



# 自然语言处理与行业应用

演讲人：一览群智 刘占亮



# 目录

- 01 自然语言处理简介
- 02 对AI的期望
- 03 NLP发展现状
- 04 我们在做什么

## 什么是自然语言

- 语言是某个符号系统上按照一定规律构成的句子和符号串的有限或无限的集合。
- 自然语言是指汉语、英语、法语等人们日常使用的语言，是自然而然的随着人类社会发展演变而来的语言。



## ► 什么是形式语言

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0}$$

$$\nabla \cdot \mathbf{B} = 0$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

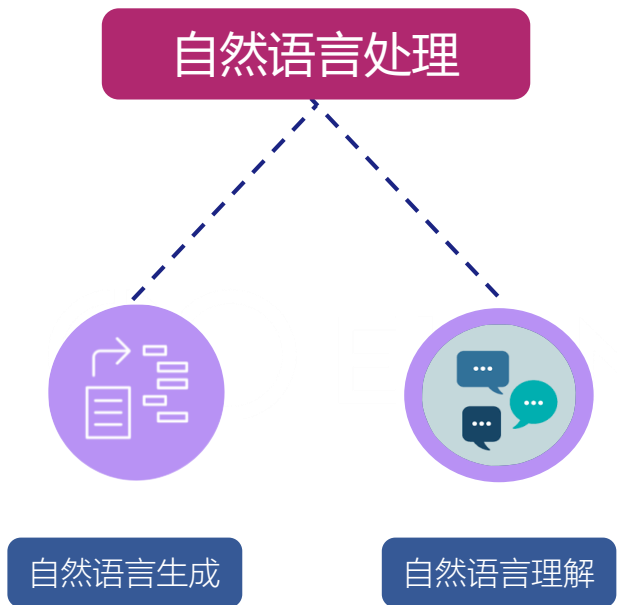
$$\nabla \times \mathbf{B} = \mu_0 \left( \mathbf{J} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right)$$

```
368 int lcd_create_map_value_to_empty(int x, int y)
369 {
370     memset(empty, 0, 8);
371     int i = 0;
372     int tmp;
373
374     tmp = percent1 / 10;
375     printf("percent1 = %d, tmp = %d\n", percent1, tmp);
376     for(i = 7; i >= 0; i--)
```

```
nf.math.sub(
  nf.math.sum(
    nf.math.pow(
      nf.math.sum(
        5,
        nf.math.to_number(
          "5.8"
        )
      ),
      2
    ),
    nf.math.to_number("1.5")
  ),
  2
),
nf.math.to_number("261.712")
);
```

与自然语言不同的是，**形式语言**（Formal Language）是为了特定应用而人为设计的语言。例如数学家用的数字和运算符号、化学家用的分子式等。编程语言也是一种形式语言，是专门设计用来表达计算过程的形式语言。

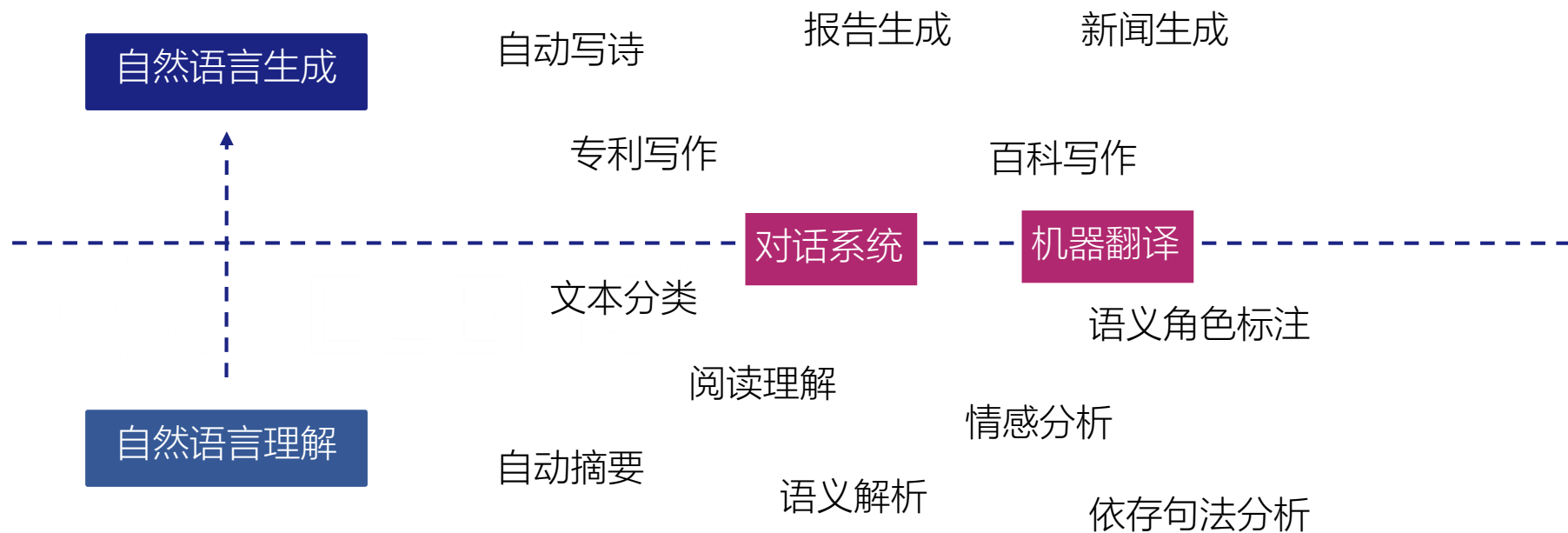
## ► 什么是自然语言处理



人类知识有80%都是由自然语言承载的。与包括编程语言在内的形式语言不同，理解自然语言需要关于外在世界的广泛知识以及运用操作这些知识的能力。

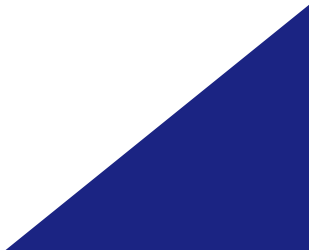
所以自然语言处理是一个人工智能完备(AI-complete)的问题，并被视为人工智能的核心问题之一。

## 自然语言处理任务





# 目录

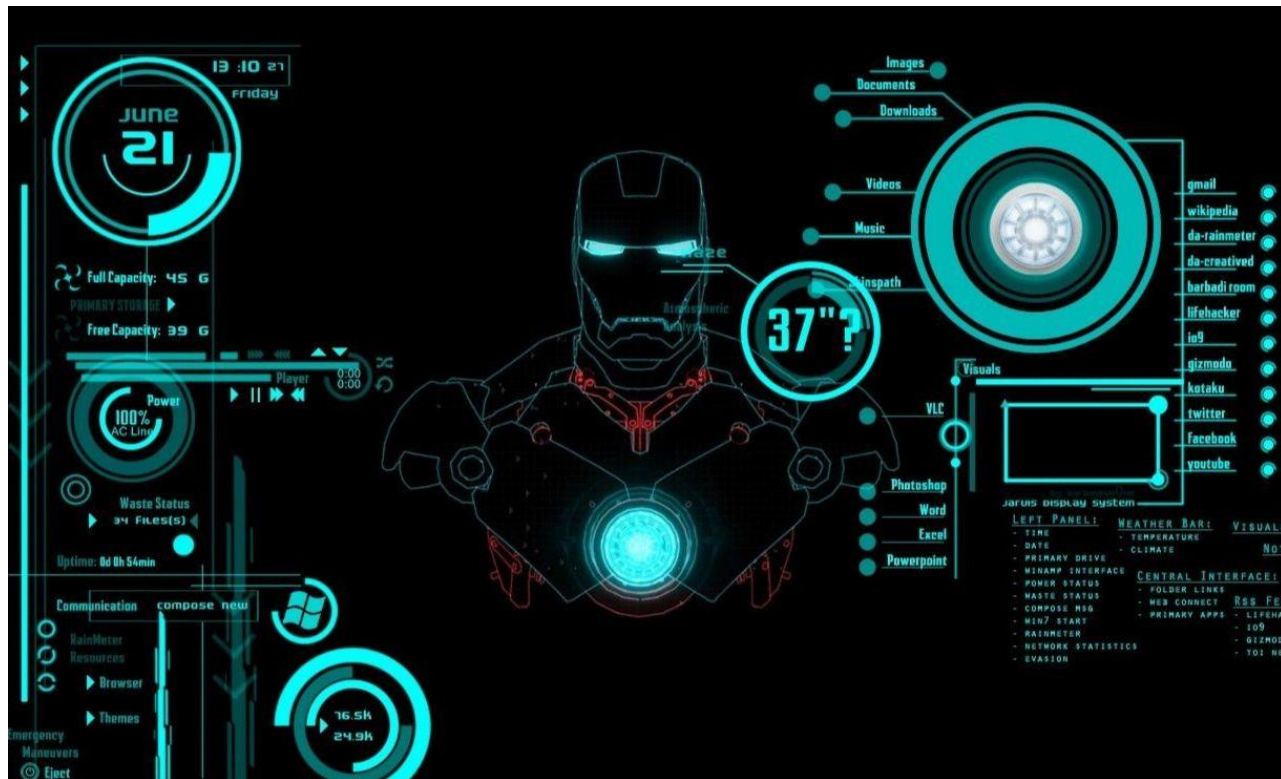
- 01 什么是自然语言处理
  - 02 对AI的期望
  - 03 NLP发展现状
  - 04 我们能做什么
- 



星球大战：1970s ~

- ▶ C-3PO：精通六百万种沟通方式，懂得各地风俗的金色机器人
- ▶ R2-D2：太空船技工和电脑接口专家，一个典型的机智、勇敢、而又鲁莽的机器人

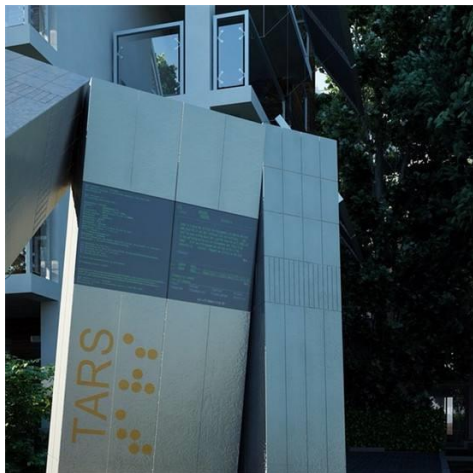




## ► 钢铁侠 2008 ~

贾维斯是《钢铁侠》中的智能管家，超智能软件，能独立思考，会帮助主人处理各种事务，计算各种信息

## ▶ 还有



TARS: 《星际穿越》



哈尔9000: 《2001太空漫游》



MOSS: 《流浪地球》

# 善解人意，无所不能

## ▶但现状是

小孩：小音箱，播放《笑起来真好看》

音箱：为你播放《balabala》

小孩：小音箱，播放错了！

音箱：好的，即将为你播放代梓琪的《错了》

.....

### 触发词：

小爱同学

天猫精灵

小度，小度

Hey, Siri

OK, Google

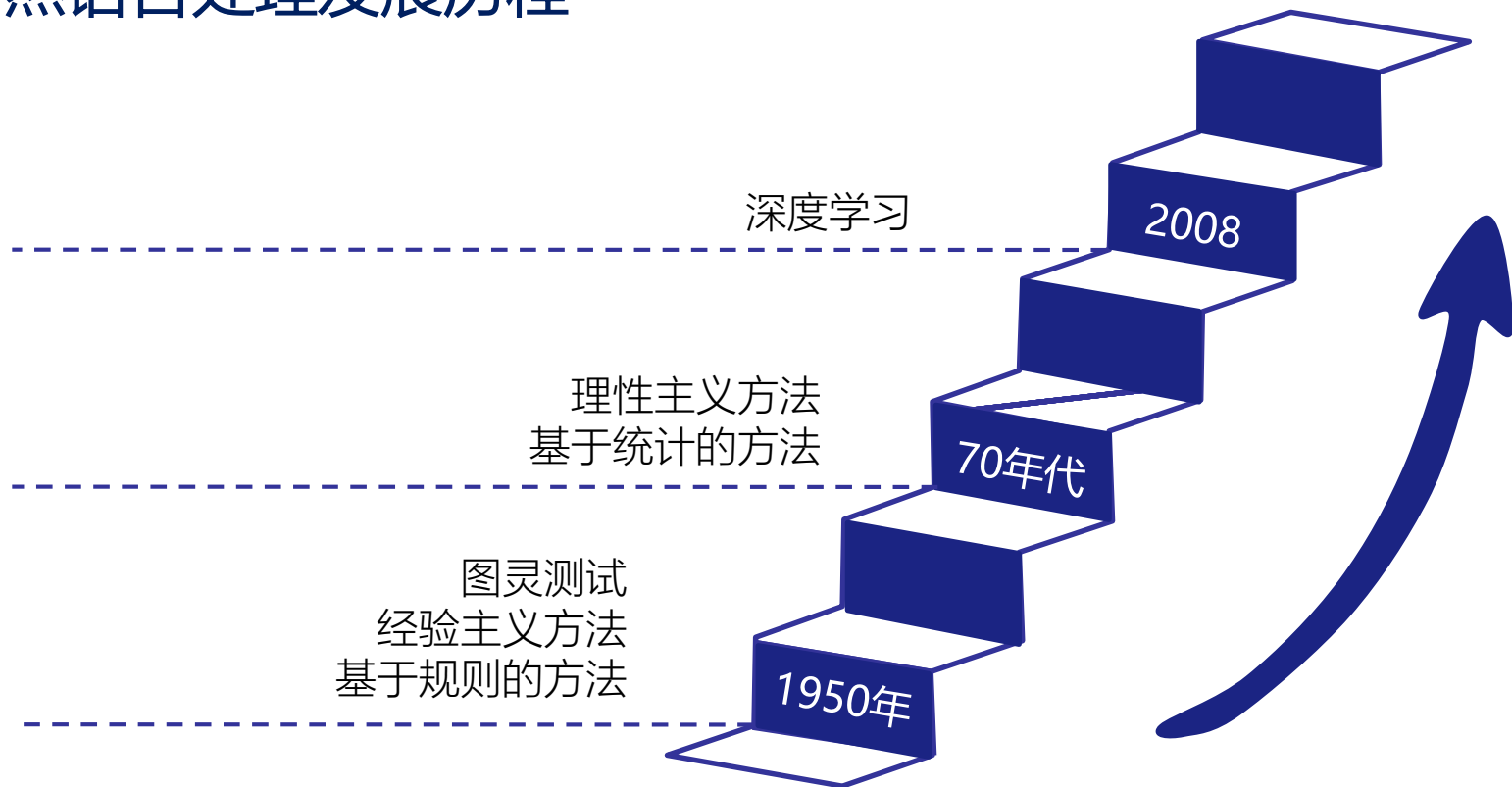
.....




# 目录

- 01 自然语言处理简介
  - 02 对AI的期望
  - 03 NLP发展现状
  - 04 我们能做什么
- 

## ► 自然语言处理发展历程





2013	Word embeddings	Word2Vec
2013	Neural networks for NLP	RNN LSTM CNN
2014	Seq2Seq Models	Machine Translation, Structure prediction
2015	Attention	Transformer(2017)
2015	Memory-based neural network	Neural Turing Machine
2018	Pretrained Language Models	ELMo, BERT.....
2019	Natural Language Generation, Reasoning, Bigger models (T5) .....	



## Can Artificial Intelligence replace the 35,000 humans at Facebook?

**This AI text generator can replace human writers. The semi-coherent ones, at least.**

**AI will soon surpass human intellect in every way: Elon Musk at WAIC**

Elon Musk's AI warning:  
Artificial Intelligence is a  
'potential danger to the public'

## ► 越来越大的模型

- 数小时训练
- 小规模语料
- 参数少
- 本地训练

- US\$ 6,912
- 大规模语料
- 亿级参数
- 4 天训练
- TPU

- US\$ 256 一小时
- 40GB语料
- 十五亿参数
- 长时间训练
- 大量TPU / GPU

- US\$ 245,000
- 大规模语料
- 十亿级参数
- 2.5 天训练
- TPU

- US\$ 百万?
- 750GB语料
- 百亿级参数
- TPU

早期

BERT

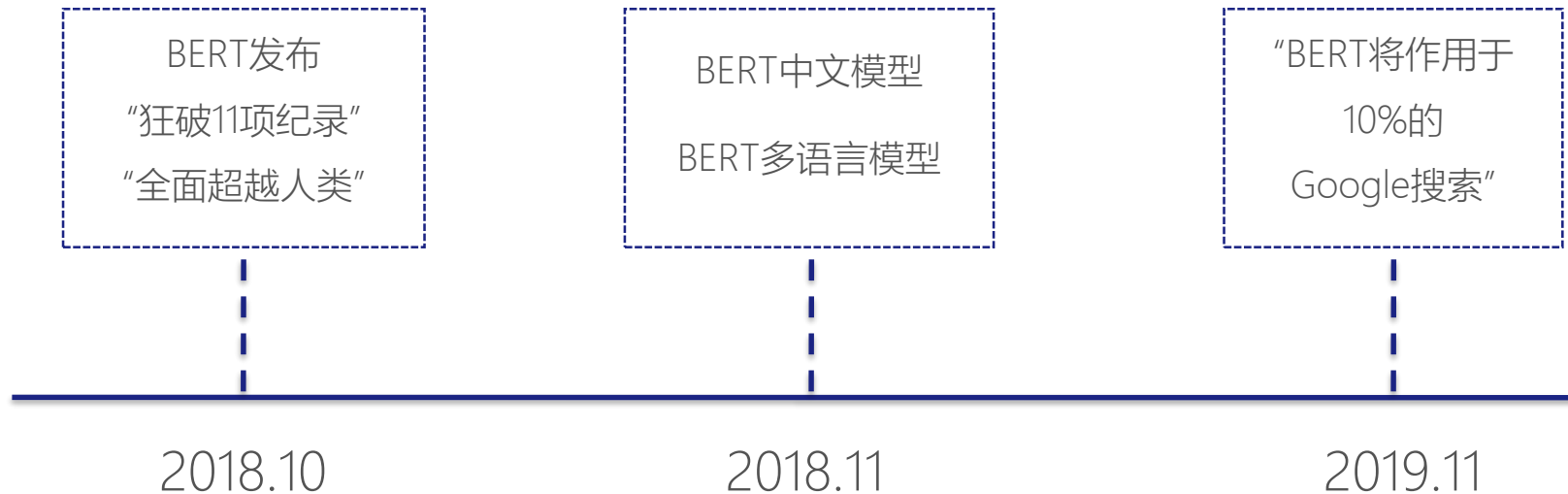
GPT2

XLNET

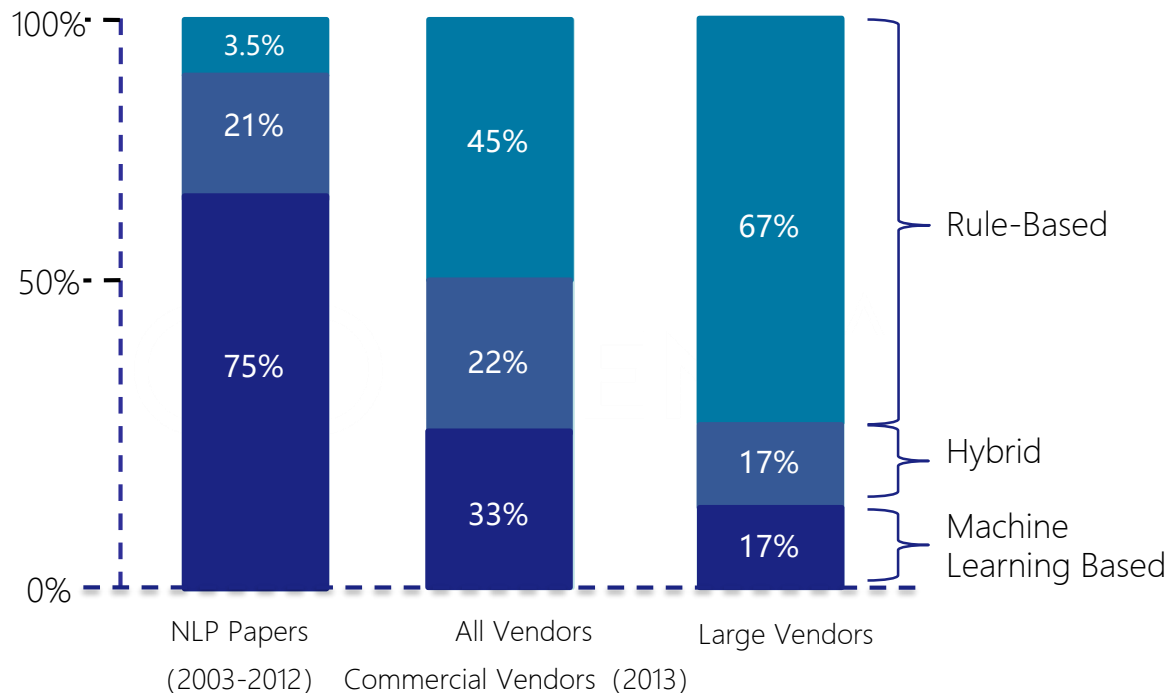
T5



## ► 大模型是否真的能投入生产



## ► Implementations of Entity Extraction



- 对2003年到2013年NLP学术会议论文进行统计，发现进行实体提取的方法是由机器学习方法主导的。
- 在工业界，基于规则的实体抽取算法还是有相当大的比重的。

## ▶ 当前技术发展的几个瓶颈

### 自然语言处理

- 模型是否真的理解了文字？
- 算法能否像人一样理解文本？
- 在足够多的数据量下，是否能学到所有语言？

### 大文档或多文档推理

- 在长篇文本中是否也能有好的推理能力？
- 能否在多文档场景下获得一致的推理能力？

1



3

2



4

### 小语料领域

- 是否需要为特定语言设计专属的算法？
- 是否能在与训练语料分布不同的场景下保持泛化能力？
- 少量语料是否也能得到好的模型？

### 数据集，问题本身和评测标准

- 没有标注数据或少量的场景下使用什么方法？
- 是否定义清楚问题本身？
- 是否有合理的评测标准？



# 目录

- 01 自然语言处理简介
  - 02 对AI的期望
  - 03 NLP发展现状
  - 04 我们能做什么
- 

# 一览群智

一览群智是一家以**自然语言处理**和**知识图谱**等认知智能技术为核心的人工智能公司，为客户提供一站式AI产品和行业解决方案。公司自主研发出智语、智图、智慧、智策四大产品，满足企业在超大规模多源异构情况下的数据治理融合、不同场景下的AI建模，和复杂决策分析需求，让AI技术快速在客户场景落地，打造客户专属的智能决策平台。



金融行业



公共安全



媒体情报



## 探索性工作 | 问答机器人

司法问答client

**小法：您想咨询什么呢？**

迷途的羔羊：你好，今天早上我的车在海淀南路被后车追尾了，肇事车还逃逸了，怎么办？

**小法：建议报警，由交警部门查看事故的现场后依法认定责任，然后再要求对方承担相应的赔偿责任。**

迷途的羔羊：肇事车逃逸了，找不到，怎么办？

**小法：可起诉肇事司机及车主，要求赔偿。**

迷途的羔羊：我起诉他，需要准备什么材料？

**小法：你好，你可以委托律师帮助你。**

input

Send

## 探索性工作 | 文本生成

### 佛经 阿弥陀佛

起始语 如是我闻。

Visual

JSON

status

ok

message

如是我闻。  
一时佛在摩竭提国。  
与大比丘众五百人俱。  
菩萨尔时从兜率天下来下欲界诸天。  
各持天上宝盖来迎菩萨。  
时波旬国诸菩萨等白净王言。  
欲诣尊者大王所设天上。  
有上善法。  
我当往彼奉见菩萨。  
时王即从菩萨下车礼拜迎接。  
既至菩萨两足已。  
右绕三匝却住一面。  
合掌恭敬作如是言

### 金庸小说 飞雪连天射白鹿

Home > docs > demo

起始语 左脚点右脚尖

Go

Visual

JSON

status

ok

message

左脚点右脚尖。黄药师笑道：“药兄请！”身子弹处，转身欲走。欧阳锋大叫：“喂，慢走，老子先到。”黄药师不答，伸手径向他扑去。欧阳锋回头一让，欧阳锋已乘势抓住他背心，提了过去。黄蓉格格一笑，已被欧阳锋抓住了右脚，喝道。黄蓉笑道：“你笑甚么？”欧阳锋一笑，又将他提着放上楼梯。洪七公笑道：“小心他的臭花样，原来你的臭花样精的一个小丫头。”黄蓉微微一笑，伸左手在郭靖肩上一捺，欧阳锋立足不定，身子微微向前一扑。洪七公倒转身子，将郭靖放下。黄蓉大叫：“慢来，慢来。”欧阳锋知道他有意相让，只得放开他左足，双拳齐出，向下急挺，要想挣脱他足上来。哪知洪七公这一拉竟然拿捏不住，只是向他猛扑，只使出降龙十八掌中“飞龙在天”，欧阳

request

左脚点右脚尖

## 探索性工作 | 算法领先

实体识别	F1-Score	分词	F1-Score	词性标注	F1-Score
Ontonotes	82.11	CityU	98.15	CTB9	94.01
Resume	96.72	MSR	98.32	CTB6	95.17
Weibo	72.3	AS	97.1	文本分类	Accuracy
MSRA	95.69	PKU	96.7	ChnSenti	96.33
依存句法分析	UAS	LAS		语义角色标注	F1-Score
CPTB 5.1	90.7	89.3		CoNLL 09	84.38



## 低资源语种 | 维吾尔语NLP

[illegible][illegible]

翻译结果

ترجمه ناتمامست

分词

سۆز تومۇرى بىلەن قوشۇمچە ..

实体识别

ئەمەلىي گەۋدەنى پەرقلەندۈر...

关系抽取

پاچىلىقنىڭ تاللاش

关键词抽取

پاچىلىقنىڭ سۆزلەرنى تاللاش

抽取个数 10

关键词 权重

关键词权重排序图

关键词	权重
节点	0.038
节点	0.023
节点	0.021
节点	0.019
节点	0.018
节点	0.018
节点	0.017
节点	0.017

关键词权重排序图

- 支持维汉翻译
- 维语分词
- 维语实体抽取
- 维语关键词抽取
- 维语关系抽取
- .....

### مقدمه

تاریخچه و اهمیت

فصل اول: کلیات

فصل دوم: مبانی

فصل سوم: روش‌ها

فصل چهارم: یافته‌ها

فصل پنجم: نتیجه‌گیری

## مقدمه

### نتیجه‌گیری

خلاصه و پیشنهاد

فصل اول: کلیات

فصل دوم: مبانی

فصل سوم: روش‌ها

فصل چهارم: یافته‌ها

فصل پنجم: نتیجه‌گیری

## مقدمه

### نتیجه‌گیری

خلاصه و پیشنهاد

翻译结果  
 تەرجىمە نەتىجىسى  
 分词  
 سۆز تومۇرى بىلەن قوشۇمچىلا ..  
 实体识别  
 ئەمەلىي گەۋدىسى پەرقلەندۈر...  
 关系抽取  
 باغلىنىشلىقنى تاللاش  
 关键词抽取  
 ئاچقۇچلۇق سۆزلەرنى تاللاش

关系抽取

关系抽取示意图

图例：

- 人物：红色
- 地点：黄色
- 机构：蓝色

节点：

- 自治区党委组织部
- 党委组织部
- 组织部

边：

- 自治区党委组织部 与 党委组织部
- 自治区党委组织部 与 组织部
- 党委组织部 与 组织部

## 语义解析 | FMR

输入

壹万〇两拾一点一

五与5.8的和的1.5次方

解析

```
nf.math.sum(  
  nf.math.sum(  
    nf.math.mul(1, 10000),  
    nf.math.sum(  
      nf.math.mul(2, 10),  
      1  
    )  
  ),  
  nf.math.decimal(1)  
);
```

```
nf.math.pow(  
  nf.math.sum(  
    5,  
    nf.math.to_number("5.8")  
  ),  
  nf.math.to_number("1.5")  
);
```

执行

denotation: 10021.1

denotation: 35.49242172633476

# 语义解析 | 数字解析语法示例

## 中文数字语法

```

1 <cn_unit>="一"{nf.I(1)}|"二"{nf.I(2)}|"九"{nf.I(9)}|"壹"{nf.I(1)}|"贰"{nf.I(2)}...;
2 <cn_zero>="零"{nf.I(0)}|"〇"{nf.I(0)};
3 <cn_digit>=<cn_unit>{nf.I($1)}|<cn_zero>{nf.I($1)}|<digits>{nf.I($1)};
4 <numbers>=<cn_digit>{nf.I($1)}|<cn_digit> <numbers>{nf.util.concat($1, $2)};
5 <cn_e1>="十"{nf.I(10)}|"拾"{nf.I(10)}; <cn_e2>="百"{nf.I(100)}|"佰"{nf.I(100)};
6 <cn_e3>="千"{nf.I(1000)}|"仟"{nf.I(1000)}; <cn_e4>="万"{nf.I(10000)};
7 <cn_e8>="亿"{nf.I(100000000)}|"万万"{nf.I(100000000)};
8
9 <cn_e1s>=<cn_e1>{nf.I($1)}|<cn_e1> <cn_unit>{nf.math.sum($1, $2)}
10 |<cn_unit> <cn_e1>{nf.math.mul($1, $2)};
11 |<cn_unit> <cn_e1> <cn_unit>{nf.math.sum(nf.math.mul($1, $2), $3)}
12 |<cn_unit>{nf.I($1)}|<cn_zero>{nf.I($1)};
13 <cn_e2s>=<cn_unit> <cn_e2>{nf.math.mul($1, $2)}|<cn_e1s>{nf.I($1)}
14 |<cn_unit> <cn_e2> <cn_unit>{nf.math.sum(nf.math.mul($1, $2), nf.math.mul(10, $3))}
15 |<cn_unit> <cn_e2> <cn_e1s>{nf.math.sum(nf.math.mul($1, $2), $3)}
16 |<cn_unit> <cn_e2> <cn_e1s>{nf.math.sum(nf.math.mul($1, $2), $4)};
17 <cn_e3s>=<cn_unit> <cn_e3>{nf.math.mul($1, $2)}|<cn_e2s>{nf.I($1)}
18 |<cn_unit> <cn_e3> <cn_unit>{nf.math.sum(nf.math.mul($1, $2), nf.math.mul(100, $3))}
19 |<cn_unit> <cn_e3> <cn_e2s>{nf.math.sum(nf.math.mul($1, $2), $3)}
20 |<cn_unit> <cn_e3> <cn_zero> <cn_e2s>{nf.math.sum(nf.math.mul($1, $2), $4)};
21 <cn_e4s>=<cn_e3s> <cn_e4>{nf.math.mul($1, $2)}|<cn_e3s>{nf.I($1)}
22 |<cn_e3s> <cn_e4> <cn_unit>{nf.math.sum(nf.math.mul($1, $2), nf.math.mul(1000, $3))}
23 |<cn_e3s> <cn_e4> <cn_e3s>{nf.math.sum(nf.math.mul($1, $2), $3)}
24 |<cn_e3s> <cn_e4> <cn_zero> <cn_e3s>{nf.math.sum(nf.math.mul($1, $2), $4)};
25 <cn_e8s>=<cn_e4s> <cn_e8>{nf.math.mul($1, $2)}|<cn_e4s>{nf.I($1)}
26 |<cn_e4s> <cn_e8> <cn_unit>{nf.math.sum(nf.math.mul($1, $2), nf.math.mul(10000000, $3))}
27 |<cn_e4s> <cn_e8> <cn_e4s>{nf.math.sum(nf.math.mul($1, $2), $3)}
28 |<cn_e4s> <cn_e8> <cn_zero> <cn_e4s>{nf.math.sum(nf.math.mul($1, $2), $4)};
29
30 <cn_decimal>="点" <numbers>{nf.math.decimal($2)};
31 <number>=<cn_e8s>{nf.I($1)}|<cn_decimal>{nf.I($1)}|<cn_e8s> <cn_decimal>{nf.math.sum($1, $2)};
32 <number>=<cn_e2>{nf.I($1)}|<cn_e3>{nf.I($1)}|<cn_e4>{nf.I($1)}|<cn_e8>{nf.I($1)};

```

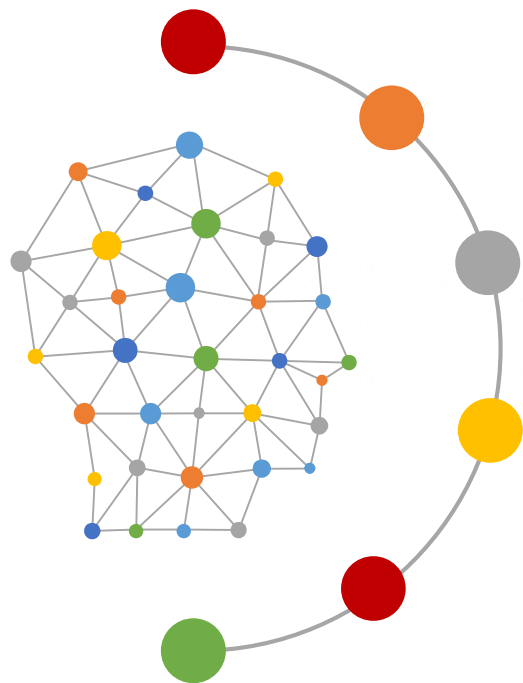
## 英文数字语法

```

1 <unit>="zero"{nf.I(0)}|"one"{nf.I(1)}|"two"{nf.I(2)}|"three"{nf.I(3)} ...
2 <ten>="ten"{nf.I(10)}|"eleven"{nf.I(11)}|"twelve"{nf.I(12)} ... |"nineteen"{nf.I(19)};
3 <ten>="twenty"{nf.I(20)}|"thirty"{nf.I(30)}|"forty"{nf.I(40)} ... |"ninety"{nf.I(90)};
4
5 <e1>=<unit>{nf.I($1)}|<ten>{nf.I($1)}|<ten>{nf.I($1)}
6 |<ten> <unit>{nf.math.sum($1, $2)}
7 |<ten> "-" <unit>{nf.math.sum($1, $3)};
8 <e2>="hundred"{nf.I(100)}; <e3>="thousand"{nf.I(1000)};
9 <e6>="million"{nf.I(1000000)}; <e9>="billion"{nf.I(1000000000)};
10
11 <e2s>=<e1>{nf.I($1)}|<e2>{nf.I($1)}|<e1> <e2>{nf.math.mul($1, $2)}
12 |<e1> <e2> <e1>{nf.math.sum(nf.math.mul($1, $2), $3)}
13 |<e1> <e2> "and" <e1>{nf.math.sum(nf.math.mul($1, $2), $4)};
14
15 <e3s>=<e2s>{nf.I($1)}|<e3>{nf.I($1)}|<e2s> <e3>{nf.math.mul($1, $2)}
16 |<e2s> <e3> <e2s>{nf.math.sum(nf.math.mul($1, $2), $3)}
17 |<e2s> <e3> "and" <e2s>{nf.math.sum(nf.math.mul($1, $2), $4)};
18
19 <e6s>=<e3s>{nf.I($1)}|<e6>{nf.I($1)}|<e3s> <e6>{nf.math.mul($1, $2)}
20 |<e3s> <e6> <e3s>{nf.math.sum(nf.math.mul($1, $2), $3)}
21 |<e3s> <e6> "and" <e3s>{nf.math.sum(nf.math.mul($1, $2), $4)};
22
23 <e9s>=<e6s>{nf.I($1)}|<e9>{nf.I($1)}|<e6s> <e9>{nf.math.mul($1, $2)}
24 |<e6s> <e9> <e6s>{nf.math.sum(nf.math.mul($1, $2), $3)}
25 |<e6s> <e9> "and" <e6s>{nf.math.sum(nf.math.mul($1, $2), $4)};
26
27 <en_decimal> = "point" <numbers>{nf.math.decimal($2)};
28 <en_number> = <e9s>{nf.I($1)};
29
30 <number> = <en_number>{nf.I($1)}|<en_decimal>{nf.I($1)}
31 |<en_number> <en_decimal>{nf.math.sum($1, $2)}
32 |<en_number> "and" <en_decimal>{nf.math.sum($1, $3)};

```

## ▶ AI行业应用中的小样本学习问题解决思路—智语AI平台



### 自动化机器学习

自动化机器学习，强大的ETL以及快速模型开发和部署等关键功能。极大地提高了工作效率，消除了分析过程中的瓶颈和AI应用程序的构建。

### 模型训练流程完善

通过内置的智能化算法进行数据类型检测、数据预处理、特征选择、模型选择以及参数调优，提供人机结合的模型训练及调优。

### 模型快速发布，系统灵活部署

工具/平台容器化，支持快速部署。兼容各种版本操作系统。

## ▶ AI行业应用中的小样本学习问题解决思路—智慧标注平台

### 01

管理流程完善，支持针对用户标注技能的考核、管理，适用于大规模、高质量的标注需求；

管理流程完善



### 02

标注过程流畅，尊重用户体验，提供了全文匹配、添加字典等便捷操作的功能，最大程度上便捷标注操作。

用户体验流畅



### 03

通过主动学习，大幅度提升标注效率，从手动标记变为简单确认；

基于主动学习



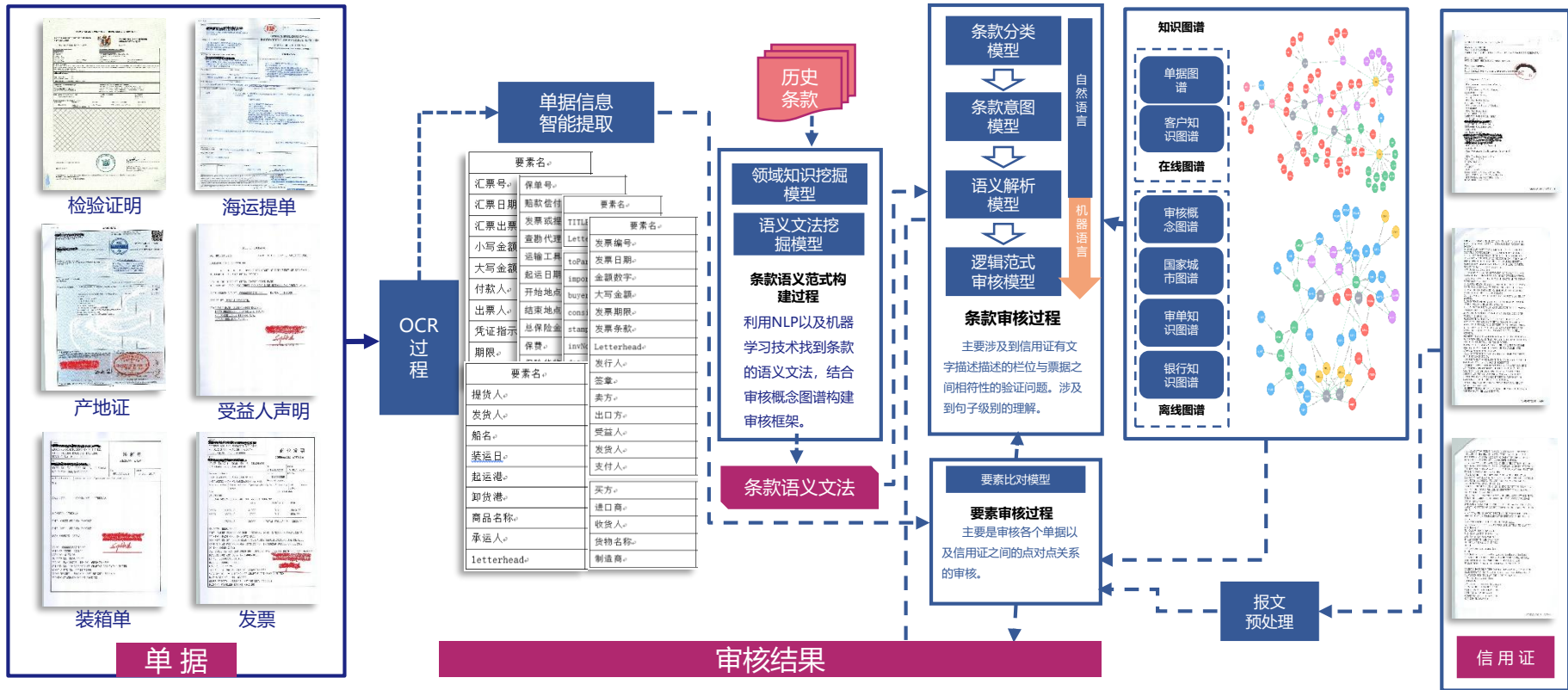
### 04

提供了诸如多人标注同一条数据、人员管理的功能，可大幅提高标注质量。

严控标注质量



## 行业应用 | 国际结算智能审单



## 行业应用 | 案件智能搜索

输入条件

1号上午经过长春桥的白色大众车

搜索

语义解析

时间: 20190901 6-12点

地点: 长春桥

颜色: 白色

品牌: 大众

要素: 车

数据库

违章数据

车辆登记数据

串并案数据

....

结果



筛选

违章记录

犯罪记录

失窃车辆

备案车辆

结果



京X 12345

张三

王五运毒案

车主

涉案

京X 12345

同车人员

伴随车

李三

津X 12345

## ► 行业应用 | 地址标准化



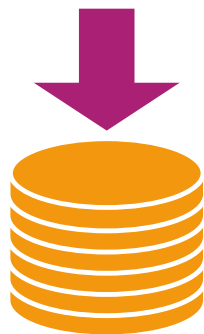


## 行业应用 | 地址标准化

原始地址

克州人民医院住院部

省级	地级	区县级	乡级	社区、居（村）委会	路	POI	其他
	克州					人民医院	住院部

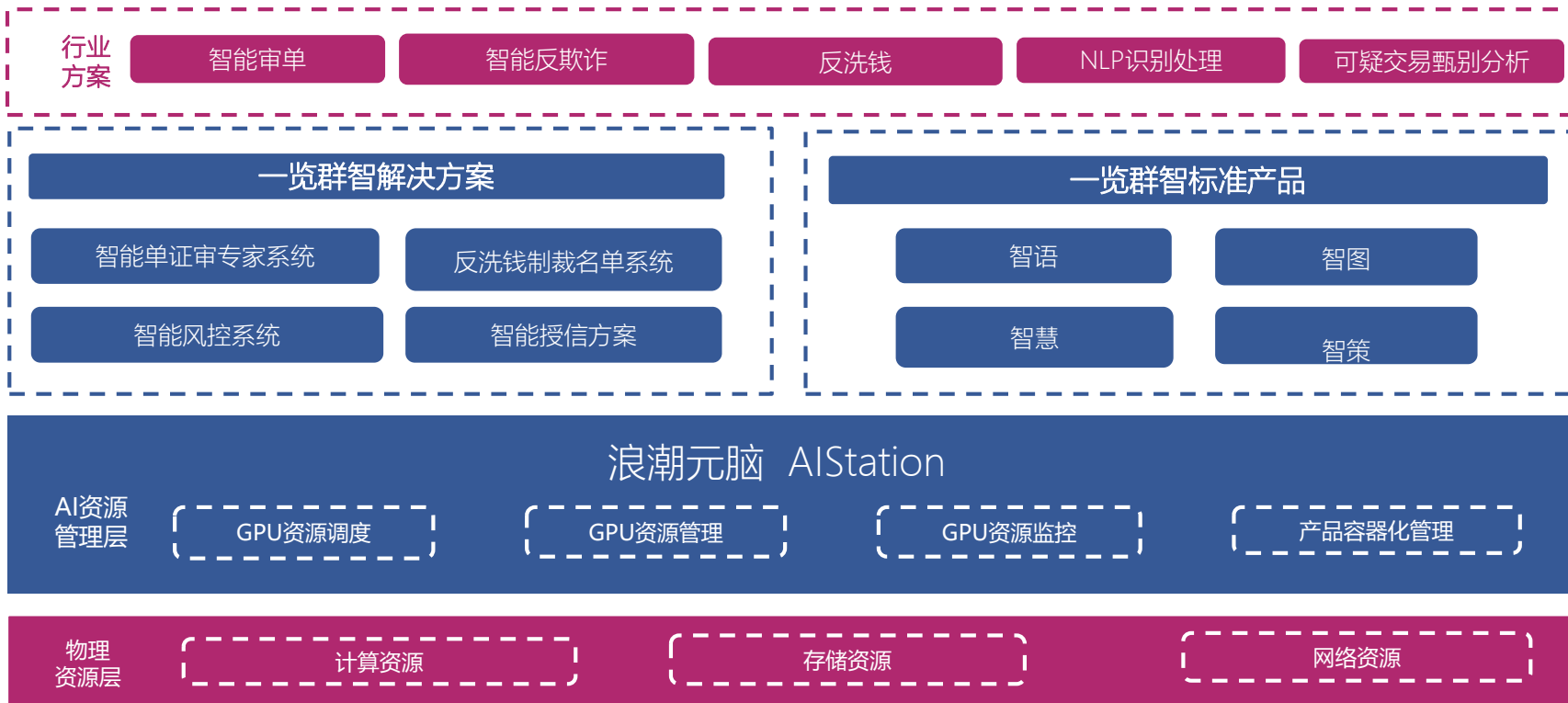


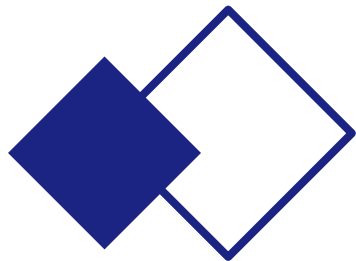
标准地址库



省级	新疆维吾尔自治区
地级	克孜勒苏柯尔克孜自治州
区县级	阿图什市
乡级	幸福路街道
社区、居（村）委会	帕米尔西路社区
路	帕米尔西路5号
POI	人民医院
其他	住院部

## ▶ 一览群智VS浪潮元脑生态战略合作





# THANKS!

让人工智能成为**生产力**

