UCLOUD

悟有所值 Ucan 下午茶 Ucan Afternoon Tea

文本智能处理的 深度学习技术

达观数据 张健

UCLOUD 专业云计算服务商

IT大咖说 知识共享平台

悟有所值 Ucan 下午茶 Ucan Afternoon Tea

目录

01 文本智能处理背景

02 深度学习与 NLP

03 深度学习用于各类型文本应用的实践方法

04 达观数据文本挖掘的实践经验

05 总结&QA

01

文本智能处理背景简介

人工智能中的文本处理细分领域简介

什么是 NLP

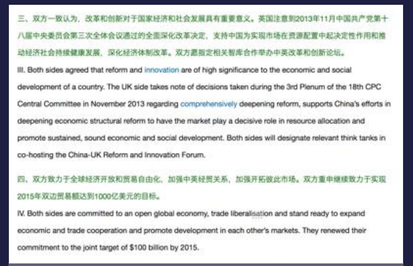
- **概念**：Natural Language Processing 自然语言处理
- **目的**：让机器理解人类的语言，是人工智能领域的重要分支，用于分析、理解和生成自然语言，方便人机交流
- **应用**：智能问答，机器翻译，文本分类，文本摘要，标签提取，情感分析，主题模型

日常工作中各类常见文本形式

法律/人事/证券等专业文本

企业合同/公文

客户评论意见



三、双方一致认为，改革和创新对于国家经济和社会发展具有重要意义。英国注意到2013年11月中国共产党第十八届中央委员会第三次全体会议通过的全面深化改革决定，支持中国为实现市场在资源配置中起决定性作用和推动经济社会持续健康发展，深化经济体制改革。双方愿指定相关智库合作举办中英改革和创新论坛。

III. Both sides agreed that reform and innovation are of high significance to the economic and social development of a country. The UK side takes note of decisions taken during the 3rd Plenum of the 18th CPC Central Committee in November 2013 regarding comprehensively deepening reform, supports China's efforts in deepening economic structural reform to have the market play a decisive role in resource allocation and promote sustained, sound economic and social development. Both sides will designate relevant think tanks in co-hosting the China-UK Reform and Innovation Forum.

四、双方致力于全球经济开放和贸易自由化，加强中英经贸关系，加强开拓彼此市场。双方重申继续致力于实现2015年双边贸易额达到1000亿美元的目标。

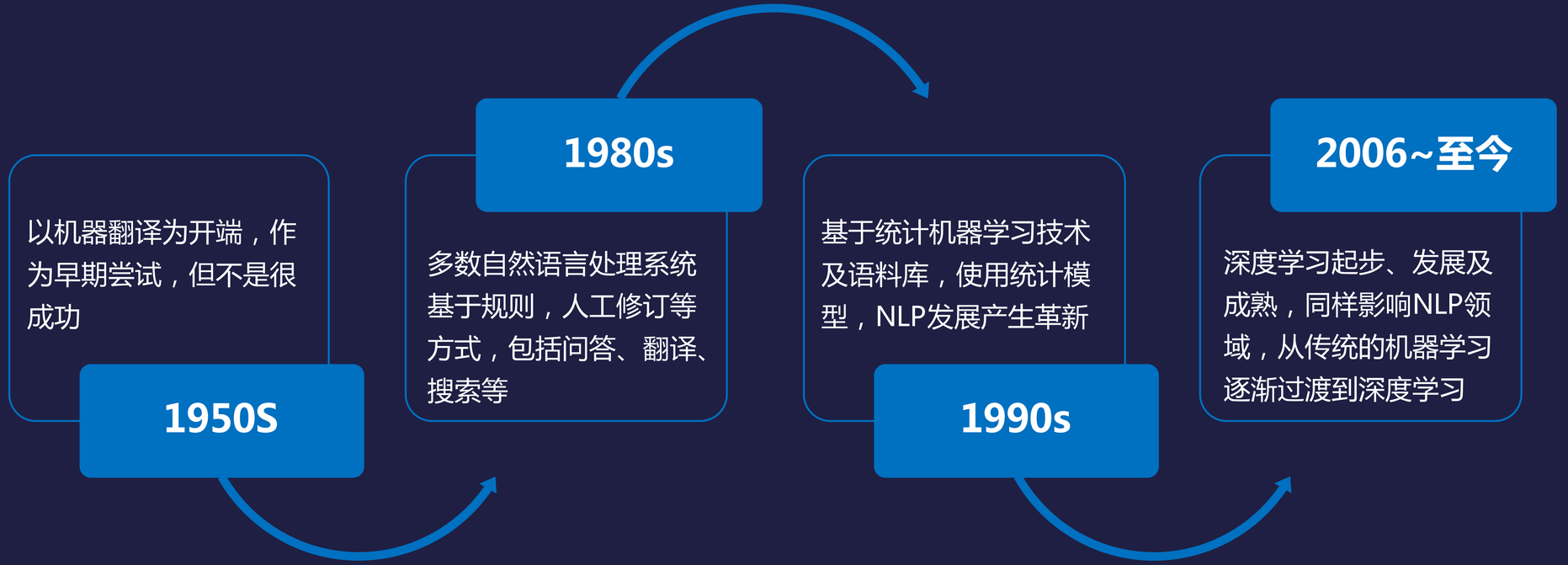
IV. Both sides are committed to an open global economy, trade liberalisation and stand ready to expand economic and trade cooperation and promote development in each other's markets. They renewed their commitment to the joint target of $100 billion by 2015.

企业产品手册

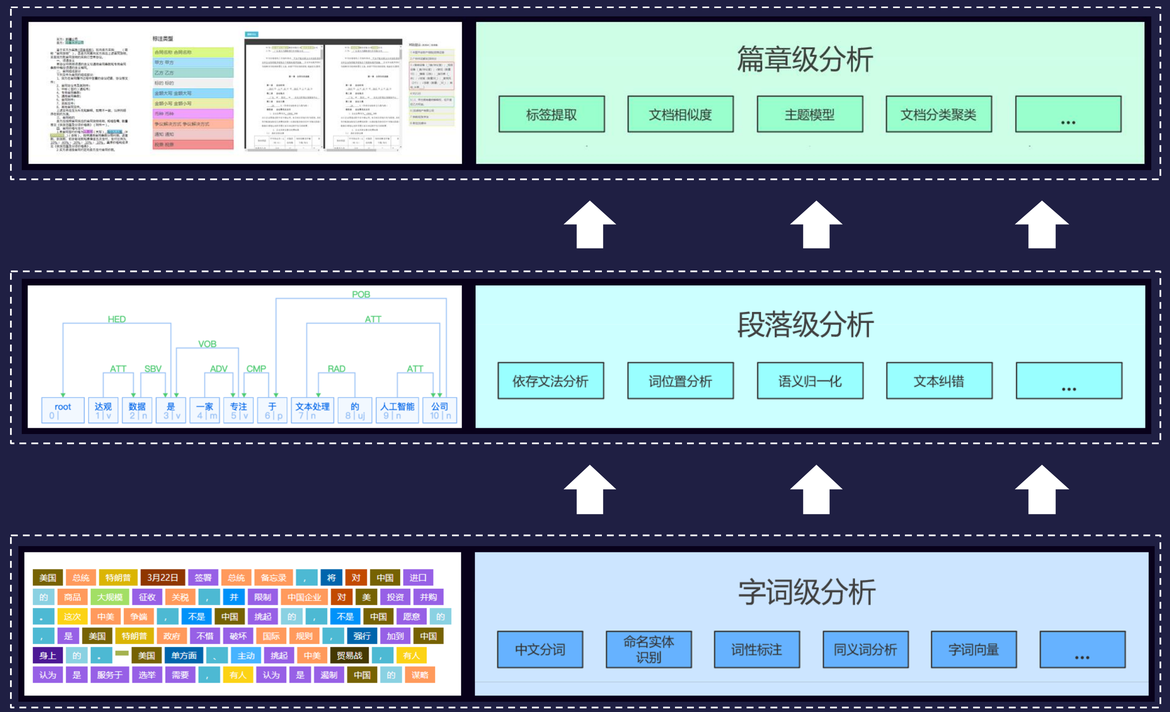
新闻文章

问答资料

NLP发展简史



NLP技术层次

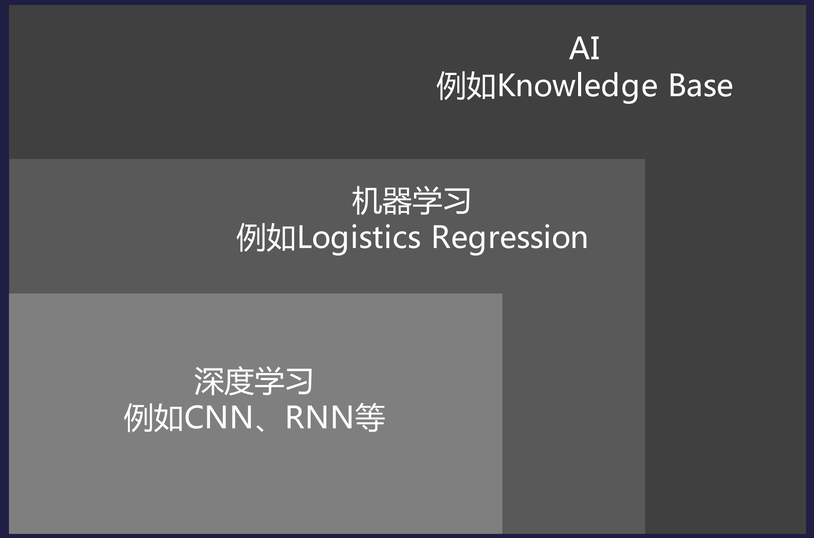


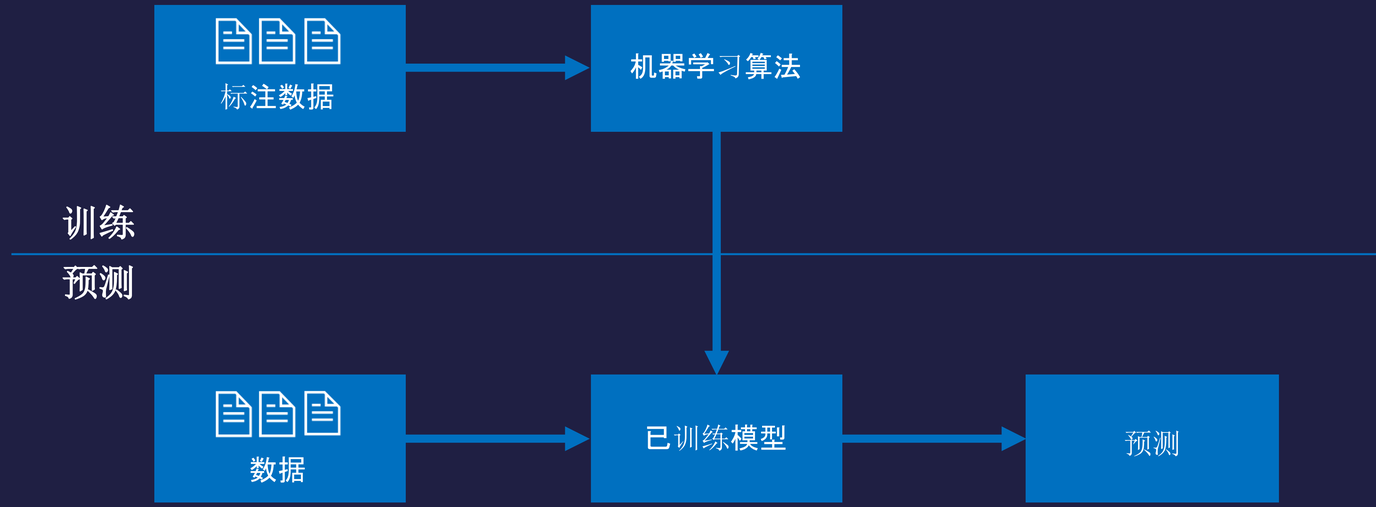
02

深度学习与 NLP

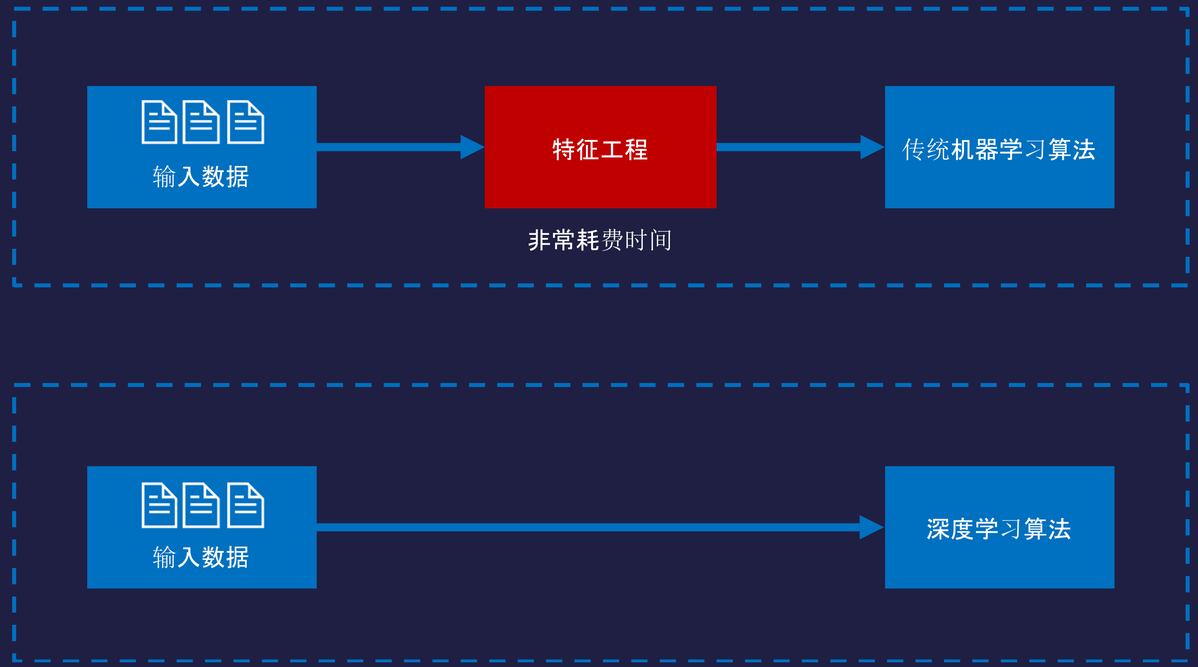
深度学习发展与应用

深度学习和机器学习



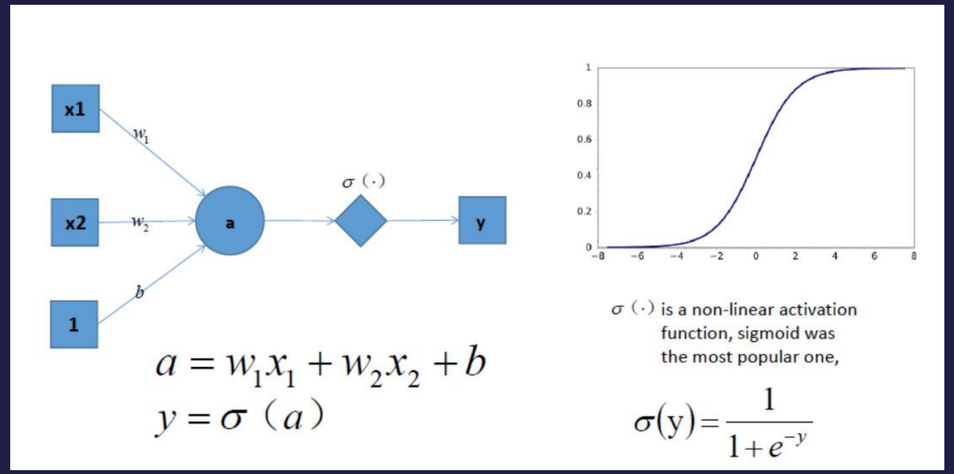


深度学习和传统机器学习

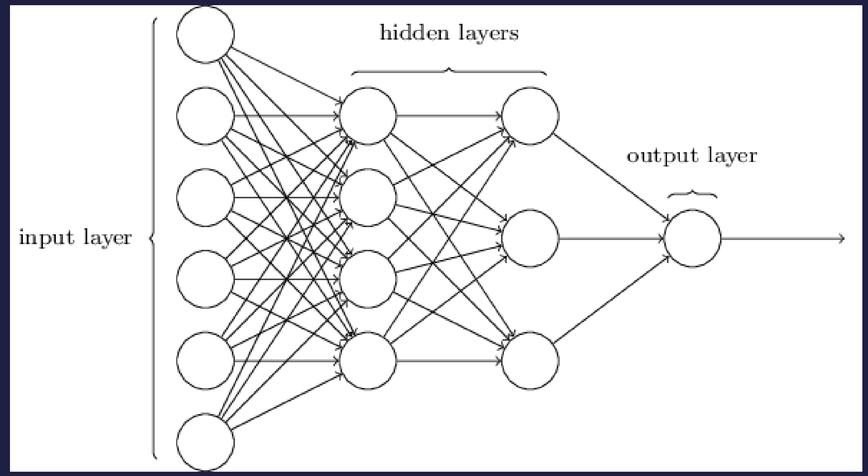


以文本分类过程举例，常见的
特征提取算法包括：
词频
TF-IDF
互信息
信息增益
期望交叉熵
主成分分析
...
特征工程需要手工寻找特征，
花费大量人力，特征的好坏往往决定最终结果

深度学习基础结构



基础神经元结构



多个神经元连接组成神经网络

字词表示

one-hot表示

计算机 [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]

电脑 [0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]

服务器 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ...]

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]

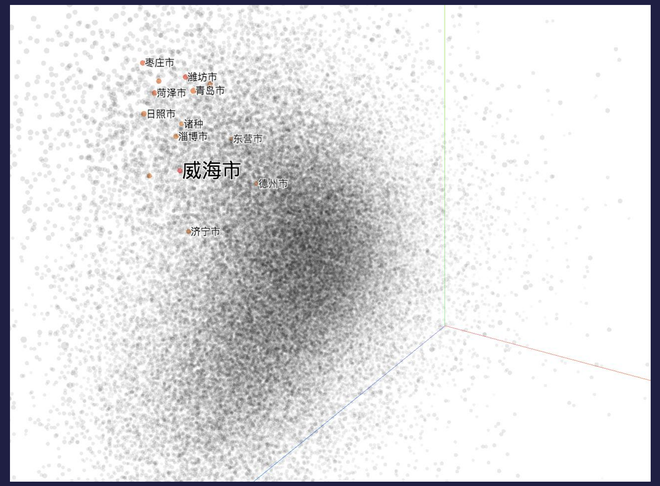
高维，稀疏，正交，无法计算语义相关性

字词表示

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

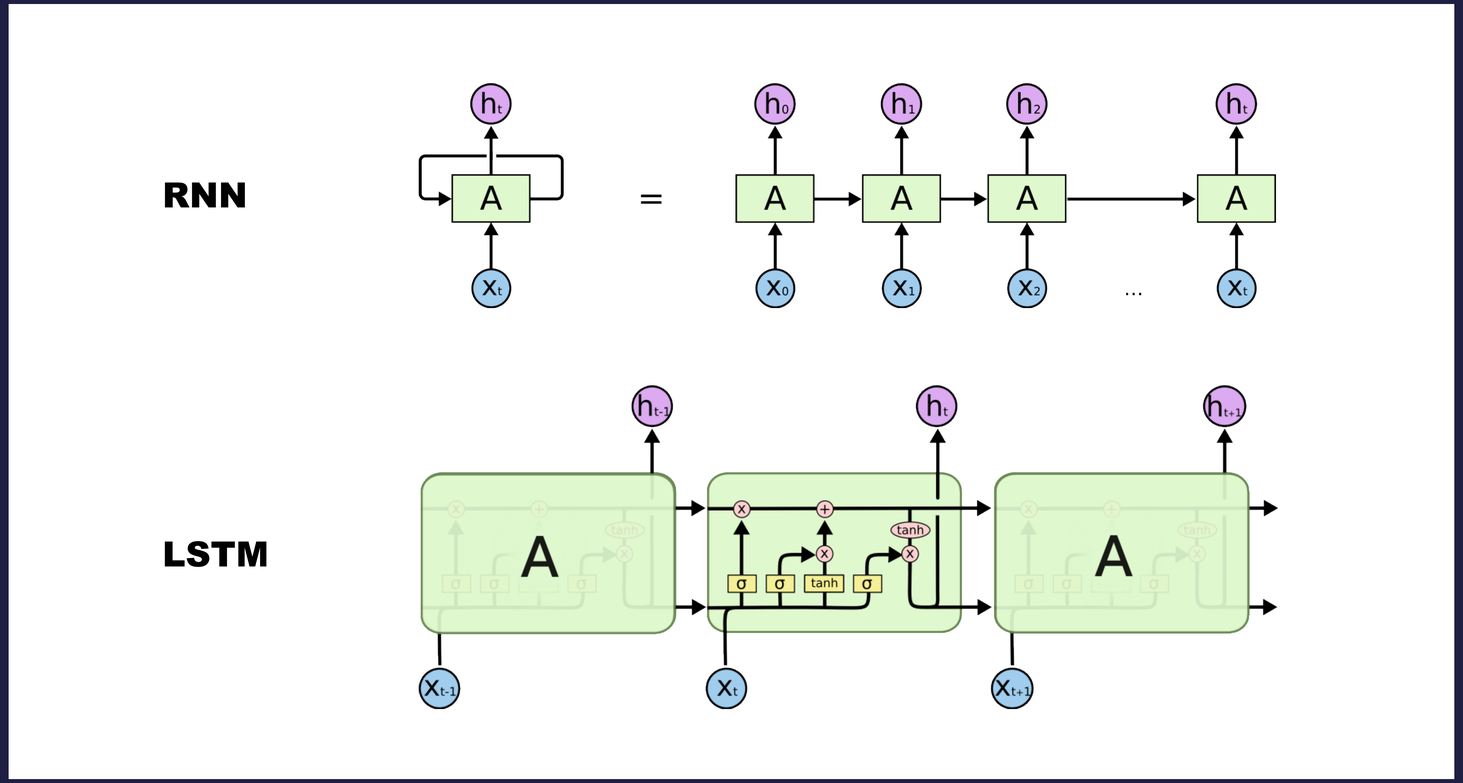
潍坊市　[-0.9602200985, 0.8771957159, 1.0565081835, ... 4.1443724632, -4.1823129654, -0.2311971784]

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



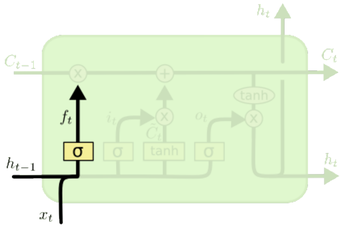
潍坊市	0.363
枣庄市	0.424
菏泽市	0.441
青岛市	0.486
泰安市	0.487
德州市	0.491
日照市	0.492
聊城市	0.492
济宁市	0.497
滕州市	0.504
淄博市	0.504
东营市	0.507

RNN与LSTM



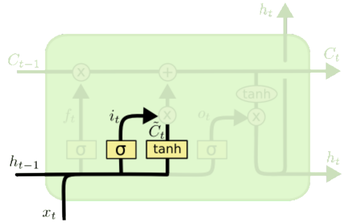
LSTM原理

1，单元状态丢弃



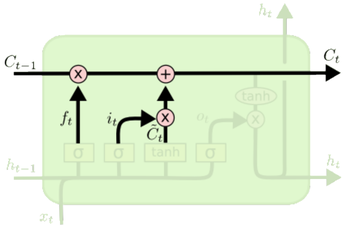
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

2，新信息选择



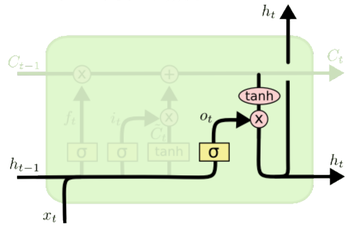
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

3，单元状态更新



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

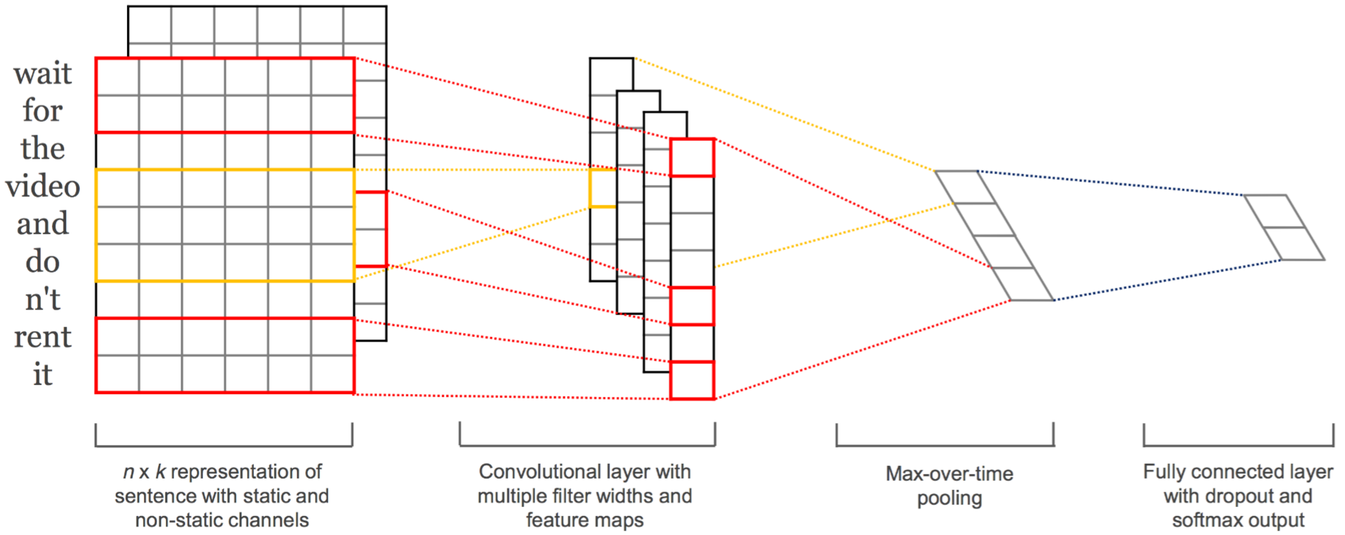
4，确定输出



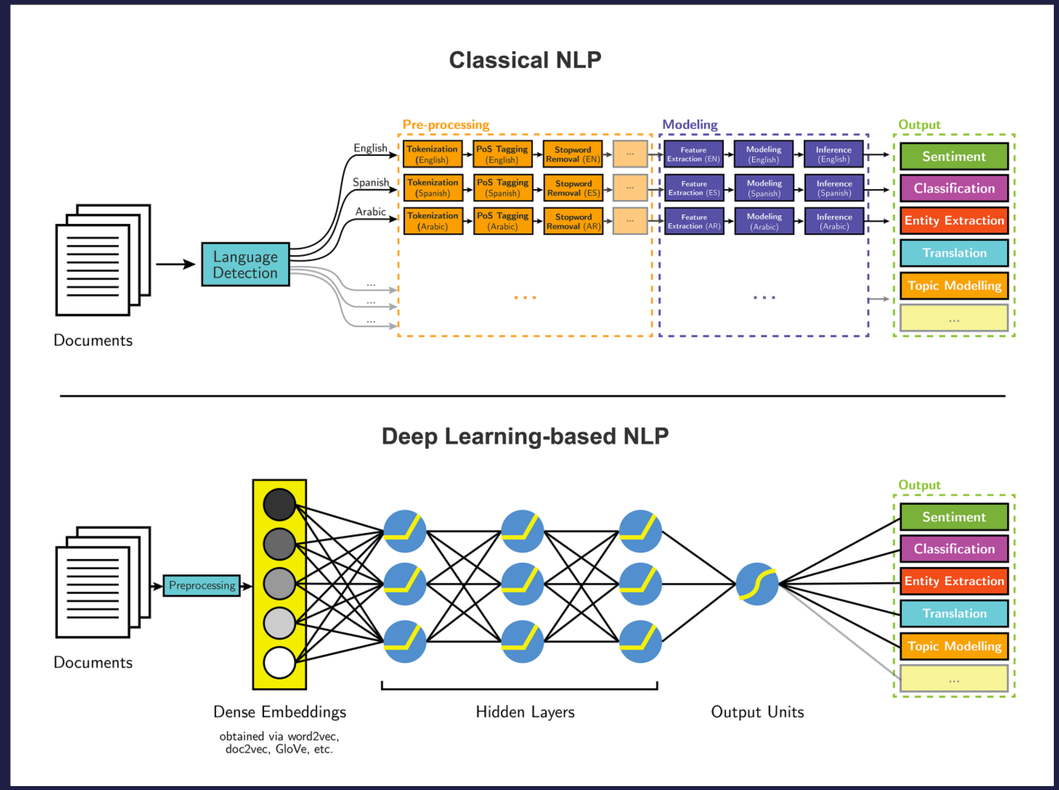
$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

CNN原理

- 单层CNN结构，GPU并行加速训练快
- 词向量解决传统方法词one-hot编码稀疏问题
- max-pooling会忽略文档结构信息



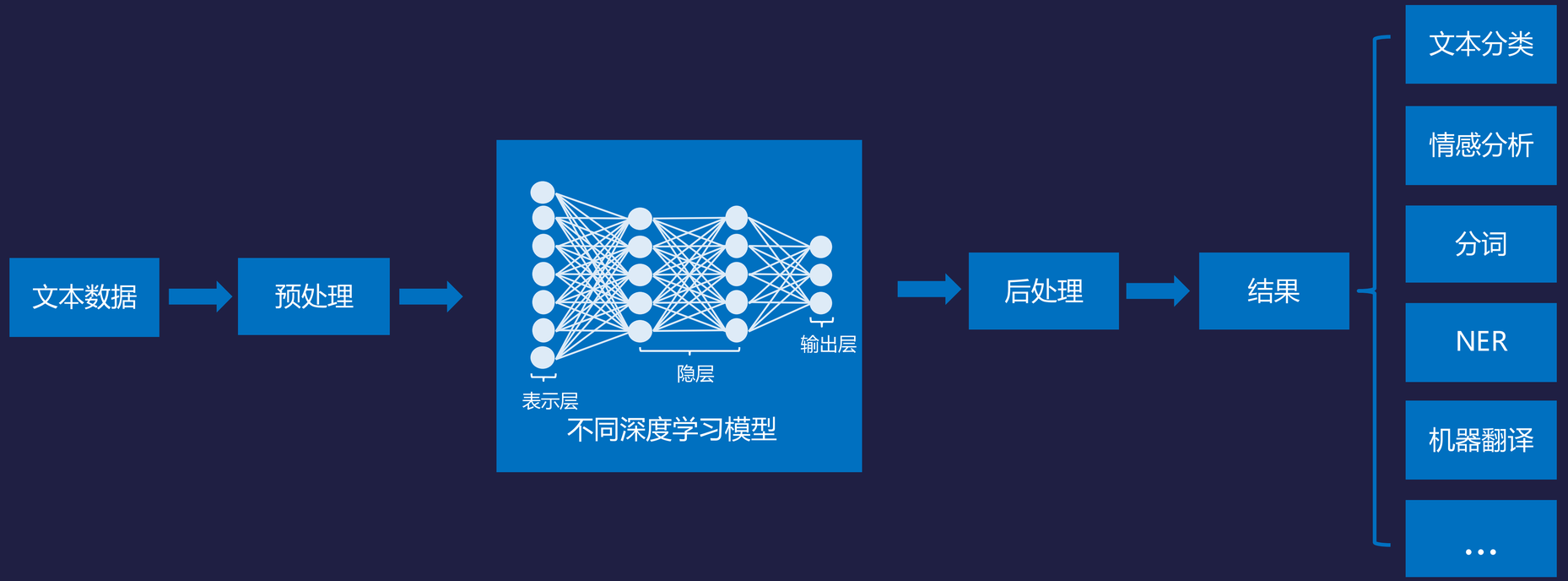
使用深度学习解决 NLP 问题



03

深度学习用于各类型文本应用的实践方法

文本挖掘各种类型应用的处理框架



文本分类

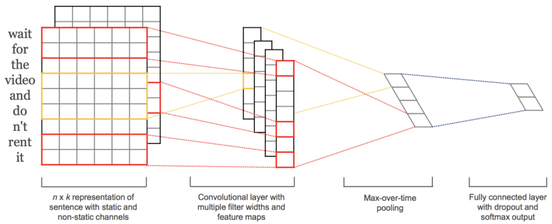
传统机器学习

- 选择分类器（朴素贝叶斯，SVM，KNN，LR，决策树）
- 特征工程构造特征
- 不同领域定制优化成本高
- 常需要分类算法融合提升效果

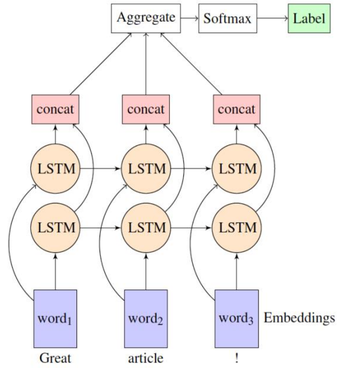
深度学习（CNN，RNN等）

- 端到端，无需大量特征工程
- 框架通用性好，满足多领域需求
- 可以使用非监督语料训练字词向量提升效果

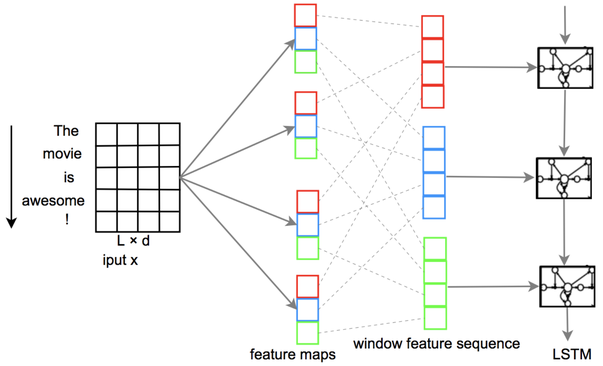
文本分类



CNN



RNN



CLSTM

Hierarchical Attention Network

- 考虑了文本结构信息
- 词->句子->篇章
- Attention结构
- 部分可解释（词句对分类的贡献大小）

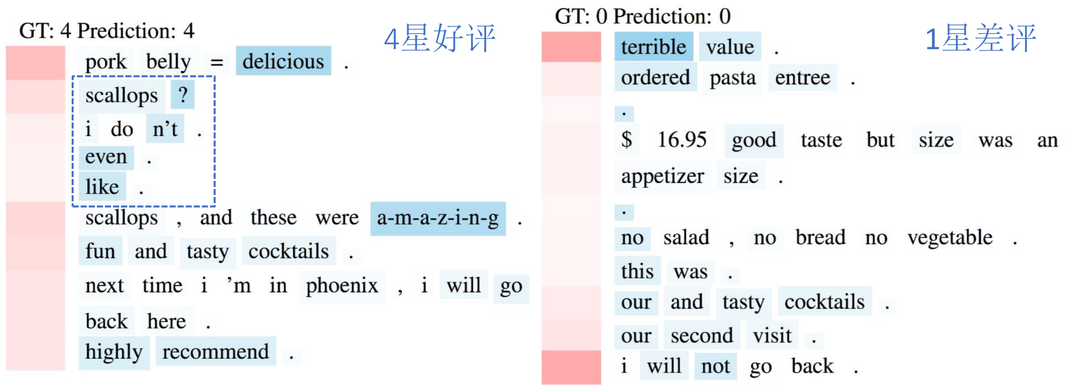
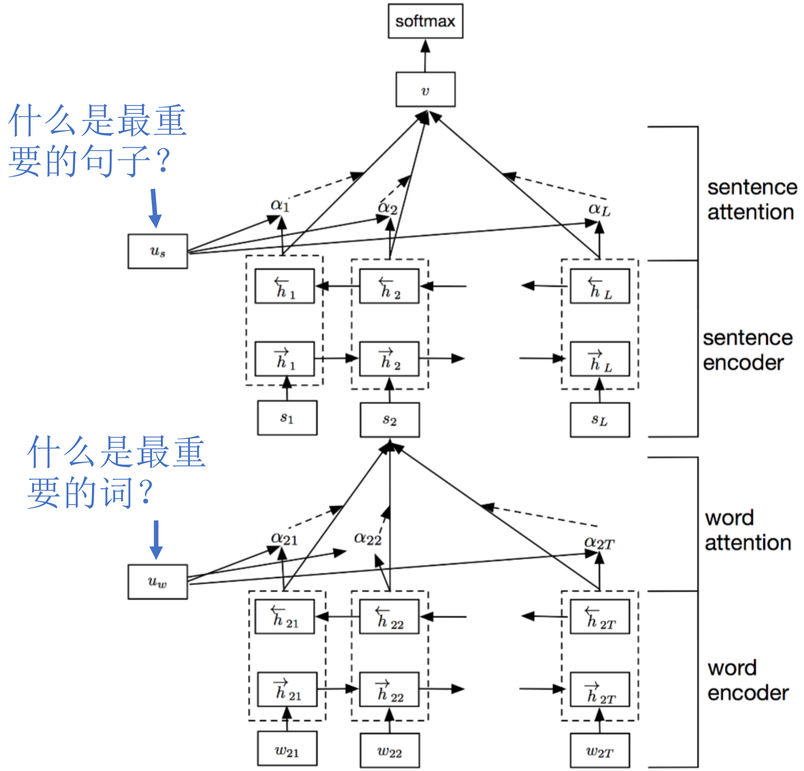


Figure 5: Documents from Yelp 2013. Label 4 means star 5, label 0 means star 1.

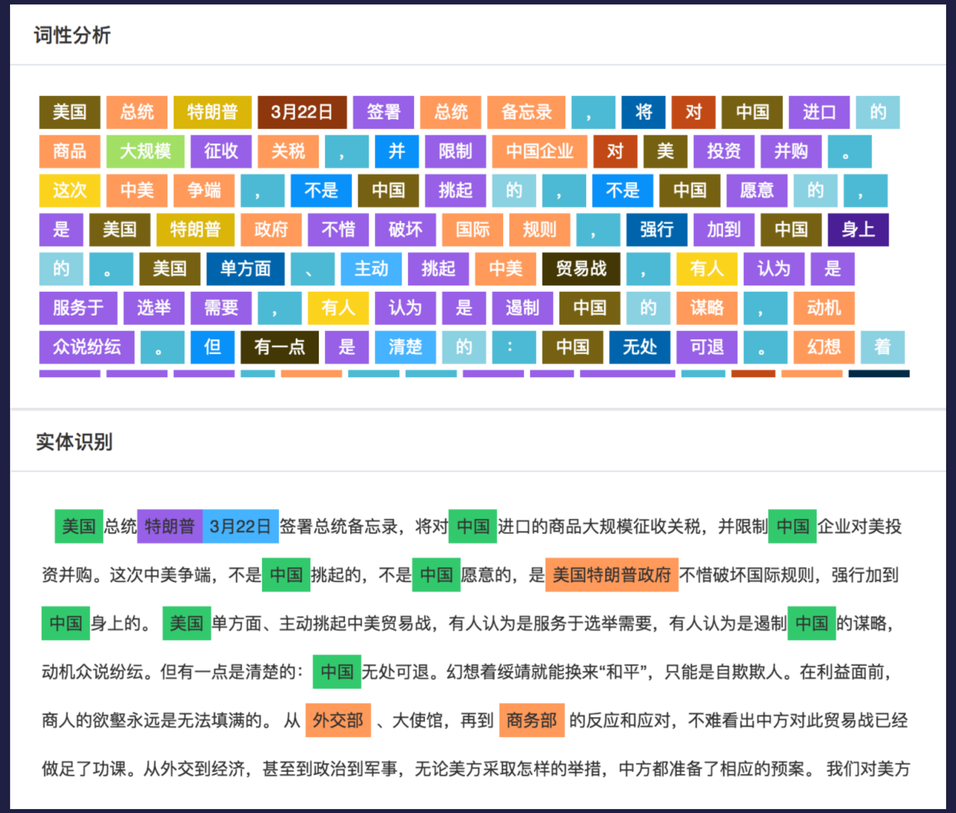


序列标注

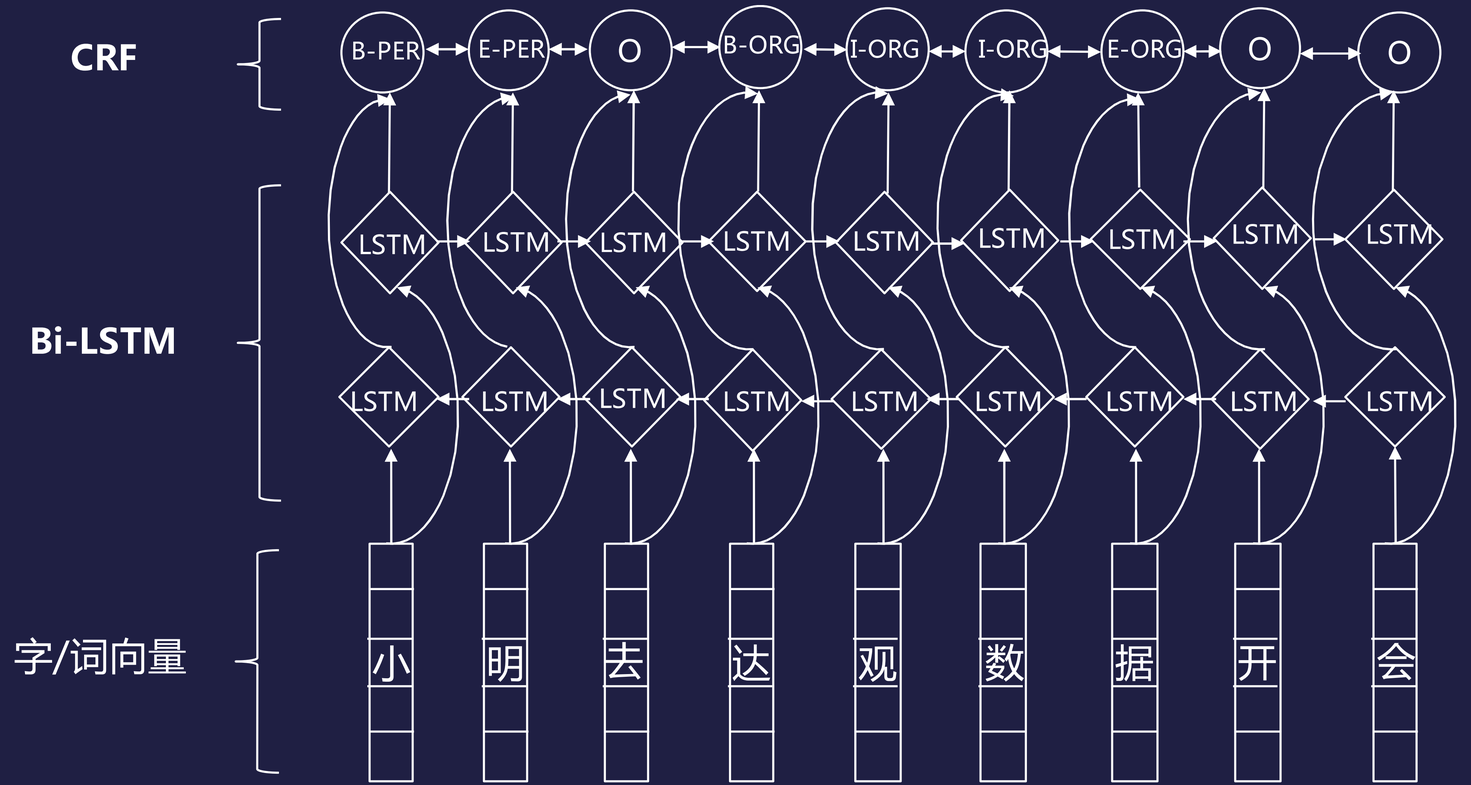
信息抽取➔4种类型分类
时间序列分析

他	来	自	达	观	数	据
S	B	E	B	M	M	E

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



序列标注



04

达观数据文本挖掘的实践经验

文本挖掘的一些常见应用需求

风险智能审核功能

文档智能抽取功能

错误智能纠正功能

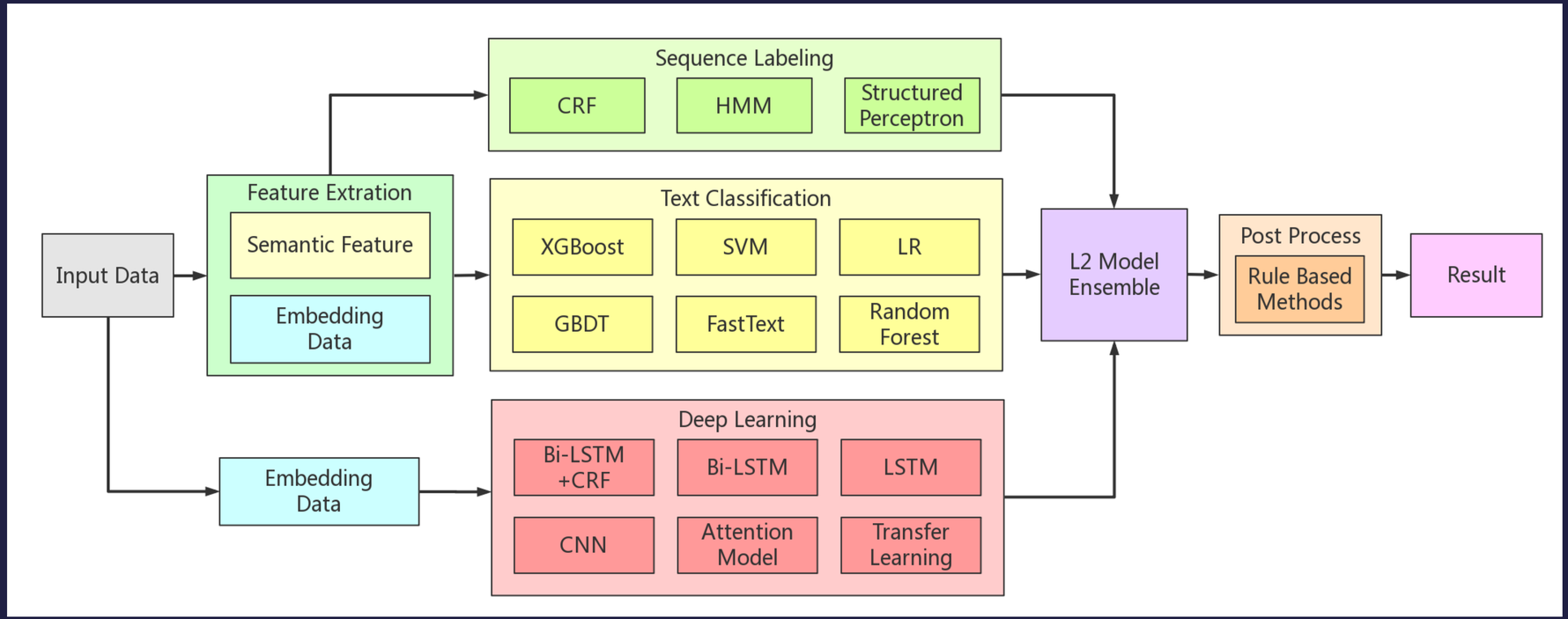
文档智能比对功能

达观智能文档审阅平台

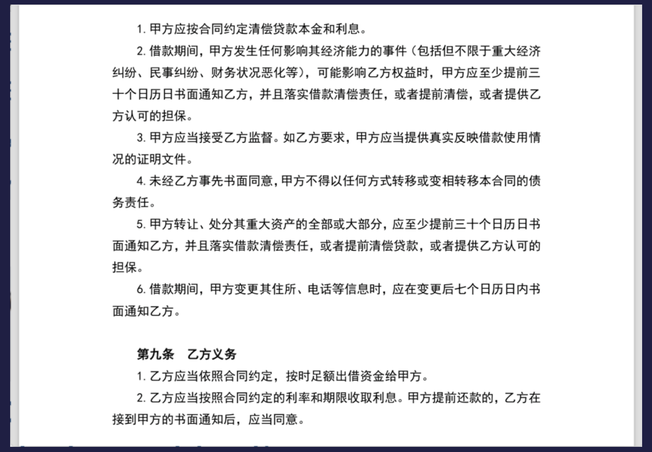
常见应用场景

- 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
- ……
- 更多场景可定制开发

智能文档审阅：核心抽取算法



智能文档审阅：段落分析



1. 甲方应按合同约定清偿贷款本金和利息。

2. 借款期间，甲方发生任何影响其经济能力的事件（包括但不限于重大经济纠纷、民事纠纷、财务状况恶化等），可能影响乙方权益时，甲方应至少提前三十个日历日书面通知乙方，并且落实借款清偿责任，或者提前清偿，或者提供乙方认可的担保。

3. 甲方应当接受乙方监督。如乙方要求，甲方应当提供真实反映借款使用情况的证明文件。

4. 未经乙方事先书面同意，甲方不得以任何方式转移或变相转移本合同的债务责任。

5. 甲方转让、处分其重大资产的全部或大部分，应至少提前三十个日历日书面通知乙方，并且落实借款清偿责任，或者提前清偿贷款，或者提供乙方认可的担保。

6. 借款期间，甲方变更其住所、电话等信息时，应在变更后七个日历日内书面通知乙方。

第九条　乙方义务

1. 乙方应当依照合同约定，按时足额出借资金给甲方。

2. 乙方应当按照合同约定的利率和期限收取利息。甲方提前还款的，乙方在接到甲方的书面通知后，应当同意。

PDF格式文本数据丢失段落信息

2. vim content.txt (vim)

```
1.甲方应按合同约定清偿贷款本金和利息。
2.借款期间，甲方发生任何影响其经济能力的事件(包括但不限于重大经济
纠纷、民事纠纷、财务状况恶化等)，可能影响乙方权益时，甲方应至少提前三
十个日历日书面通知乙方，并且落实借款清偿责任，或者提前清偿，或者提供乙
方认可的担保。
3.甲方应当接受乙方监督。如乙方要求，甲方应当提供真实反映借款使用情
况的证明文件。
4.未经乙方事先书面同意，甲方不得以任何方式转移或变相转移本合同的债
务责任。
5.甲方转让、处分其重大资产的全部或大部分，应至少提前三十个日历日书
面通知乙方，并且落实借款清偿责任，或者提前清偿贷款，或者提供乙方认可的 担保。
6.借款期间，甲方变更其住所、电话等信息时，应在变更后七个日历日内书
面通知乙方。
第九条 乙方义务
1.乙方应当依照合同约定，按时足额出借资金给甲方。
2.乙方应当按照合同约定的利率和期限收取利息。甲方提前还款的，乙方在
接到甲方的书面通知后，应当同意。
~
~
~
```

使用深度学习进行段落分析

4. vim content_new.txt (vim)

```
1.甲方应按合同约定清偿贷款本金和利息。
2.借款期间，甲方发生任何影响其经济能力的事件(包括但不限于重大经济纠纷、民事纠纷、财务状况恶化>
等)，可能影响乙方权益时，甲方应至少提前三十个日历日书面通知乙方，并且落实借款清偿责任，或者提>
前清偿，或者提供乙方认可的担保。
3.甲方应当接受乙方监督。如乙方要求，甲方应当提供真实反映借款使用情况的证明文件。
4.未经乙方事先书面同意，甲方不得以任何方式转移或变相转移本合同的债务责任。
5.甲方转让、处分其重大资产的全部或大部分，应至少提前三十个日历日书面通知乙方，并且落实借款清偿
责任，或者提前清偿贷款，或者提供乙方认可的 担保。
6.借款期间，甲方变更其住所、电话等信息时，应在变更后七个日历日内书面通知乙方。
第九条 乙方义务
1.乙方应当依照合同约定，按时足额出借资金给甲方。
2.乙方应当按照合同约定的利率和期限收取利息。甲方提前还款的，乙方在接到甲方的书面通知后，应当同
意。
~
~
~
~
~
~
~
~
```

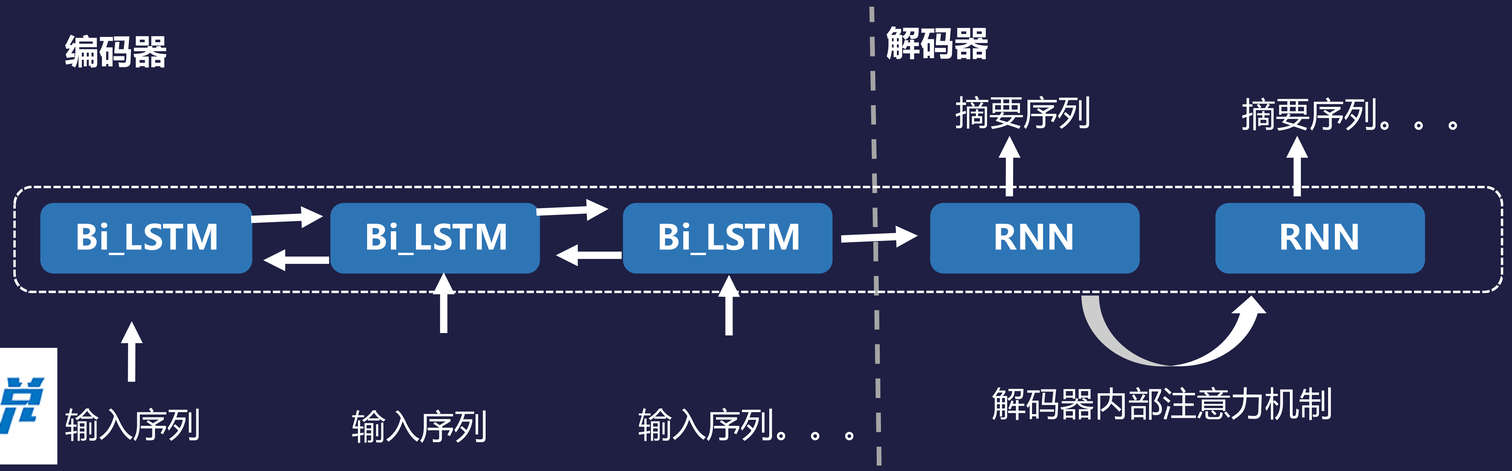
生成式摘要

生成式摘要的深度学习网络基本结构

- 编码器/解码器结构，都是神经网络结构
- 输入的原文经过编码器变成向量
- 解码器从向量里面提取关键编码信息，组合成生成式摘要

深度学习内部注意力机制的引入

- 内部注意力机制在解码器里面做
- 关注已生成词，解决长序列摘要生成时，个别字词重复出现的问题

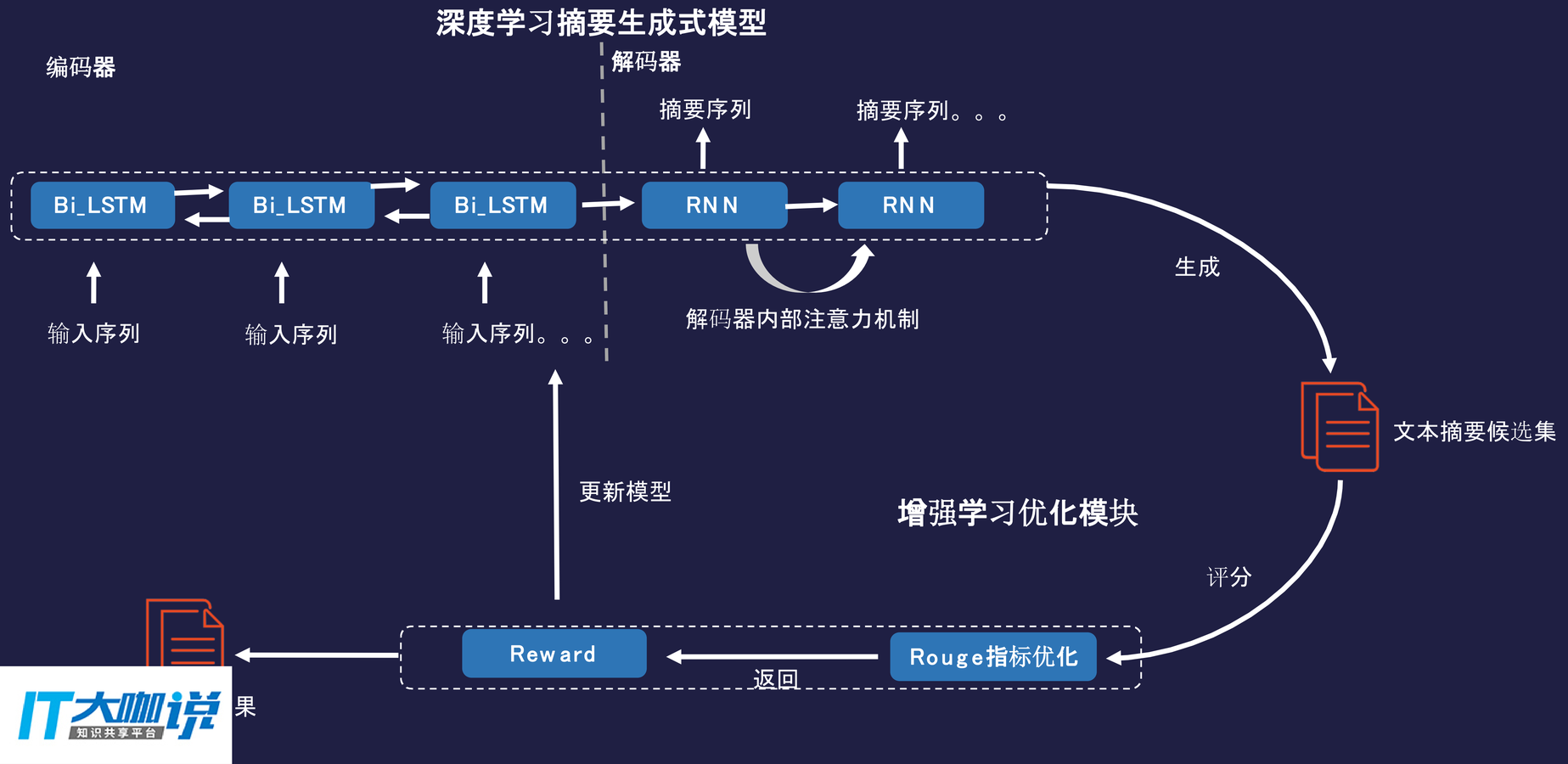


生成式摘要

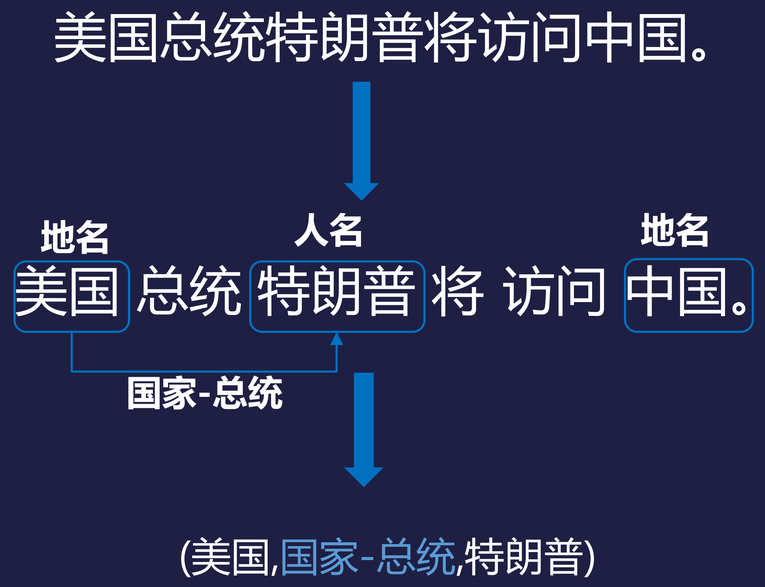
强化学习和深度学习相结合的学习方式

- 最优化词的联合概率分布：MLE（最大似然），有监督学习。在这里生成候选的摘要集。
- ROUGE指标评价：不可导，无法采用梯度下降的方式训练，考虑强化学习，鼓励reward高的模型，通过给与反馈来更新模型。最终训练得到表现最好的模型。

生成式摘要



知识图谱关系抽取：联合学习方法



难点：结构复杂

知识图谱关系抽取：基于深度学习

基于参数共享的方法

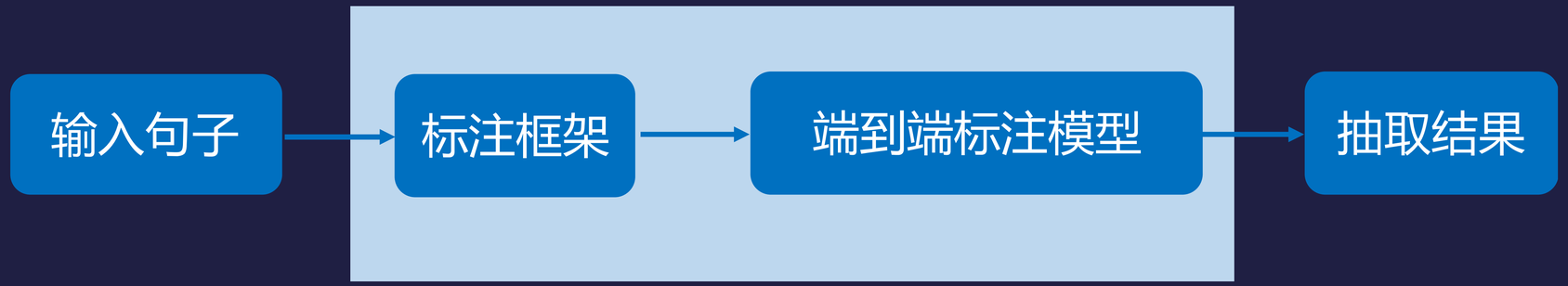
对于输入句子通过共用的 word embedding 层，然后接双向的 LSTM 层来对输入进行编码。然后分别使用一个 LSTM 来进行命名实体识别（NER）和一个 CNN 来进行关系分类（RC）。

基于联合标注的方法

把原来涉及到序列标注任务和分类任务的关系抽取完全变成了一个序列标注问题。然后通过一个端对端的神经网络模型直接得到关系实体三元组。

知识图谱关系抽取：基于联合标注

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



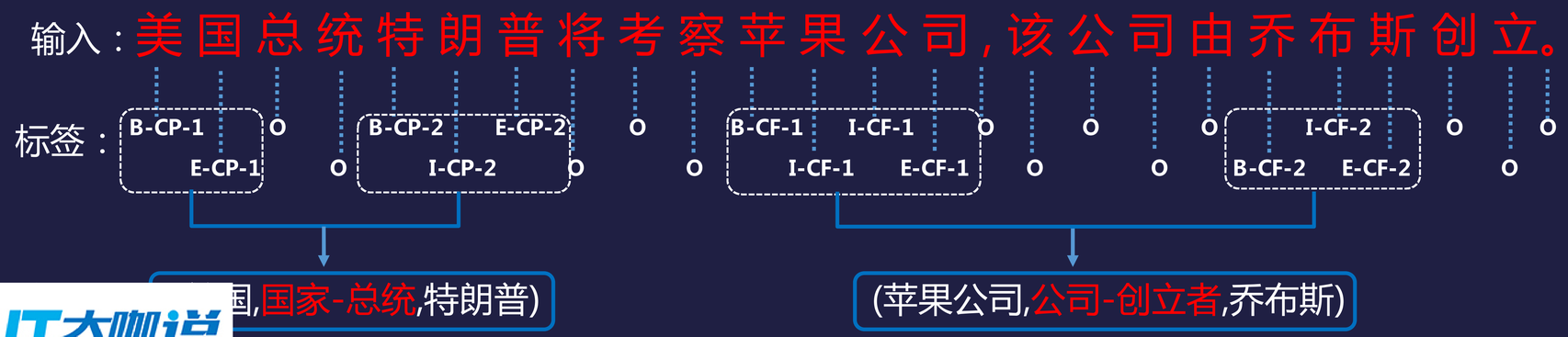
知识图谱关系抽取：基于深度学习

三类标签

- 单词在实体中的位置{B(begin),I(inside),E(end),S(single)}
- 关系类型{CF,CP,...}
- 关系角色{1(entity1),2(entity2)}

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，如果一个句子包含一个以上同一类型的关系，那么就采用就近原则来进行配对。

目前这套标签并不支持实体关系重叠的情况。



05

总结&QA

总结： 深度学习用于文本挖掘的优缺点

优点：

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

缺点：

1，小数据量效果不一定好
2，调参工作量有时不亚于特征工程
3，客户部署硬件环境限制

总结：一些实践经验

1，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

2，初始阶段可以使用传统机器学习模型快速尝试，作为baseline版本

3，疑难问题使用端到端的方式也许会有惊喜

4，不断尝试...

悟有所值
Ucan Afternoon Tea

THANKS

UCLOUD 专业云计算服务商

IT大咖说 知识共享平台