有术语和复杂表达
- 文本数据具有高质量、结构化的特点

English:

You have been threatening to tell me the truth about my husband since the day we met, and I'm ready to listen.

Spanish:

comp.sys.mac.hardware

1. 任务数据资源

◆ 文本摘要任务数据集

★ CNN 每日邮报数据集 (CNN / DailyMail Dataset)

- 包含来自 CNN 和每日邮报新闻网站的文章及其对应的摘要
- 有 286817个训练对、13368 个验证对和 11487 个测试对
- 源文档平均有 766个单词和 29.74 个句子，摘要平均有 53 个单词和 3.72 个句子

★ Reddit TIFU 数据集 (Reddit TIFU dataset)

- 基于 Reddit 论坛的 TIFU 子板块
- 包含用户发布的内容以及对应的简短总结
- 提供了较为真实的非正式语体文本，尤其适用于短文本和噪声较高的社交数据环境

Article:

LONDON, England (Reuters) -- Harry Potter star Daniel Radcliffe gains access to a reported £20 million ($41.1 million) fortune as he turns 18 on Monday, but he insists the money won't cast a spell on him. Daniel Radcliffe as Harry Potter in "Harry Potter and the Order of the Phoenix" …

Highlights:

Harry Potter star Daniel Radcliffe gets £20M fortune as he turns 18 Monday . Young actor says he has no plans to fritter his cash away . Radcliffe's earnings from first five Potter films have been held in trust fund .

1. 任务数据资源

◆ 机器阅读理解任务数据集

★ 斯坦福问答数据集 (Stanford Question Answering Dataset (SQuAD))

- 包含来自维基百科的文章及其对应的自然语言问题，每个问题的答案均为文章中的一段文本。
- SQuAD1.1：包含 500 多篇文章的 10万多个问答对
- SQuAD2.0：整合了 SQuAD1.1 中的 10 万个问题，同时加入了 5 万多 “不可回答” 问题

文本	从建筑风格上看，这所学校具有天主教特色。主楼金色圆顶的顶端矗立着圣母玛利亚的金色雕像。主楼正前方、面向主楼的位置，有一尊双臂高举的基督铜像，底座刻有铭文“Venite Ad Me Omnes”（“你们都到我这里来”）。在主楼旁边是圣心大教堂，而在大教堂的正后方是石窟，这是一个供人祈祷与默想的圣母场所，它仿照法国卢尔德的石窟建造，圣母玛利亚据说曾于1858年在那里显现给圣女伯尔纳德特·苏比鲁斯。在主干道的尽头（与三座雕像和金色圆顶排成一条直线）则是一尊简洁的现代石雕圣母像。
问题	1858年，圣母玛利亚据称在法国卢尔德向谁显现?
答案	圣女伯尔纳德特·苏比鲁斯

1. 任务数据资源

◆ 机器阅读理解任务数据集

★ DuReader

- 2017 年百度发布的大规模中文阅读理解数据集
- 包含大量以往较少研究的是非和观点类问题，每个问题对应多个答案
- 总计20 万问题、100 万原文和 42 万答案

问题	智齿一定要拔吗
问题类型	YesNo-类型
答案 1	[Yes]因为智齿很难清洁的原因，比一般的牙齿容易出现口腔问题，所以医生会建议拔掉
答案 2	[Depend]智齿不一定非得拔掉，一般只拔出有症状表现的智齿，比如说经常引起发炎...
文档 1	为什么要拔智齿?智齿好好的医生为什么要建议我拔掉?主要还是因为智齿很难清洁...
…	…
文档 5	根据我多年的临床经验来说,智齿不一定非得拔掉.智齿阻生分好多种...

1. 任务数据资源

◆ 问答任务数据集

★ 维基问答语料库 (WikiQA)

- 问题源自必应查询日志，可反映用户的真实信息需求
- 每个问题关联到一个可能包含答案的维基百科页面，页面的摘要部分作为候选答案

★ Hotpot 问答数据集 (HotpotQA)

- 针对复杂问答任务设计的数据集，包含多跳推理和跨文档问答，要求模型能够综合多个信息源来回答问题
- 总计约 11 万个有效样本，其中包含约 2 万个单跳训练样本、约 7 万个多跳训练样本，以及约 2 万个验证和测试样本

1. 任务数据资源

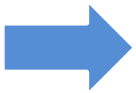
◆ 问答任务数据集

Hotpot 问答数据集示例

《七堂极简物理课》是由一位自 2000 年起在法国工作的意大利物理学家撰写的。这位物理学家从哪一年开始在法国工作?

文档 1: Guido Caldarelli
Guido Caldarelli（1967 年 4 月 8 日生于罗马）是一位……
文档2: Aldo Pontremoli
Aldo Pontremoli（1896 年 1 月 19 日 – 1928 年 5 月 25 日）是一位意大利物理学家，曾担任……
文档 ...
文档 6: Carlo Rovelli
Carlo Rovelli（1956 年 5 月 3 日出生）是一位意大利……
文档 ...
文档 10: 《七堂极简物理课》
《七堂极简物理课》是由意大利物理学家 Carlo Rovelli 撰写的一本书。该书最初于 2014 年以意大利语出版，后来被翻译成 41 种语言。

文档10: 《七堂极简物理课》
《七堂极简物理课》是意大利物理学家 Carlo Rovelli 撰写的一本书……



文档6: Carlo Rovelli
Carlo Rovelli……曾在意大利和美国工作，自 2000 年起在法国工作……

内 容 提 要

2.1 数据资源概述

2.2 统计时代语料资源

2.3 深度学习时代数据资源

1. 任务数据资源
2. 预训练数据资源
3. 预训练微调数据资源

2. 预训练数据资源

预训练数据：

在预训练语言模型+精调（第三范式）时期，模型训练采用“预训练+精调”的模式，即由预训练语言模型参数和下游任务参数共同构成。在这一阶段，需要大量的预训练数据对预训练模型进行训练，同时需要少量的标注任务数据对模型的任务参数进行微调。其中，所需的任务标注数据集可复用第二范式的任务数据。

注：第四范式所用数据与第三范式基本相同，区别是第四范式微调不是调任务模型参数，而是微调预训练语言模型

2. 预训练数据资源

◆ **网页数据：** 网页数据是预训练语料中最常见和使用最广泛的数据资源，包含大量的互联网网页文本信息。为大语言模型学习真实场景的语言规律提供了重要的数据资源。

★ **Common Crawl （英文为主）：**

https://commoncrawl.org/

https://github.com/karust/gogetcrawl

➢ 一个从 2008 年至今一直在定期更新的庞大的非结构化的多语言网页数据集，包含原始网页数据、元数据和提取的文本数据等，总数据量达到 PB 级别

2. 预训练数据资源

★ **RefinedWeb （英文为主）**

- 在 Common Crawl 数据集的基础上通过筛选和去重得到
- 使用的源数据是从 2008 年到 2023 年 6 月的所有 Common Crawl 网页记录，共约 5×10^12 个词元的高质量英文文本
- 开源部分有 6×10^11 个词元，数据量约 500GB

★ **C4 （英文为主）**

- 基于 2019 年 4 月的 Common Crawl 数据集，经过多重过滤处理，去除了无用、有害以及非英文文本
- 包含超过 1.56×10^11 个词元，数据量约 800GB

★ **ChineseWebText （中文为主）**

- 基于 Common Crawl 精心筛选的中文数据集
- 该数据集汇集了 2021～2023 年间的网页快照，总计 1.42TB 数据
- 每篇文本都附有定量的质量评分

2. 预训练数据资源

◆ **书籍：** 书籍数据也是预训练常用的数据类型之一。与网页相比，书籍具有更长的文本，蕴含更多的语言篇章信息，能帮助语言模型学习语言的长程依赖关系。包括小说、传记、教科书等。

★ **BookCorpusOpen**

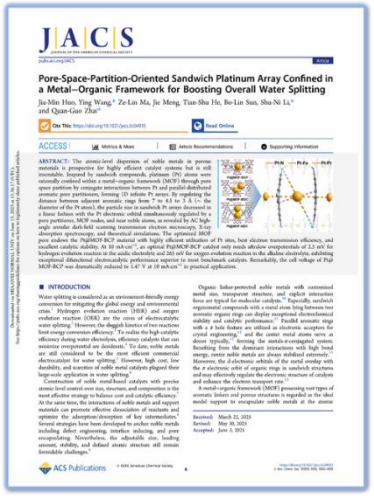
- 由多伦多大学创建的免费书籍数据集
- 包含17 868 本图书，涵盖16 种不同的主题

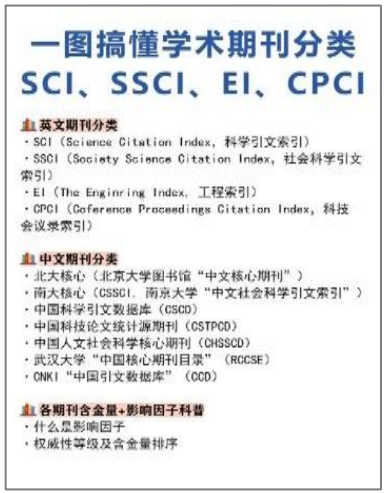
★ **Project Gutenberg**

- 最早的数字图书馆，目前还在持续更新中
- 收录西方文学作品，包括小说、诗歌、戏剧等。作品以英语为主，但也涵盖法语、德语等多种语言，在其官方网站上可免费下载

2. 预训练数据资源

◆ **学术资料：**包括但不限于学术论文、期刊文章、会议论文、研究报告、专利等。具有高度的专业性和学术严谨性。可以提供更准确的专业知识信息，帮助模型理解学术领域中的术语和专业知识。





2. 预训练数据资源

★ arXiv 数据集(arXiv Dataset)

- 由收录了物理学、数学、计算机科学、生物学和经济学等众多领域预印本论文的集合
- 发布了机器可读的 arXiv 论文数据集，共包含约 1.7× 106 篇文章，每篇文章都包含文本、图表、作者、引文、分类以及其他元数据等信息，总数据量约 1.1TB

★ S2ORC 数据集 (The Semantic Scholar Open Research Corpus)

- 由语义学术论文构成
- 通过清洗、过滤并转换为适合预训练的文本格式，共包含 1.36×10^8 篇论文

2. 预训练数据资源

◆ **维基百科：**维基百科涵盖了历史、科学、文化艺术等多个领域。其特点是自由内容、自由编辑，其支持多种语言并不断更新定期发布其数据库副本。

★ **Wikipedia**

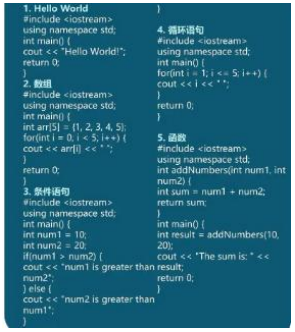
https://huggingface.co/datasets/legacy-datasets/wikipedia

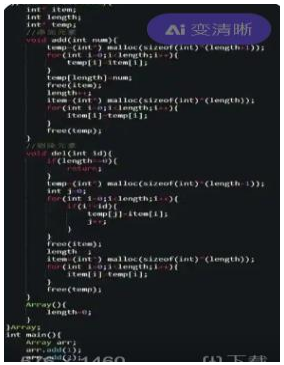
https://github.com/noanabeshima/wikipedia-downloader

维基百科数据集是基于Wikipedia数据转储构建的。这份数据集涵盖了20种不同的语言，每个示例都包含一篇完整的维基百科文章的内容。数据处理过程中，已经清理了不需要的部分，如参考文献、脚注等，以确保数据的纯净性和可用性。

2. 预训练数据资源

- **代码：** 指用编程语言编写的程序代码，具有高度结构化与专业性的特点。对于预训练语言模型，引入代码的数据进行训练不仅有助于提高模型的编程能力， 还可以增强模型的结构化推理能力， 提升模型理解和生成编程语言的能力。





2. 预训练数据资源

★ BigQuery

- 谷歌发布的企业数据仓库，是包含社交、经济、医疗、代码等众多领域的公共数据集
- 代码类数据重点收录了六种精选编程语言数据，可为预训练语言模型提供高质量的代码语料

★ The Stack

- The Stack数据集是BigCode项目的一部分,该数据集包含6TB的合法开源代码文件，覆盖了30种编程语言。
- 这些代码文件是从公开存档的GitHub仓库中收集的，并经过筛选以满足目标扩展和许可要求。

2. 预训练数据资源

◆ **混合型数据集:**为了便于使用，很多研究机构对于多种来源的数据集合进行了混合，发布了一系列包括多来源的文本数据集合。这些混合数据集往往融合了新闻、社交媒体内容、维基百科条目等各种类型的文本，减少了重复清洗数据、选择数据的繁重工程。

★ **悟道数据集 (WuDaoCorpora)**

- 由北京智源人工智能研究院构建，包括文本数据集、多模态图文数据集和中文对话数据集,覆盖教育、科技等 50 多个行业的数据标签
- 经清洗和去隐私后，数据量为 5TB，其中包含 200GB 的开源数据集

★ **书生·万卷 1.0 多模态预训练语料 (WanJuan)**

- 上海人工智能实验室发布,由多种不同来源数据组成，包括网页、书籍等
- 包含约 5×10^8 个文档，已经过数据格式统一和细粒度的清洗，总数据量超过 1TB

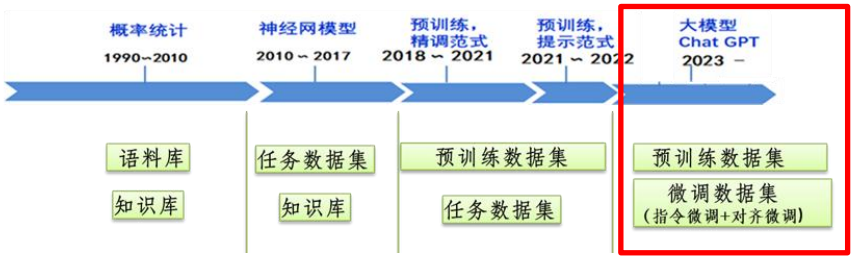
内 容 提 要

2.1 数据资源概述

2.2 统计时代语料资源

2.3 深度学习时代数据资源

1. 任务数据资源
2. 预训练数据资源
3. 预训练微调数据资源



3. 预训练微调数据资源

预训练微调数据：

在预大语言模型（第五范式）时期，大模型训练采用“预训练+后训练(微调)”模式，即先通过预训练方式对大语言模型进行预训练，然后采用微调的方式让模型给出符合人类认知的输出，其中微调阶段又分为指令微调和对齐微调，指令微调主要让模型按人类的特定习惯给出输出，对齐微调主要是让模型输出的内容符合人类价值观和偏好。

在这一过程中，预训练阶段需要大量的预训练数据对模型进行训练（数据集可用前节介绍的“预训练数据资源”）；微调阶段指令微调需要用标注的指令微调数据集进行微调，对齐阶段需要对齐数据集进行微调。

3. 预训练微调数据资源

◆ **指令微调数据集:** 该类数据集是有标注数据集，主要对大语言模型进行符合人类习惯输出的微调训练。一般有人工标注数据集和用模型合成方法生成的数据集

★ **Dolly（人工标注）**

- 由 Databricks 公司发布
- 包含 15000 个人工标注的数据实例
- 主题涉及 InstructGPT 论文中提到的 7 个领域，包括头脑风暴、分类、封闭式质量保证、生成、信息抽取、开放式质量保证和总结等

★ **OpenAssistant（人工标注）**

- 人工创建的多语言对话语料库，共有 91 829条用户提示、69614 条助手回复
- 包含 35 种语言的语料，每条语料基本都附有人工标注的质量评级

3. 预训练微调数据资源

★ **Self-Instruct-52K（合成）**

- 人工收集创建 175 个种子任务
- 利用 GPT-3 生成 52000条指令条指令和 82000个实例数据

★ **Alpaca-52K （合成）**

- 基于Self-Instruct-52K的175 个种子任务
- 利用 OpenAI text-davinci-003 模型生成 52000个不重复的指令
- 每一条指令仅生成一个实例

3. 预训练微调数据资源

★ 其它指令微调数据集

类别	集合	时间	# 样本数量	来源
任务	Nat. Inst.	2021 年 04 月	193K	Allen Institute for AI
	FLAN	2021 年 09 月	4.4M	Google
	P3	2021 年 10 月	12.1M	BigScience
	Super Nat. Inst.	2022 年 04 月	5M	Allen Institute for AI
	MVPCorpus	2022 年 06 月	41M	Renmin University of China
	xP3	2022 年 11 月	81M	BigScience
	OIG	2023 年 03 月	43M	LAION-AI
	UnifedSKG	2022 年 03 月	812K	The University of Hong Kong
对话	HH-RLHF	2022 年 04 月	160K	Anthropic
	HC3	2023 年 01 月	87K	SimpleAI
	ShareGPT	2023 年 03 月	90K	TechCrunch
	Dolly	2023 年 04 月	15K	Databricks
	OpenAssistant	2023 年 04 月	161K	LAION-AI
	InstructWild v2	2023 年 04 月	111K	National University of Singapore
	LIMA	2023 年 06 月	1K	Meta AI
合成	Self-Instruct	2022 年 12 月	82K	University of Washington
	Alpaca	2023 年 03 月	52K	Standford
	Guanaco	2023 年 03 月	535K	-
	Baize	2023 年 04 月	158K	University of California, San Diego
	Belle	2023 年 04 月	1.5M	LianjiaTech
	Alpaca-GPT4	2023 年 04 月	52K	Microsoft
	Evol-Instruct	2023 年 06 月	52K	Microsoft
	UltraChat	2023 年 06 月	675K	Tsinghua University

3. 预训练微调数据资源

◆ **人类对齐数据集:** 除了指令微调之外，将大语言模型与人类价值观和偏好对齐也非常重要。现有的对齐目标主要聚焦于三个方面：有用性、诚实性和无害性，自针对上述对齐目标数据集一般进行了标注，

★ **HH-RLHF**

（https://huggingface.co/datasets/Anthropic/hh-rlhf）

- 包含两类标注数据，大语言模型的有用性和无害性
- 整个数据集共包含约 1.69×10^5个开放式对话，每个对话信息助手将会为每个用户查询提供两个回答，若一个回答被选择则另一个回答被拒绝。

3. 预训练微调数据资源

★ Math-Step-DPO-10K

- 为提升大语言模型在多步数学推理任务中的表现而构建的高质量数据集
- 包含10,795 条推理偏好对

★ 其它人类对齐数据集

数据集	时间	# 样本数量	来源	对齐目标
Summarize from Feedback	2020 年 09 月	193K	OpenAI	有用性
SHP	2021 年 10 月	385K	Standfordnlp	有用性
WebGPT Comparisons	2021 年 12 月	19K	OpenAI	有用性
Stack Exchange Preferences	2021 年 12 月	10M	HuggingFaceH4	有用性
HH-RLHF	2022 年 04 月	169K	Anthropic	有用性、无害性
Sandbox Alignment Data	2023 年 05 月	169K	Google	有用性、诚实性、无害性
CValues	2023 年 07 月	145K	Alibaba	无害性
PKU-SafeRLHF	2023 年 10 月	330K	PKU-Alignment	有用性、无害性

开源社区资源

◆ **资源社区：**除了数据资源外，还有大语言模型相关代码库开源社区，为用户提供高效、易用且可重复的自然语言处理技术解决方案。

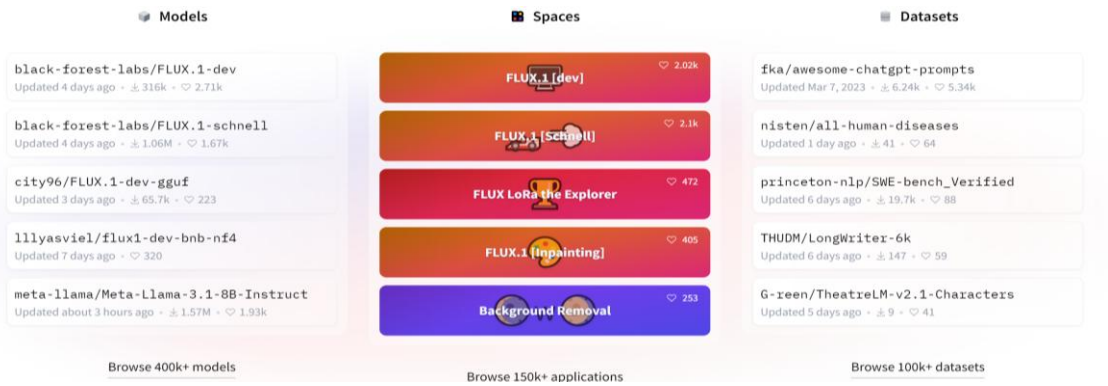
★ **Huggingface**

https://huggingface.co/

- Hugging Face 是一个广泛应用于人工智能和机器学习领域的平台，提供丰富的模型、数据集以及工具，特别是针对自然语言处理（NLP）任务。Hugging Face拥有超过12万个模型和2万个数据集。

开源社区资源

- Hugging Face 提供的主要资源包括：
- Model库：提供了大量模型资源，如预训练模型BERT、GPT等，以及大语言模型Lllama、Bloom等。
- Dataset库：用于构建、共享和读取各种NLP数据集，支持文本分类、问答、翻译等多种任务的数据集。
- Tokenizer库：用于对输入文本进行分词处理，是进行NLP任务的基础。
- Spaces：托管无限数量使用机器学习工具和模型构建的应用。

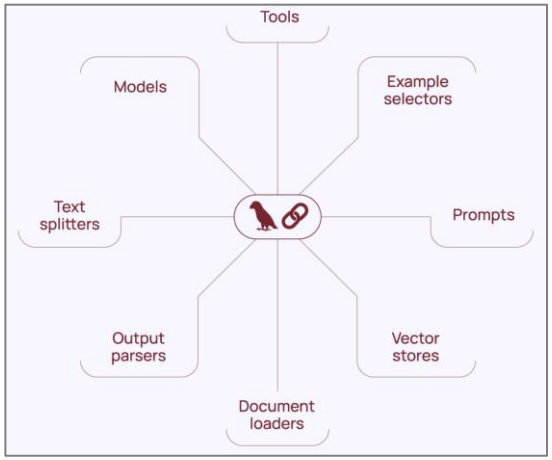


开源社区资源

★ LangChain

https://www.langchain.com/

LangChain是一个开源框架，提供了构建基于LLM的AI应用所需的各种模块和工具，旨在简化使用大型语言模型（LLMs）创建应用程序的过程。



开源社区资源

Langchain有6大核心模块：

- Models：模型，是各种类型的模型和模型集成。
- Prompts：提示，包括提示管理、提示优化和提示序列化。
- Memory：记忆，用来保存和模型交互时的上下文状态。
- Indexes：索引，用来结构化文档，以便和模型交互。包括文档加载程序、向量存储器、文本分割器和检索器等。
- Agents：代理，决定模型采取哪些行动，执行并且观察流程，直到完成为止。
- Chains：链，一系列对各种组件的调用。

使用Langchain中不同组件的特性和能力，可以构建不同场景下的应用，如聊天机器人、基于文档的问答、知识管理、个人助理、Agent智能体等等。

参考文献：

宗成庆，统计自然语言处理（第2版）课件

赵鑫 李军毅 等， 大语言模型 ， Copyright © RUC AI Box

https://www.langchain.com/

https://huggingface.co/

在 此 表 示 感 谢!

中国科学院大学
University of Chinese Academy of Sciences

谢 谢！

Thank you