达观数据
DATA GRAND

智能文本信息抽取技术

达 观 数 据　　高 翔

达观数据：专注于文本智能处理的高科技创业企业

达观数据
DATA GRAND

达观数据成立于2015年，总部位于上海浦东软件园，同时在北京、深圳、成都、西安设立产品和解决方案中心，专注于为客户提供文本智能化处理的软件系统

达观运用先进的自然语言处理（NLP）技术，提供的智能系统能够自动对文本进行抽取、审核、纠错、搜索、推荐、写作等操作，让计算机代替人来完成工作，大幅提高效率

先后获得宽带资本、软银赛富、真格基金、元禾重元、联想之星等国际著名投资机构的超2亿元融资，是中国文本语义分析类创业企业中获得融资金额最多的企业

CBC 宽带资本 | SAIF Partners | ZhenFund 真格基金

元禾 ORIZA | 联想之星 Legend Star | 方广资本 F&G VENTURE

钟鼎资本 EASTERN BELL CAPITAL | 众麟资本 Z CAPITAL | 掌门集团

达观数据
DATA GRAND

01

文本挖掘简介

文本智能处理属于人工智能的三大块领域之一 具有广阔应用

达观数据
DATA GRAND

文本智能处理是AI的重点和难点

在对图像、语音等感知层面的处理完成后，进一步对文本进行认知层面的自动处理，模拟人类智慧分析过程，号称是人工智能皇冠上的明珠

NLP是指让计算机代替人类自动化的进行文字（自然语言）的相关处理

达观数据
DATA GRAND

自然语言处理 NLP
Natural Language Processing

自动阅读

自然语言理解 NLU
Natural Language Understanding

自动化的审批、核对、纠错、搜索、推荐、比对、分析、评判等功能，大幅度减轻人工负担

自动写作

自然语言生成 NLG
Natural Language Generation

自动完成写作工作，如自动填表、摘要写作、润色修改、写作提示、内容扩充等

NLP发展简史

达观数据
DATA GRAND

1950S

符号主义：以机器翻译为开端，作为早期尝试，但不是很成功

1980s

语法规则：多数自然语言处理系统基于规则，人工修订等方式，包括问答、翻译、搜索等

1990s

统计学系：基于统计机器学习技术及语料库，使用统计模型，NLP发展产生革新

2006~至今

深度学习：深度学习起步、发展及成熟，同样影响NLP领域，从传统的机器学习逐渐过渡到深度学习，2018年预训练语言模型兴起，是nlp发展史又一个重大里程碑

人类运用文字的三个特点和计算机的价值

达观数据
DATA GRAND

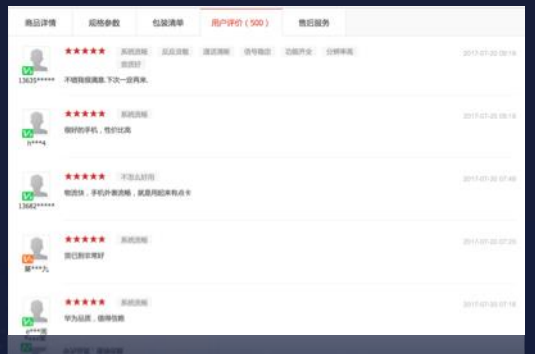
- 人短期阅读文字很快，但是长期很容易遗忘，无法记住细节 → 让计算机来进行归纳和搜索
  - 人脑从来都不擅长记忆特别具体的信息，所以对文本内容的归纳、整理、搜索、对比等请计算机来代劳能大幅度提高效率
- 人阅读文字很快，但是写作总是很慢 → 所以可以让计算机协助完成初稿写作
  - 人的阅读可以一目十行，非常迅速，1分钟能看完一篇1000字的文章，但是写作1000字的文章要几个小时，甚至几天时间
- 人从文字中解读整体意思的很容易，但操作局部内容很慢 → 让计算机协助完成细节处理
  - 请看这段文字：研表究明，乱错的文字并不一定影阅响读！但是要逐一整调文字的错误往往是费时的

应用场景：如何让计算机自动处理文本数据

达观数据
DATA GRAND

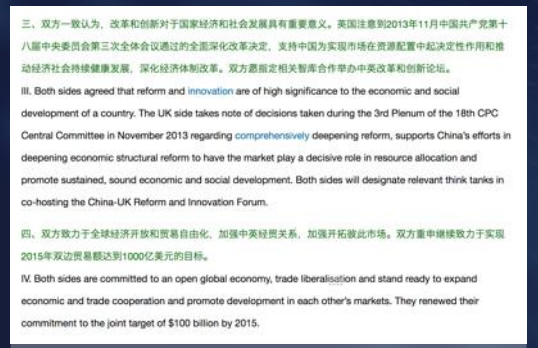
法律/人事/证券等专业文本

企业合同/公文



客户评论意见

- 常见需求：从文本中抽取出关键的人、事、数字、关系、动态、核心条款、风险内容等信息。



企业产品手册

新闻文章

问答资料

- 识别文本内容中的核心观点、或情感倾向，或比对多个文档间的关联信息，或在多个文本中进行搜索

达观数据
DATA GRAND

02

抽取算法概述

信息抽取是什么

达观数据
DATA GRAND

文本信息抽取指的是从自然语言文本中抽取指定类型的**实体**，**关系**，**事件**等事实信息，并形成结构化数据输出的文本处理技术。信息抽取的任务主要有：命名实体识别，关系抽取，事件抽取等。

命名实体识别（Named-entity recognition，NER）是信息抽取中的重要任务，一般需要抽取信息中的人物、地点、机构、时间等内容。下面以识别公司名称抽取举例说明。

识别公司名任务—符号主义

达观数据
DATA GRAND

穷举所有公司名称，配进词典用于匹配

中国石油化工股份有限公司
中国石油天然气股份有限公司
中国建筑股份有限公司
中国平安保险（集团）股份有限公司
上海汽车集团股份有限公司
中国移动有限公司
中国工商银行股份有限公司

- 公司名称太多，整理费时费力
- 公司存在各种简称，无法全部覆盖
- 每天都有新的公司产生，词典维护成本高
- 没有考虑上下文，名称存在歧义，无法精确匹配

识别公司名任务—语法规则

达观数据
DATA GRAND

穷举所有句法，抽取公司名称

- xx是xx公司
- xx作为xx公司
- xx公司成立于xx
- 新成立了xx公司

- 句子结构上下文复杂，生成准确规则困难
- 不同人书写习惯不同，句法结构太多，无法全部覆盖
- 不同规则可能存在冲突，规则量大了之后维护困难

识别公司名任务—基于统计的机器学习

达观数据
DATA GRAND

- 基于标注训练，使用统计方法进行抽取
- 通过从大量文字语料中统计上下文分布规律来进行语义分析
- 常见的序列标注机器学习方法，如HMM，CRF

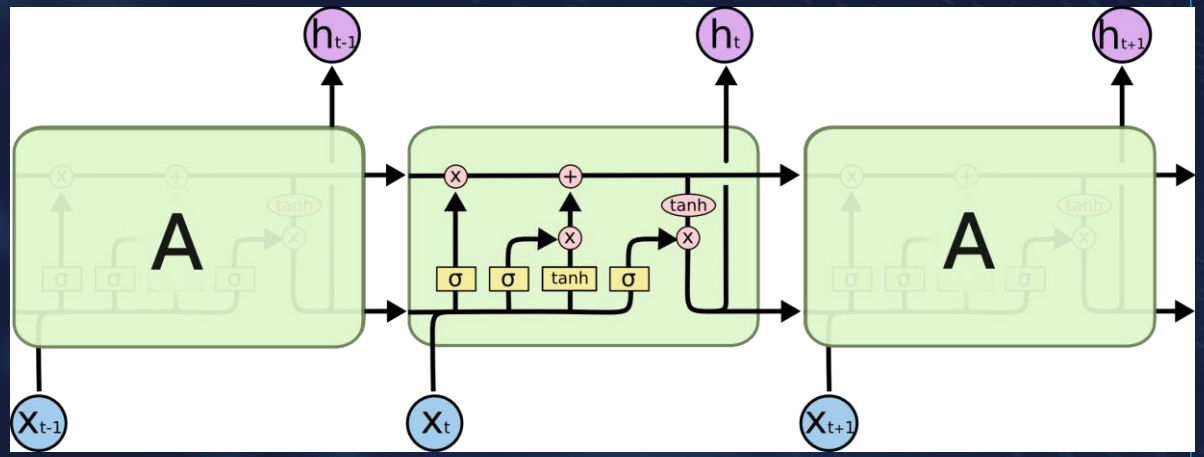
- 标注：达观数据是人工智能公司
- 预测：虚拟数据是人工智能公司
- 预测：上海的虚拟数据是人工智能公司
- 预测：上海的虚拟数据是高科技公司

- 需要人工进行一定量的数据标注
- 需要做特征工程，特征工程质量如何决定了一定标注数据量下的效果
- 训练和预测需要消耗一定的机器资源

识别公司名任务—深度学习

达观数据
DATA GRAND

- 通过深度神经网络技术，对整个篇章的内容进行整体性的表示学习
- 常见的深度学习网络结构如RNN，LSTM，DPCNN等



- 需要大量的标注数据
- 可解释性非常弱
- 训练和预测都需要消耗比较多的 机器资源

序列标注

他	来	自	达	观	数	据
S	B	E	B	M	M	E

Label Set
B: Begin　M: Middle　E: End　S: Single

将信息抽取问题转化为了4种类型的分类问题

- 经典的NLP算法通过采集大量上下文特征来进行分类
- 深度学习通过端到端学习来自动学习特征

时间序列分析的相关思路可以广泛运用

- 传统的隐马尔可夫和Viterbi算法可以运用，LSTM/GRU也可以
- 从中文分词、词性标注、命名实体识别，到特定领域的核心信息抽取，都可以运用序列标注的思想来处理

达观数据
DATA GRAND

03

传统抽取算法介绍

传统抽取算法

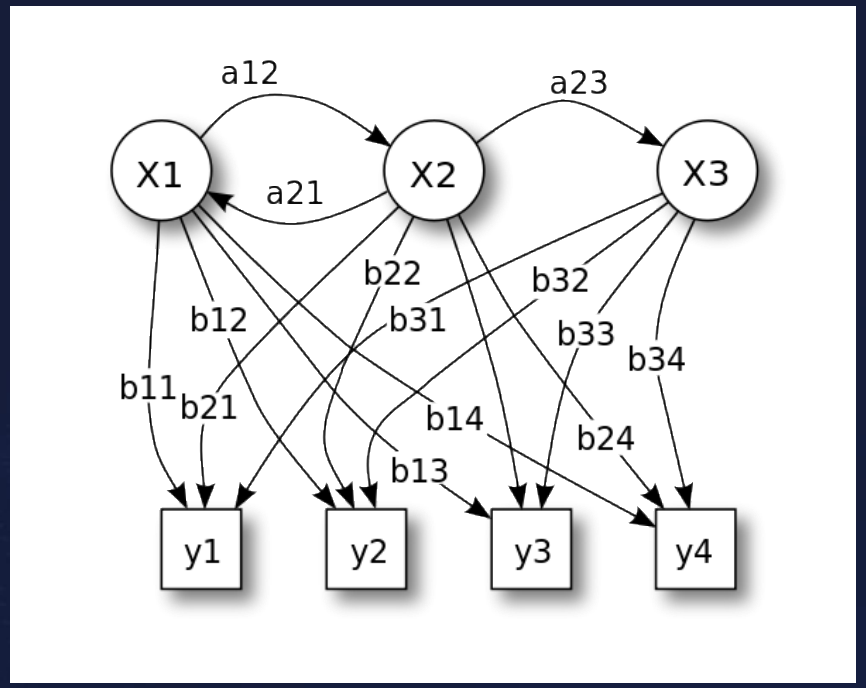
达观数据
DATA GRAND

在基于机器学习的方法中，信息抽取常被当作序列标注问题。利用大规模语料来学习出标注模型，从而对句子的各个位置进行标注。**常用模型包括生成式模型HMM、判别式模型CRF等。**HMM和CRF是结合概率论和图论的模型，也是基于统计机器学习的算法，模型都是根据训练出来的概率做最优结论选择。

HMM隐马尔可夫模型

HMM模型描述一个含有隐含未知参数的马尔可夫过程，核心包括二序列（隐藏序列、观察序列）三矩阵（初始状态矩阵、发射状态矩阵、状态转移矩阵）

观察序列	他	来	自	达	观	数	据
隐藏序列	S	B	E	B	M	M	E



初始状态矩阵


```
{
    "S": -1.2970671718,
    "B": -0.3192859282
}
```


发射状态矩阵


```
{
    "E": {
        "怀": -9.5802693146,
        "挂": -11.1763656101,
        "耀": -11.2449282815,
        "涉": -10.6499987291,
        "谈": -8.4366922205,
```


状态转移矩阵


```
{
    "E": {
        "S": -1.8035633406,
        "B": -1.8672131654
    },
    "S": {
        "S": -2.1705700081,
        "B": -1.8369231177
```

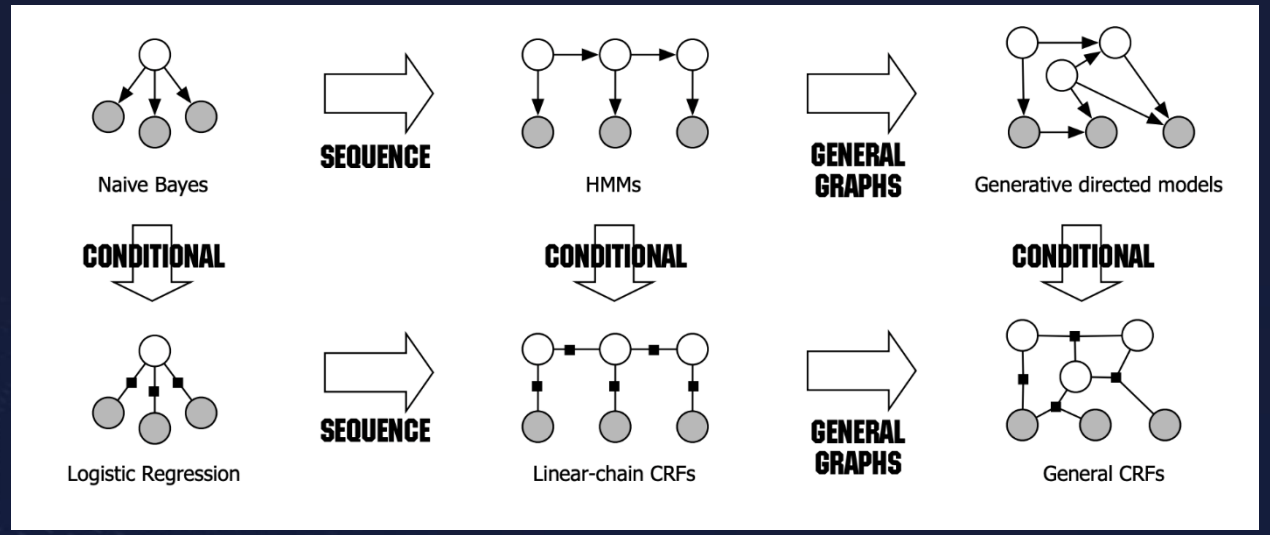

学习算法
监督学习：极大似然估计
非监督学习：Baum-Welch

解码算法
Viterbi

CRF条件随机场

达观数据
DATA GRAND

CRF是目前信息抽取的主流模型，它的目标函数不仅考虑输入的状态特征函数，而且还包含了标签转移特征函数。CRF的优点在于其为一个位置进行标注的过程中可以利用丰富的内部及上下文特征信息。




```
"tags": [
    "B_nz",
    "E_nz",
    "M_nz",
    "O",
    "S_nz"
]
```

```
"feature_template": [
    "U00:%x[-3,0]",
    "U01:%x[-2,0]",
    "U02:%x[-1,0]",
    "U03:%x[0,0]",
    "U04:%x[1,0]",
    "U05:%x[2,0]",
    "U06:%x[3,0]",
    "U07:%x[-1,0]/%x[0,0]",
    "U08:%x[0,0]/%x[1,0]",
    "B"
],
```

```
"feature_func_weight": {
    "U06:径": [
        -0.3300855925293462,
        -0.2881151933425375,
        0.9348927899674289,
        -0.3163016610989254,
        -0.0003903429977172
    ],
    "U06:待": [
        -0.0761171148843781,
        -0.3304252678324269,
        -0.0258093469791894,
        0.4334372103636684,
        -0.0010854806687564
    ],
```

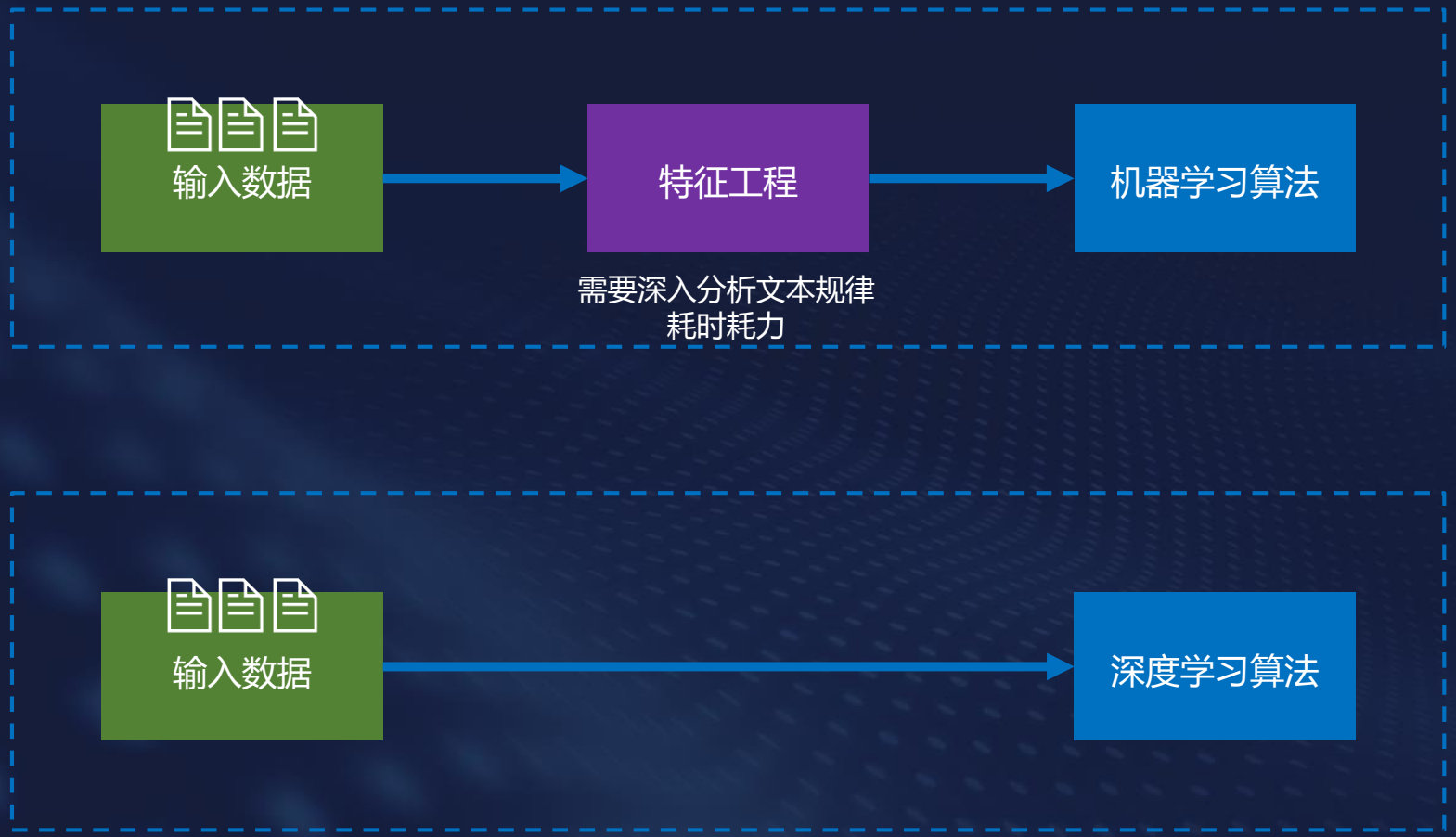
```
"trans_func_weight": {
    "B_nz": {
        "B_nz": -1.6447157870113511,
        "S_nz": -0.5746089198775536,
        "E_nz": 5.448357196490l385,
        "O": -7.017360470166356,
        "M_nz": 5.470838223726445
    },
    "S_nz": {
        "B_nz": -0.7763750119890296,
        "S_nz": -0.2810642872369533,
        "E_nz": -0.6441535798346163,
        "O": 0.7711532845280676,
        "M_nz": -0.425280812972025
    },
```

达观数据
DATA GRAND

04

基于深度学习的抽取算法

文本挖掘处理过程：经典机器学习 VS 深度学习



- 采用经典机器学习方法来进行文本处理，需要进行非常多的特征抽取工作

  TF-IDF
  互信息
  信息增益
  期望交叉熵
  主成分分析
  …… ……

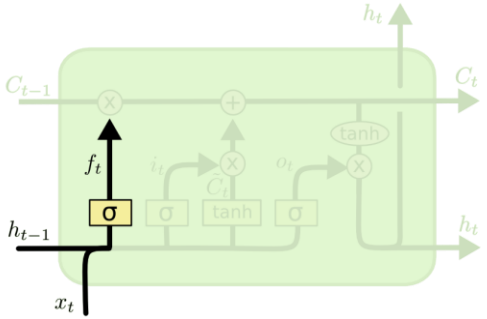
- 特征工程往往需要手工寻找，花费大量人力，特征的好坏往往决定效果。特征依赖对文本内容的理解甚至领域知识

- 深度学习把文本视作一个序列输入的信号，通过网络进行信息的组合和规模抽取，优点是可以省略特征工程

LSTM原理

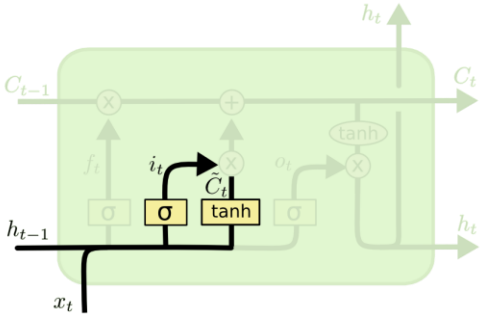
达观数据
DATA GRAND

1，单元状态丢弃



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

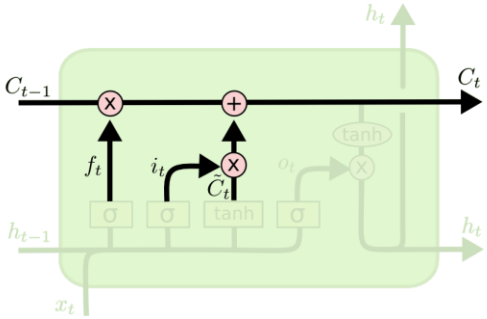
2，新信息选择



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

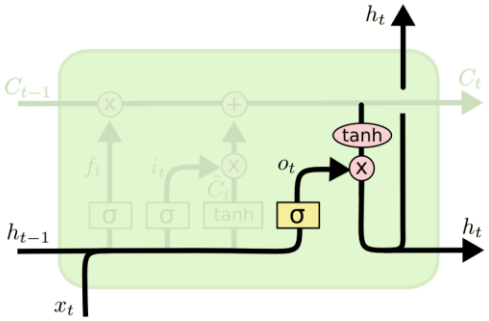
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

3，单元状态更新



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

4，确定输出

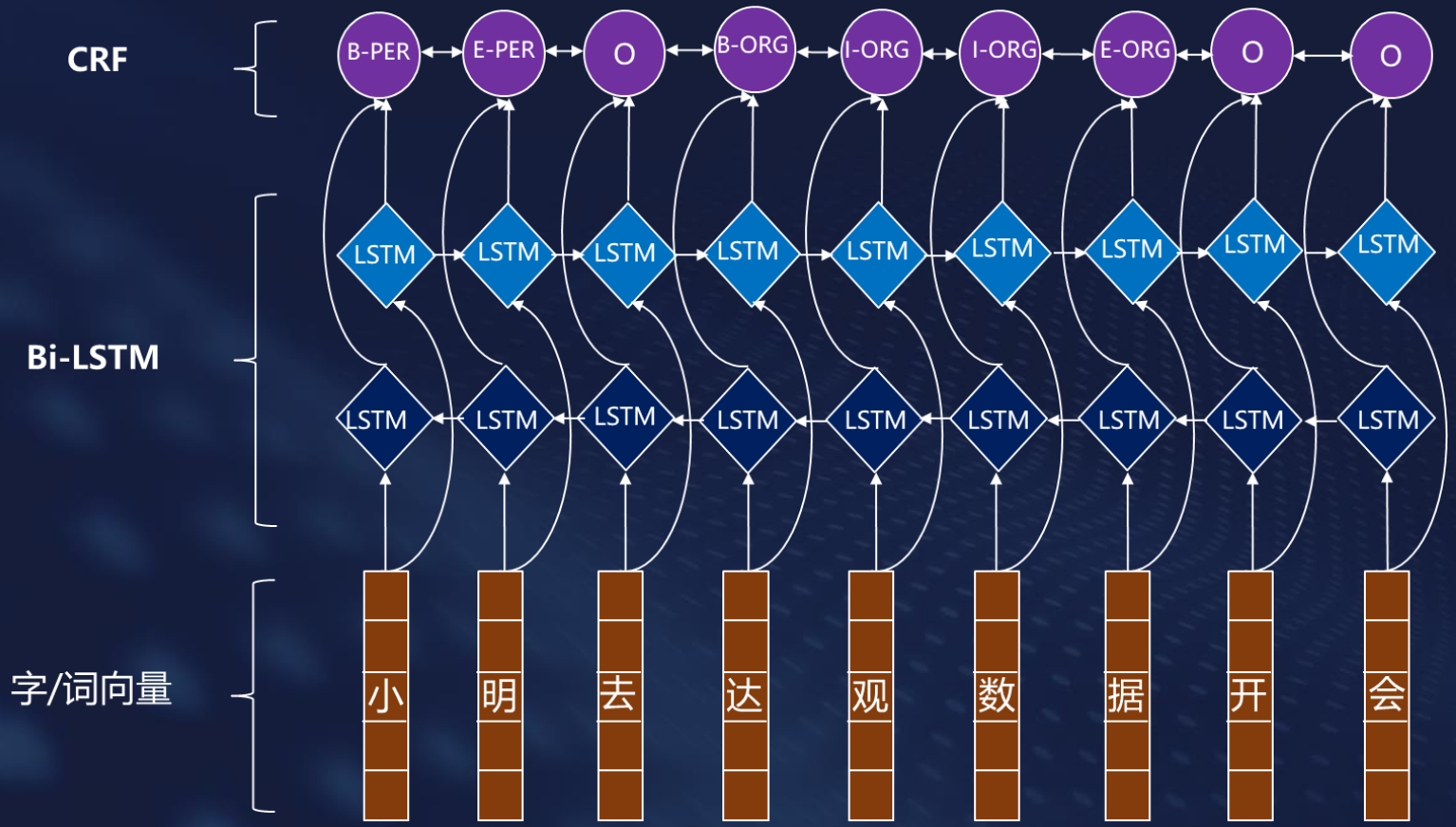


$$o_t = \sigma\left(W_o \ [h_{t-1}, x_t] \ + \ b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

基于深度学习的信息抽取方法

达观数据
DATA GRAND



- Bi-LSTM双向网络分别从前往后和从后往前进行序列信号的记忆和传递是常见做法

  华为发布了新一代的麒麟处理X
  X鲜和美国签订了新一轮的谅解备忘录

- CRF等经典方法结果可控性好，在序列标注时，在顶层用CRF对Bi-LSTM的结果进行二次操作可得到更好的结果

- 信号输入层，对中文进行embedding能起到非常好的效果

- 对英文先进行卷积CNN操作往往能抽取出单词的前后缀等信息，对提升效果有帮助

Vector Representations字词表示

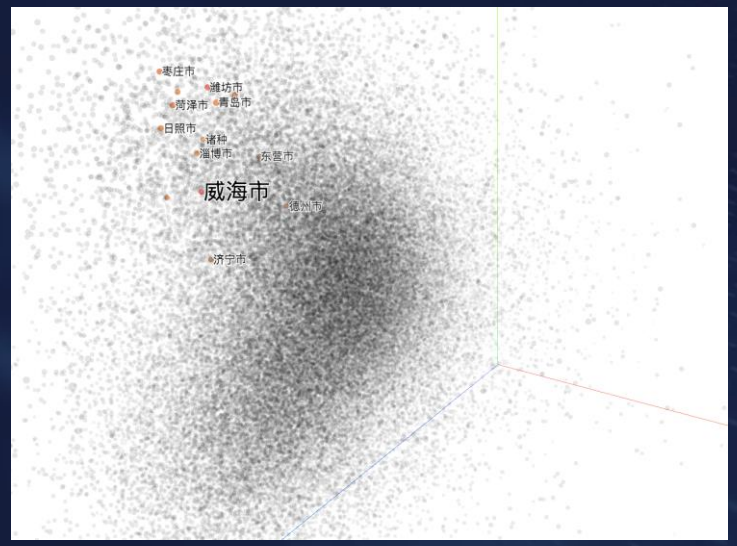
达观数据
DATA GRAND

- 使用Word2Vec进行低维连续空间的映射（CBOW/Skip-Gram）进行词汇embedding

威海市　[-2.0795249939, 1.4055569172, 1.9540510178, ... -0.651816964, -6.1333961487, -0.5107190013]

潍坊市　[-0.9602200985, 0.8771957159, 1.0565081835, ... 4.1443724632, -4.1823129654, -0.2311971784]

枣庄市　[-2.5211799145, -0.6317474842, -0.052895709, ... 2.8651976585, -3.9351148605, 1.3284717798]



潍坊市	0.363
枣庄市	0.424
菏泽市	0.441
青岛市	0.486
泰安市	0.487
德州市	0.491
日照市	0.492
聊城市	0.492
济宁市	0.497
滕州市	0.504
淄博市	0.504
东营市	0.507

- 不仅词向量之间的矢量距离可以体现语义相关性，并且矢量还有一定的传递性

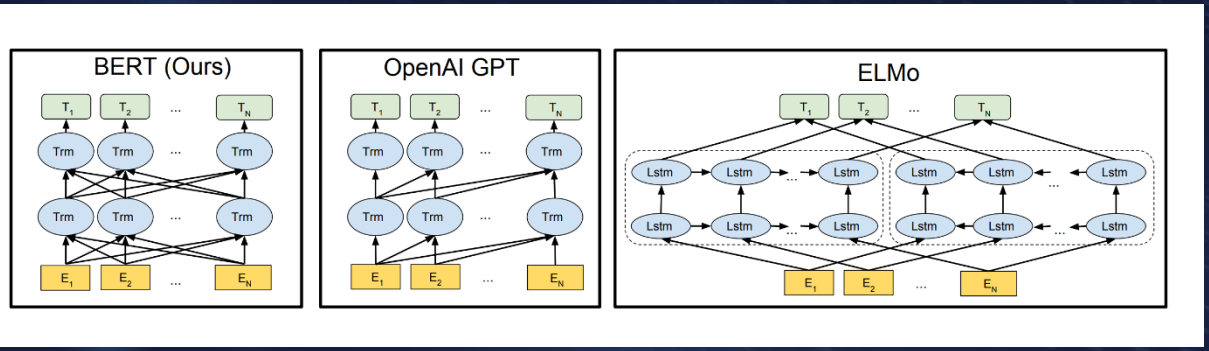
Vector（山东省）- Vector(威海市） = Vector（广东省）- Vector(佛山市)
Vector（皇帝）- Vector(皇后） + Vector（女人） = Vector(男人)

自然语言处理中的预训练模型

达观数据
DATA GRAND

静态表示：Word2Vec、Glove

动态表示：ELMO、GPT、BERT、ERNIE、MASS、XLNet...




```
>>> texts = ['苹果好吃', '苹果手机']
>>> np_result = bert_extractor.extract(texts)
Content to be convert by Bert Server length as 2
Bert Server returns in :0.1928408145904541 seconds
>>> print("numpy array:{}".format(np_result))
numpy array:[[[ 0.9161428  -0.3754939  -0.783268   ... -0.28063536  0.00458466
   0.00976186]
  [ 0.7300041  -0.8192198  -1.1590743  ... -0.10063308  0.5126673
   0.1534313 ]
  [-0.6886213  -0.03489333 -2.016646   ...  0.4749214   0.76233804
   0.12930985]
  [ 0.1035582  -0.3602897  -2.0146933  ...  0.62840873  0.45774418
  -0.05852213]]

 [[ 0.4175677  -0.15601903 -0.05043844 ... -0.43618444 -0.84303737
   0.00353375]
  [ 0.6092674  -1.0740471  -0.19154301 ...  0.2748021  -0.38275814
  -0.5265086 ]
  [-0.17926726 -0.95220983  0.09756972 ...  0.08836851  0.1829703
   0.3473437 ]
  [-0.12692973 -0.7641636  -0.2247644  ...  0.91655016 -0.32559547
  -0.05220592]]]
>>>
```

达观数据
DATA GRAND

05

抽取算法在达观数据的应用实践进阶

文本自动化处理的需求非常普遍

达观数据
DATA GRAND

常见应用场景

- 财务报表账目信息抽取
- 商业票据关键信息识别
- 应标书信息自动导出
- 基金合同差异核对
- 投资报告项目信息自动提取
- 法律文书风控要素审核
- 新闻稿文字校对
- 政府补贴项目申请表内容核准
- ……

文档智能审核功能

文档智能抽取功能

文档智能搜索功能

文档智能撰写功能

达观智能文档审阅平台

数据增强

达观数据
DATA GRAND

中国科学院院士朱清时、闵恩泽,美国环保署绿色化学项目负责人保罗·阿纳斯塔斯博士、美国化学会绿色化学研究所乔·布林博士、美国麻省理工学院斯坦菲尔德教授作了大会邀请报告。

研讨会收到国内外学术报告80余篇。

近年来,共青团中央组织过多批包括少数民族共青团干部在内的基层共青团干部到团中央机关挂职锻炼,产生了积极影响。为进一步深化跨世纪青年人才工程,加快培养民族地区少数民族共青团干部的步伐。

中国科学院院士朱清时、闵恩泽,美国环保署绿色化学项目负责人保罗·阿纳斯塔斯博士、美国化学会绿色化学研究所乔·布林博士、美国麻省理工学院斯坦菲尔德教授作了大会邀请报告。研讨会收到国内外学术报告80余篇。

研讨会收到国内外学术报告80余篇。近年来,共青团中央组织过多批包括少数民族共青团干部在内的基层共青团干部到团中央机关挂职锻炼,产生了积极影响。为进一步深化跨世纪青年人才工程,加快培养民族地区少数民族共青团干部的步伐。

中国科学院院士朱清时、闵恩泽,美国环保署绿色化学项目负责人保罗·阿纳斯塔斯博士、美国化学会绿色化学研究所乔·布林博士、美国麻省理工学院斯坦菲尔德教授作了大会邀请报告。研讨会收到国内外学术报告80余篇。近年来,共青团中央组织过多批包括少数民族共青团干部在内的基层共青团干部到团中央机关挂职锻炼,产生了积极影响。为进一步深化跨世纪青年人才工程,加快培养民族地区少数民族共青团干部的步伐。

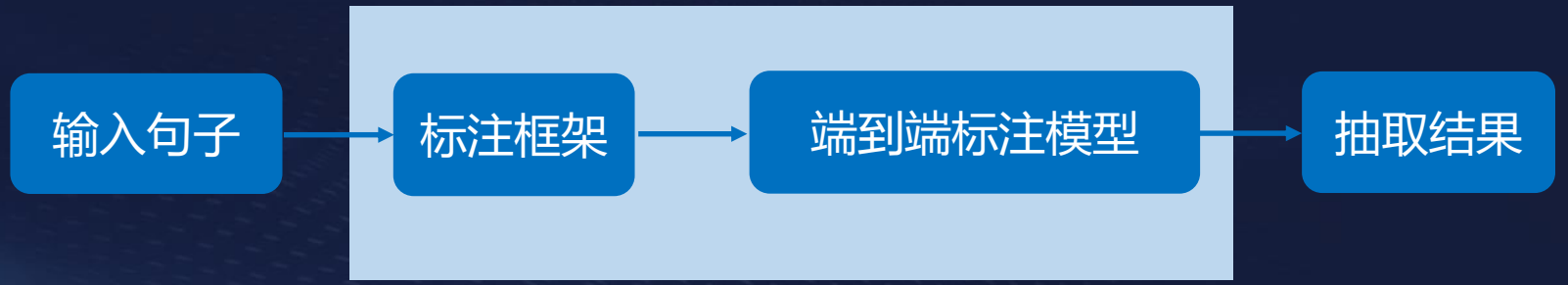
非监督Embedding训练

达观数据
DATA GRAND

知识图谱关系抽取：基于联合标注

达观数据
DATA GRAND

- 将抽取问题转换成标注任务，训练一个端到端标注模型来抽取关系
- 根据标签序列，将同样关系类型的实体合并成一个三元组作为最后的结果



输入：美 国 总 统 特 朗 普 将 考 察 苹 果 公 司，该 公 司 由 乔 布 斯 创 立。

标签：

B-CP-1				E-CP-1	O		O	B-CP-2			E-CP-2	O		O	B-CF-1			I-CF-1	O		O		O	I-CF-2				O		O
										I-CP-2		O		O		I-CF-1		E-CF-1		O		O	B-CF-2		E-CF-2			O	

输出：(美国,国家-总统,特朗普)　(苹果公司,公司-创立者,乔布斯)

总结：在实际工程中运用深度学习挖掘文本的思考

达观数据
DATA GRAND

优点：

- 可以使用非监督数据训练字词向量，提升泛化能力
- 端到端，提供新思路
- 一些模型结构能够克服传统模型缺点

缺点：

- 小数据量情况下难以保证效果
- 调参工作量有时不亚于特征工程
- 客户部署硬件环境限制

思考：

- 在业务场景下，尽量收集并理解数据，分析问题本质，选择合适模型
- 初始阶段可以使用传统机器学习模型快速尝试，再引入深度学习技术
- 疑难问题使用端到端的方式也许会有惊喜
- 关注最新的前沿技术（对抗网络，强化学习，迁移学习）
- 数据决定效果上限，模型逼近此上限
- 不断尝试，从挫折中总结规律

达观数据
DATA GRAND

“达观杯”
文本智能信息抽取挑战赛

06月28日-08月31日

¥30000元 /队
一等奖 x 1支队伍

¥10000元 /队
二等奖 x 2支队伍

¥5000元 /队
三等奖 x 3支队伍

¥3000元 /队
优胜奖 x 4支队伍

TOP30
达观数据全职和实习工作的
面试直通机会

赛事官方QQ群

807070500

达观数据
DATA GRAND

达观数据

文本智能处理专家