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

# 循环神经网络

AI DISCOVERY


# 目录

AI DISCOVERY



AI DISCOVERY


# 目录

AI DISCOVERY



AI DISCOVERY


# 语言处理技术

AI DISCOVERY

基本概念

词级分析

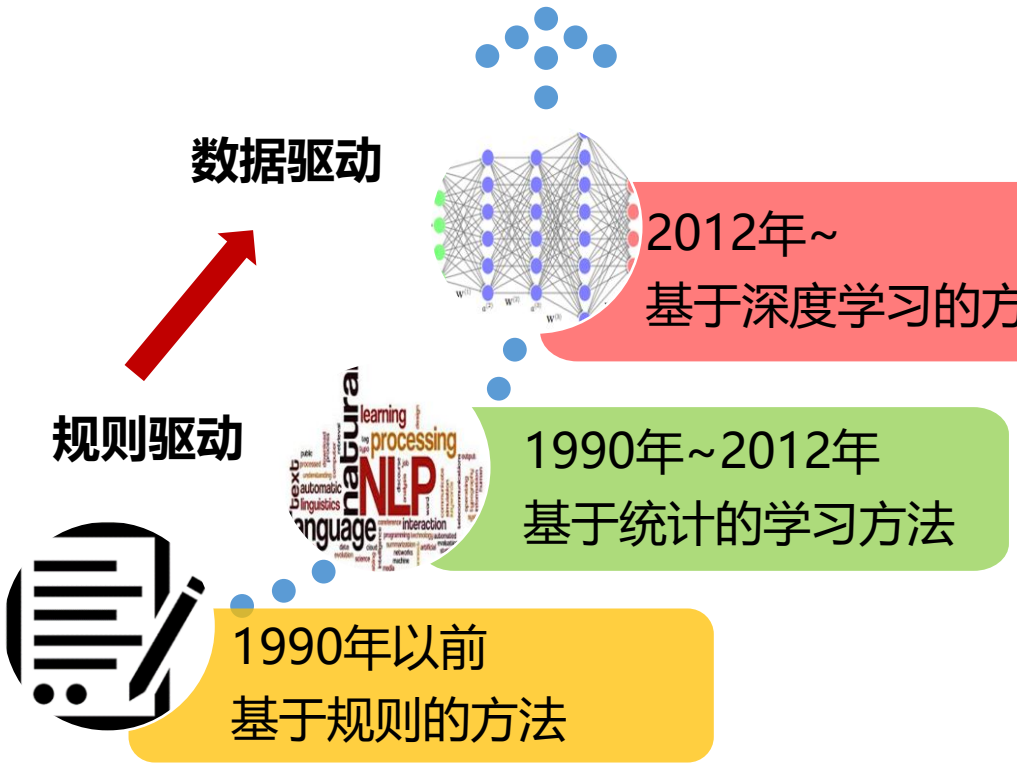
句章级分析

应用分析

AI DISCOVERY


# 什么是自然语言处理

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。自然语言处理技术发展经历了**基于规则的方法**、**基于统计学习的方法**和**基于深度学习的方法**三个阶段。自然语言处理由浅入深的四个层面分别是**形式**、**语义**、**推理**和**语用**，当前正处于由语义向推理的发展阶段。





# 研究内容

AI DISCOVERY

**词法(Lexical)学**：研究词的**词素(morphemes)构成、词性**等

- **形态(morphological) 分析**

employers —stemming→ employ + ~er + ~s

employers —lemmatize→ employer + ~s

词素（morphemes）→ 词 （word）？

⇓

词根、前缀、后缀、词尾

- **未登录词(out of vocabulary word)识别**

宅男，推特　模式口，新奥尔良

方舟子，罗姆尼，钓鱼岛

阿里巴巴...

- **中文分词(segmentation)**

你 家 用 的什么样的电脑？

ambiguities

家用 电脑。

你的 牙 刷 了吗？我的 牙刷 不见了。

把 手 举起来！这个 把手 是木制的。

- **词性标注(POS tagging)**

哥白尼说

哥白尼日心说

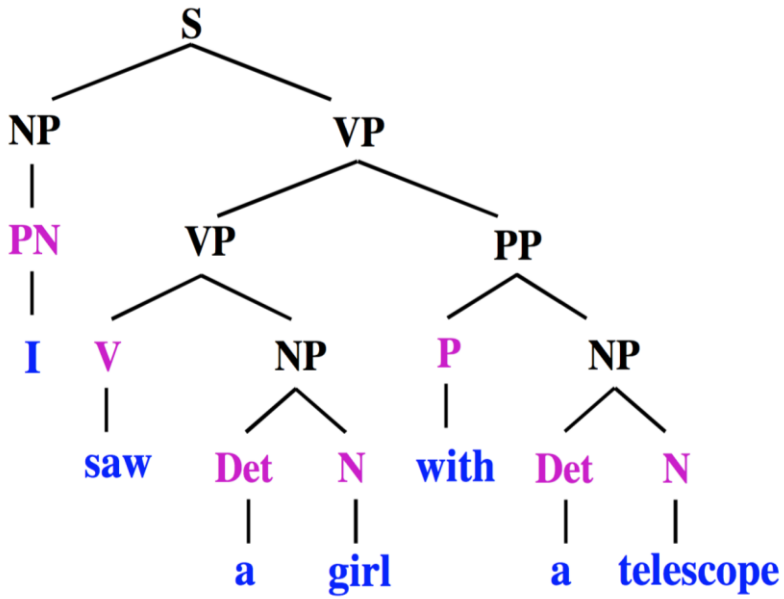
AI DISCOVERY

# 研究内容

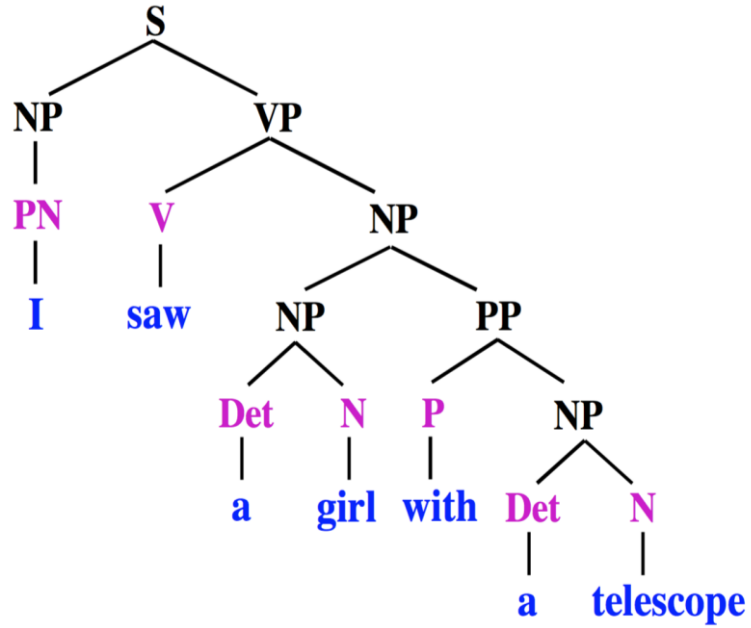
AI DISCOVERY

**句法(Syntax)学**：研究句子结构成分之间的相互关系和组成句子序列的规则

- I saw a girl with a telescope.




I saw （a girl with a telescope.）
我看见一个戴望远镜的女孩。




I saw （a girl） with a teles
我用望远镜看见一个女孩。

- “高个子”和“个子高”
  - ✓ 包含两个部分“高”和“个子”，但这两种表述具有不同的关系类型
  - ✓ **“高个子”**可以充当名词性元素，可以与动词“喜欢”构成一个更大的句子成分；
  - ✓ 而**“个子高”**是一个基本的具有主谓的句子结构。

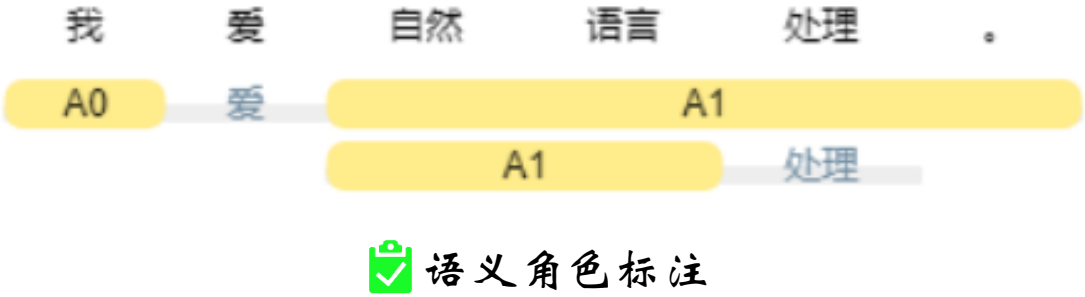

# 研究内容

AI DISCOVERY

**语义(Semantics)学**：研究如何从一个语句中词的意义，以及这些词在该语句的句法结构中的作用来推导出该语句的意义。

**这句话说了什么？含义**

- 苹果和梨 / 苹果和华为
- 火烧圆明园/ 驴肉火烧




“我”是施事者（“爱”动作的施行者），“自然语言处理”是受事者（“爱”动作的承受者）

AI DISCOVERY


# 研究内容

AI DISCOVERY

**语用(Pragmatics)学**：研究在不同上下文中的语句的应用，以及上下文对语句理解所产生的影响。

Lee打算怎么去商店?

- Lee decided to go to the store. He walked into the carport, only to find his **bike** have a flat tire.
- Jack took out a match. He **lit** a candle.

lit取什么含义？点燃 or 照亮

AI DISCOVERY


# 语言处理技术

AI DISCOVERY

基本概念

词级分析

句章级分析

应用分析

AI DISCOVERY


# 词法分析

AI DISCOVERY

**词**是自然语言中能够独立运用的最小单位，是语言信息处理的基本单位。

**词法分析**是对自然语言的形态(morphology) 进行分析，判断词的结构、类别和性质。

主要任务包括：

**自动分词（segmentation）**

**命名实体识别（Named Entity Recognition）**

**词性标注（Part-of-Speech tagging, POS tagging）**

AI DISCOVERY


# 自动分词

AI DISCOVERY

## 中文为什么要进行分词?

与大部分印欧语系的语言不同，汉语是以字为基本的书写单位，词语之间没有明显的区分标记，需要人为切分。

**例子**：我路过南京市长江大桥

我 / 路过 / 南京 / 市 / 长江 / 大桥

我 / 路过 / 南京市 / 长江大桥

} 不同的分词粒度

中文分词的核心任务是要确定词边界，将句子分解为最小意义单元，即将中文字序列转换为词序列。

中文分词是很多自然语言处理系统中的基础模块和首要环节。

AI DISCOVERY


# 中文分词

AI DISCOVERY

## 分词面临的主要问题

汉语分词困难重重

① **分词规范：**易受主观语感约束，产生不同的切分结果

② **歧义切分**

有些歧义无法在句子内部解决，
需要结合篇章上下文

- ✓ 交集歧义

  研究 / 生命 / 的 / 起源　　研究生 / 命 / 的 / 起源

- ✓ 组合歧义

  门 / 把 / 手 / 弄 / 坏 / 了　　门 / 把手 / 弄 / 坏 / 了

③ **未登录词识别**

包括中外人名、中国地名、机构组织名、事件名、货币名、缩略语、派生词、各种专业术语以及在不断发展和约定俗成的一些新词语。

确定词汇边界：PMI互信息，熵等

确定新词语义：领域词扩展等，LDA，word2vec

AI DISCOVERY


# 分词算法

AI DISCOVERY

## 基于规则的分词方法

简单易行，但歧义消解能力差

- **基本思想**：按照一定策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行匹配
- **主要方法**：最大匹配、逆向最大匹配、双向最佳匹配、逐词遍历

## 基于统计的分词方法

效果依赖于训练语料的规模和质量

- **基本思想**：上下文中相邻的字同时出现的次数越多，就越有可能构成一个词，字与字相邻出现的概率或频率能较好地反映成词的可信度。
- **主要方法**：N 元文法模型（N-gram）、隐马尔可夫模型（Hiden Markov Model，HMM）、最大熵模型（ME）、条件随机场模型（Conditional Random Fields，CRF）等

## 基于理解的分词方法

需要大量的语言知识和信息

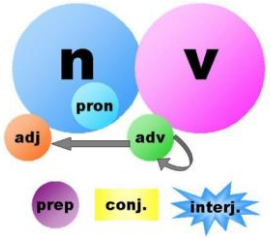
- **基本思想**：在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象，让计算机模拟人对句子的理解来进行分词。
- **主要方法**：专家系统分词、神经网络分词（LSTM,CNN）

AI DISCOVERY


# 词性标注

AI DISCOVERY

**词性标注(Part of speech tagging)**是为分词结果中的每个单词标注一个正确的词性，也即确定每个词是名词、动词、形容词或者其他词性的过程；主要任务是消除词性兼类歧义。




| 我_r | 毕业_v | 于_p | 北京_ns | 清华大学_ni | 。_wp |
|---|---|---|---|---|---|

在任何一种自然语言中，词性兼类问题都普遍存在。

**英语中：**

1) Time **flies** like an arrow.

2) I want you to **web** our annual report.

对 Brown 语料库的统计，55％词兼类。
《现代汉语八百词》兼类占 22.5％。

**汉语中：**

1) 形同音不同，如: 好(hao3, 形容词)、 好(hao4, 动词)

2) 同形同音但意义不同，如: 会 (会议,名词)、会(能够、动词)

3) 具有典型意义的兼类词，如: 典型(名词或形容词)

4) 上述组合，如: 行(xing2, 动词/ 形容词；hang2, 名词/量词)

AI DISCOVERY


# 词性标注算法

AI DISCOVERY

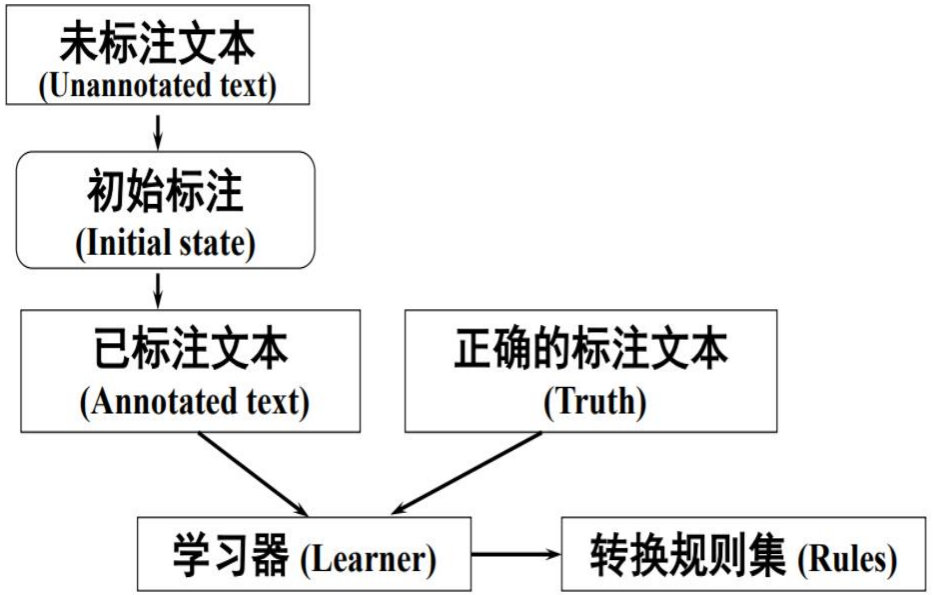
## 基于规则的方法：

- 根据词语的结构建立词性标注规则
  - ✓ 词缀(前缀、后缀)：绿油油（形容词）、一片片(数量词)
  - ✓ 重叠词规则：看看、瞧瞧、高高兴兴、热热闹闹 ...
- 基于机器学习的自动规则提取方法
  - ✓ 初始词性赋值；
  - ✓ 对比正确标注的句子,自动学习结构转换规则
  - ✓ 利用转换规则调整初始赋值

**基于统计模型的方法：** 最大熵、HMM、CRF

**综合方法：** ★

- ✓ 统计概率引导，辅以规则消歧




AI DISCOVERY


# 词法分析工具

AI DISCOVERY

**BosonNLP**：http://bosonnlp.com/dev/center

**IKAnalyzer**：http://www.oschina.net/p/ikanalyzer

**NLPIR**：http://ictclas.nlpir.org/docs

**SCWS中文分词**：http://www.xunsearch.com/scws/docs.php

结巴分词：https://github.com/fxsjy/jieba

盘古分词：http://pangusegment.codeplex.com/

庖丁解牛：https://code.google.com/p/paoding/

搜狗分词：http://www.sogou.com/labs/webservice/

腾讯文智：http://www.qcloud.com/wiki/API%E8%AF%B4%E6%98%8E%E6%96%87%E6%A1%A3

新浪云：http://www.sinacloud.com/doc/sae/python/segment.html

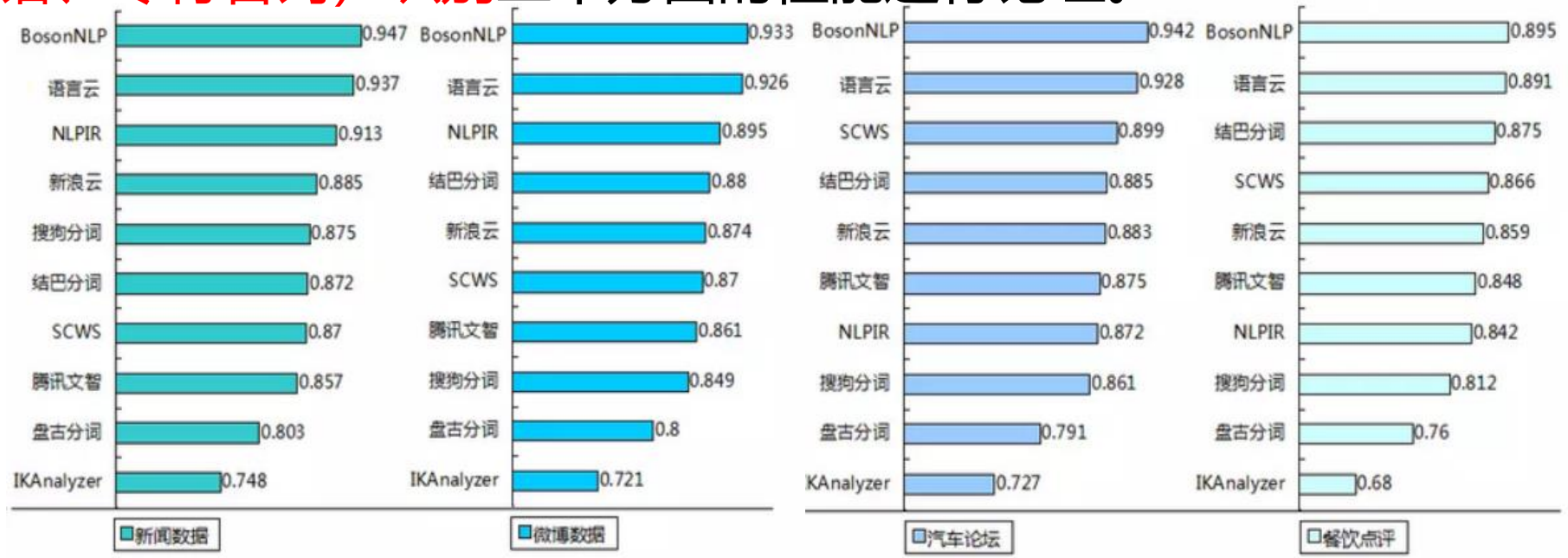
语言云：http://www.ltp-cloud.com/document

AI DISCOVERY

# 工具测评

AI DISCOVERY

**分词效果测评**：测试了以上11家中文分词引擎，在新闻、微博、论坛和点评四类数据上，综合分词准确度、歧义词切分、未登陆词（新涌现的通用词、专业术语、专有名词）识别三个方面的性能进行论证。




从分词精度来看，哈工大的语言云表现的稳定一直在第二，BostonNLP分词更好，一直在这个领域保持第一。

AI DISCOVERY


# 语言处理技术

AI DISCOVERY

基本概念

词级分析

句章级分析

应用分析

AI DISCOVERY

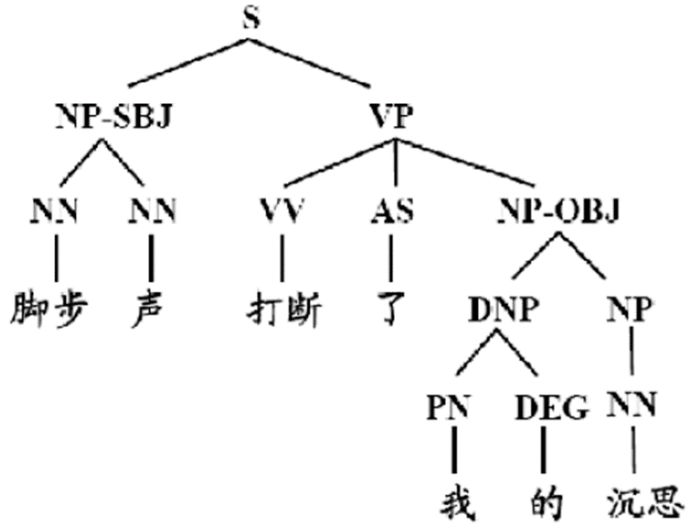
# 短语（句子）结构分析

AI DISCOVERY

**短语结构分析（phrase structure parsing）**对输入的单词序列判断其是否符合给定的语法，分析出合乎语法的句法结构。用树状结构表示，称为句法分析树。

**主要任务**

- 判断输入的字符串是否属于某种语言
- 消除句中的词法和结构等方面的歧义
- 分析句子内部结构，如成分构成、上下文关系等




**主要方法**：基于PCFG的短语结构分析

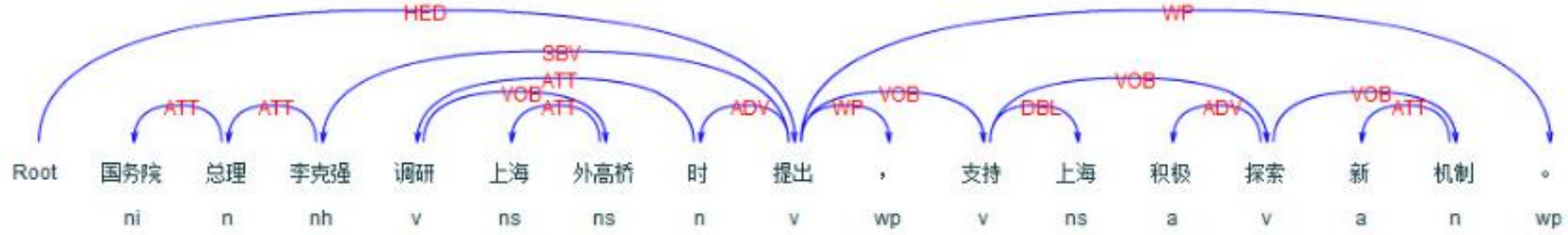
由于句法结构分析的语法集较为固定和呆板，目前的句法分析已经从句法结构分析转向依存句法分析。

AI DISCOVERY


# 依存句法分析



**依存句法分析（dependency parsing）**通过分析语言单位内成分之间的依存关系揭示其句法结构。 直观来讲，主要是识别句子中的主谓宾、定状补这些语法成分，并分析各成分之间的关系。




从分析结果中我们可以看到，句子的核心谓词为“提出”，主语是“李克强”，提出的宾语是“支持上海…”，“调研…时”是“提出”的（时间）状语，“李克强”的修饰语是“国务院总理”，“支持”的宾语是“探索 新机制”。有了上面的句法分析结果，我们就可以比较容易的看到，“提出者”是“李克强”，而不是“上海”或“外高桥”，即使它们都是名词，而且距离“提出”更近。


# 依存句法分析



## 基于图的依存句法分析方法：

- ✓ 将依存句法分析问题看成从完全有向图中寻找最大生成树的问题
- ✓ 通常采用基于动态规划的解码算法，也有一些学者采用柱搜索(beam search)来提高效率。

## 基于转移的依存句法分析方法：

- ✓ 将依存树的构成过程建模为一个动作序列，将依存分析问题转化为寻找最优动作序列的问题。

## 多模型融合的依存句法分析方法：

- ✓ 融合以上两种方法，如stacked learning


# 语义分析

AI DISCOVERY

**词汇级语义分析**指如何理解某个词汇的含义，包含两个方面：

**1. 词义消歧。**如何自动获悉某个词存在着多种含义，以及假设已知某个词具有多种含义，如何根据上下文确认其含义。

**基于词典的方法**：**英文词义标注语料库**： Semcor(普林斯顿大学)、Senseval 评测语料库等；**中文词义标注语料库**：哈尔滨工业大学和北京大学分别基于 HowNet 和北大“现代汉语语义词典”标注了词义消歧语料库。

**基于机器学习的方法**：综合待消解词的词汇、句法、语义等特征，结合机器学习算法进行。

**2. 词表示。**如何表示并学习一个词的语义，以便计算机能够有效地计算两个词的相似度。

目前流行的词义表示方式是**词嵌入（Word Embedding**，又称词向量）。基本想法是：通过训练将某种语言中的每一个词映射成一个固定维数的向量，将所有这些向量放在一起形成一个词向量空间，而每一向量则可视为该空间中的一个点，引入“距离”概念，则可以根据词之间的距离来判断它们之间的（词法、语义上的）相似性。

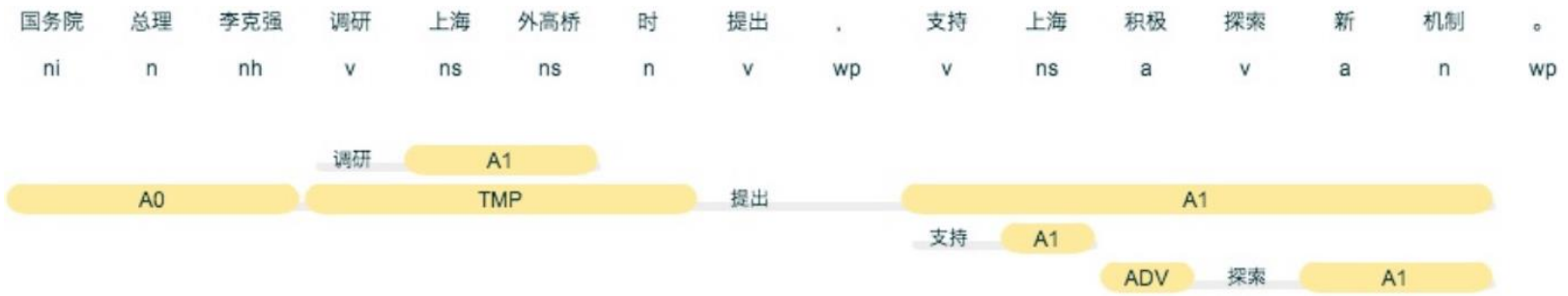
AI DISCOVERY

# 语义分析

AI DISCOVERY

**句子级语义分析**试图根据句子的句法结构和句中词的词义等信息，推导出能够反映这个句子意义的某种形式化表示。

**1. 浅层语义分析。**语义角色标注(Semantic Role Labeling，简称 SRL)是一种浅层语义分析方法，任务是找出句子中谓词的相应语义角色成分，包括核心语义角色（如施事者、 受事者等）和附属语义角色（如地点、时间、方式、原因等）。




其中有三个谓词提出，调研和探索。以探索为例，积极是它的方式（一般用ADV表示），而新机制则是它的受事（一般用A1表示）。

AI DISCOVERY


# 语义分析



**2. 深层语义分析。**主要围绕着句子中的谓词，为每个谓词找到相应的语义角色将整个句子转化为某种形式化表示。

以下是GeoQuery数据集中的句子及对应的一阶谓词逻辑语义表达式：

中文：列出 在 科罗拉多 州 所有 的 河流
英文：Name all the rivers in Colorado
语义表达式：answer(river(loc_2(stateid('colorado'))))

**基于知识库的方法：**知识库（如DBpedia、Freebase、Yoga 等），通过三元组等形式记录了一系列的事实。语义分析通过某种转换技术，将句子分析为一系列知识库中已定义的元组，并构成实体关系图。


# 语言处理技术

AI DISCOVERY

基本概念

词级分析

句章级分析

应用分析

AI DISCOVERY


# 文本分类与聚类

AI DISCOVERY

随着互联网的高速发展，海量文本数据不断产生，如何面对浩如烟海的数据进行分类、组织和管理，已经成为一个具有重要用途的研究课题，广受学术界和工业界关注。


**文本分类(Text Classification)**根据给定文档的内容或主题，自动分配预先定义的类别标签。

**文本聚类(Text Clustering)**根据文档之间的内容或主题相似度，将文档集合划分成若干个子集，每个子集内部的文档相似度较高，而子集之间的相似度较低。


**应用场景**：新闻自动分类、电子商务评价分类 、垃圾邮件识别等。


# 文本分类与聚类

AI DISCOVERY

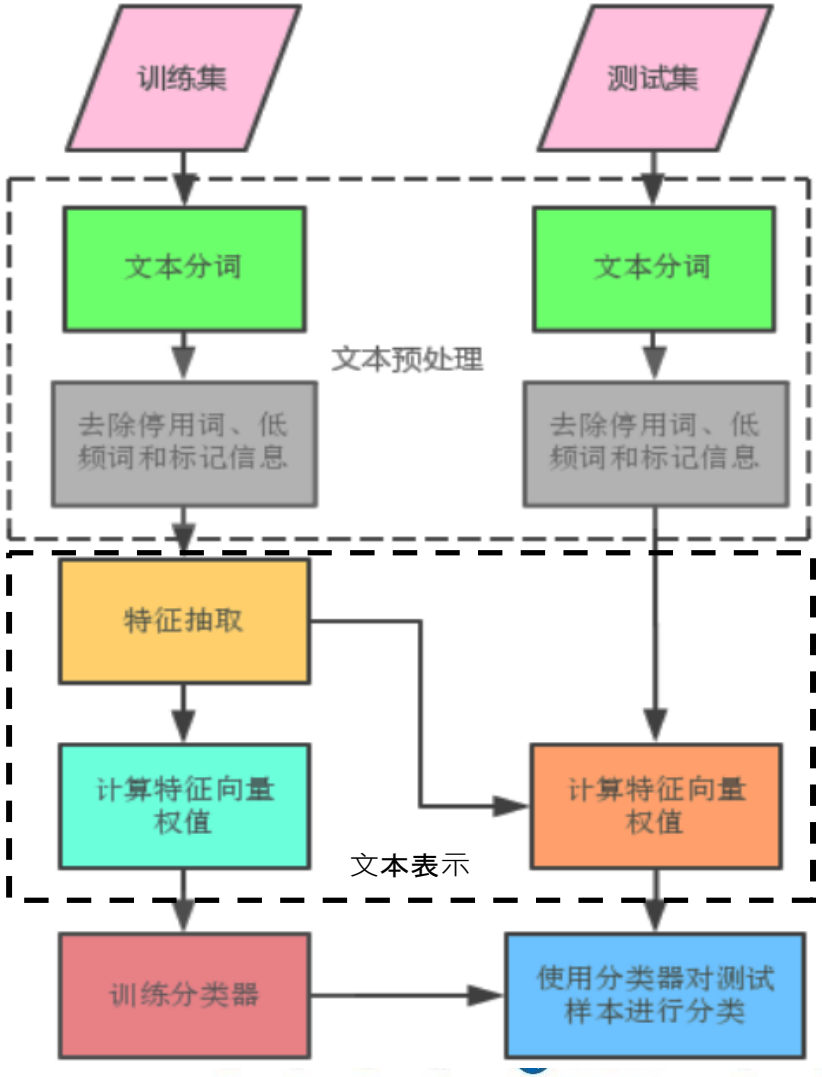
## 文本特征提取和降维

大规模文本中可能出现的词项非常多，但并不是所有词项都可以作为文本特征。为选取有效文本特征，降低特征空间维度，提高分类聚类的效果与效率，主要采用以下特征降维方法。

**特征选择(Feature Selection)**：文档频率、互信息、信息增益、$\chi^2$统计量等

**特征转换(Feature Transformation)**：主成分分析(PCA)、线性判别分析(LDA)

**话题分析(Topic Analysis)**：潜在语义分析(LSA)、基于概率的潜在语义分析(PLSA)、隐狄利克雷分布(LDA)




图为有监督分类方法的一般过程，无监督分类/聚类方法将训练部分去掉即可


# 文本分类与聚类

AI DISCOVERY

## 文本分类模型

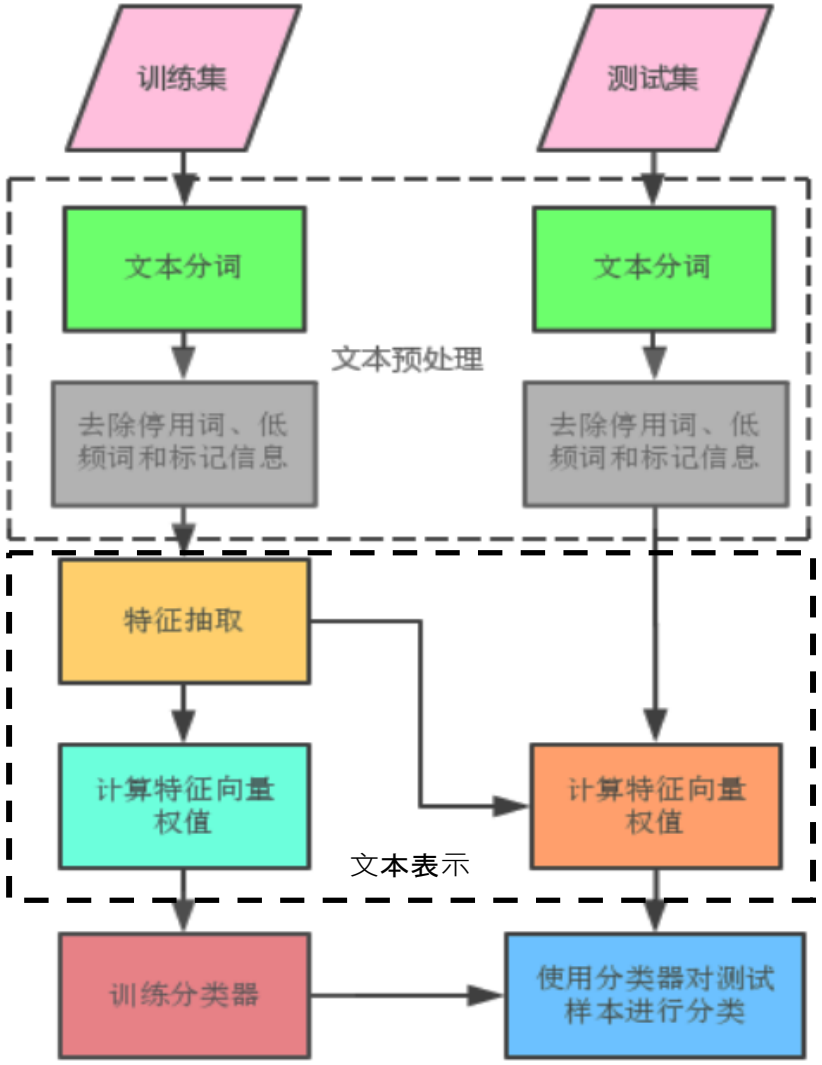
**基于机器学习的分类**：朴素贝叶斯 (Naive Bayes)、支持向量机(SVM)、最大熵分类器

**基于神经网络的方法**：多层感知机(MLP) 、卷积神经网络(CNN) 、循环神经网络(RNN)

## 文本聚类模型

**基于距离的聚类**：通过相似度函数计算语义关联度，然后根据语义关联度进行聚类，如K-means

**基于概率模型的聚类**：假设每篇文章是所有主题上的概率分布，典型的主题模型包括 PLSA 和 LDA 等




图为有监督分类方法的一般过程，无监督分类/聚类方法将训练部分去掉即可

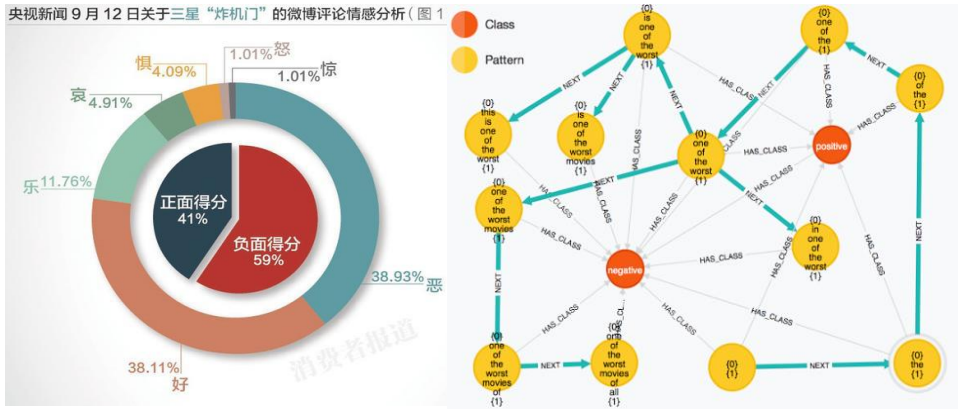

# 情感分析

AI DISCOVERY

**情感分析(sentiment analysis)**是指根据文本所表达的含义和情感信息将文本划分成褒义、贬义或多种类型，是对文本作者倾向性和观点、态度的划分，有时也称倾向性分析（opinion analysis）。

按粒度可分为词汇级、句子级和篇章级的情感分析，核心任务主要包含观点性及倾向性识别、观点要素抽取等任务。


营销分析




舆情分析


淘宝评论分析


# 情感分析

AI DISCOVERY

**基于词典的情感分析方法**通过制定一系列的情感词典和规则，对文本进行拆句、分析及匹配词典，计算情感值进行文本的情感倾向判断。

① 对文本进行句子拆解
② 分析句子中出现的词语并按照情感词典匹配
③ 处理否定逻辑及转折逻辑
④ 计算整句情感词得分（根据词语不同，极性不同，程度不同等因素进行加权求和）
⑤ 根据情感得分输出句子情感倾向性

如果是对篇章或者段落级别的情感分析任务，按照具体的情况，可以以对每个句子进行单一情感分析并融合的形式进行，也可以先抽取情感主题句后进行句子情感分析，得到最终情感分析结果。

1. 常见英文情感词库：MPQA、sentiWordNet等；
2. 常见中文情感词库：知网、台湾大学的情感极性词典；
3. 几种情感词典构建方法：基于互信息、图模型等方法。

**基于机器学习的情感分析方法**将情感分析作为一个分类问题来处理，基本流程与文本分类一致，采用支持向量机（SVM）、深度学习（CNN，RNN，LSTM）等模型方法。

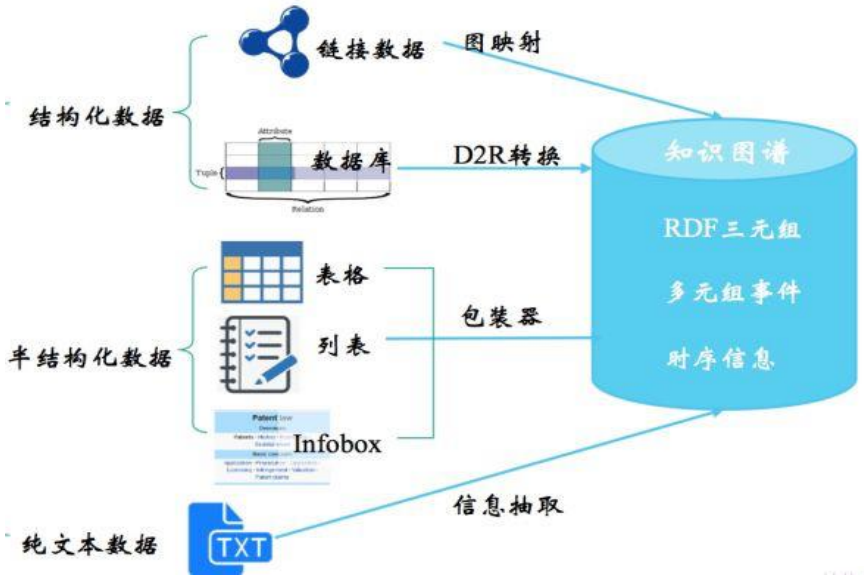
AI DISCOVERY

# 信息抽取

AI DISCOVERY

互联网的迅速普及和发展，信息资源极大丰富，但“信息过载”问题日趋严重，迫切需要快速、准确获取信息的技术手段，信息抽取技术应运而生。

信息抽取是指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。





**主要研究内容：**

- ✓实体识别与抽取
- ✓实体消歧
- ✓关系抽取
- ✓事件抽取

AI DISCOVERY


# 信息抽取——命名实体识别

AI DISCOVERY

**实体识别与抽取包括命名实体识别和开放域实体识别。**

**命名实体识别(Named Entity Recognition)**是识别文本中具有特定意义的实体，如人名、机构名、地名等专有名词和有意义的时间等，是信息检索、问答系统等技术的基础任务

小明在夏威夷度假。
命名实体：小明——人名、夏威夷——地名


## 难点解析

- 命名实体类型多样，数量众多，各种复姓、国外译名、网络红人、虚拟人物和昵称等实体不断涌现，难以建立大而全的标注语料库
- 命名实体构成结构复杂，长度不一，没有严格的规律可以遵循
- 在不同文化、领域、场景下，命名实体的外延有差异，界限不清晰，如“人名也经常出现在地名和组织名称中”

AI DISCOVERY


# 信息抽取——命名实体识别

AI DISCOVERY

**开放域实体识别**指给定某一类别的实体实例，从网页中抽取同一类别其他实体实例，特点在于不限定实体类别，不限定目标文本。

给定<中国,美国,俄罗斯>(称为“种子”)，找出其他国家<德国,英国,法国……>

**基本思路**：种子词与目标词在网页中具有**相同或者类似的上下文**（包括网页结构和上下文）。因此需要首先利用种子词提取模板，随后利用模板提取更多同类实体。

**主要方法**

- 基于Query Log的抽取方法：通过分析种子在查询日志中的上下文学得模板，再利用模板找到同类别实例。构造候选与种子上下文向量，计算相似度。
- 基于Web Page的抽取方法：利用种子、网页、模板、候选构造一个图，综合考虑网页和模板的质量，使用Random Walk算法为候选打分并排序。
- 融合多个数据源的抽取方法：采用网页、查询日志、维基百科多种数据源，针对不同数据源,选取不同特征分别进行实例扩展,对结果进行融合，针对不同数据源选取不同的模板和特征，使用不同特征计算候选的置信度。

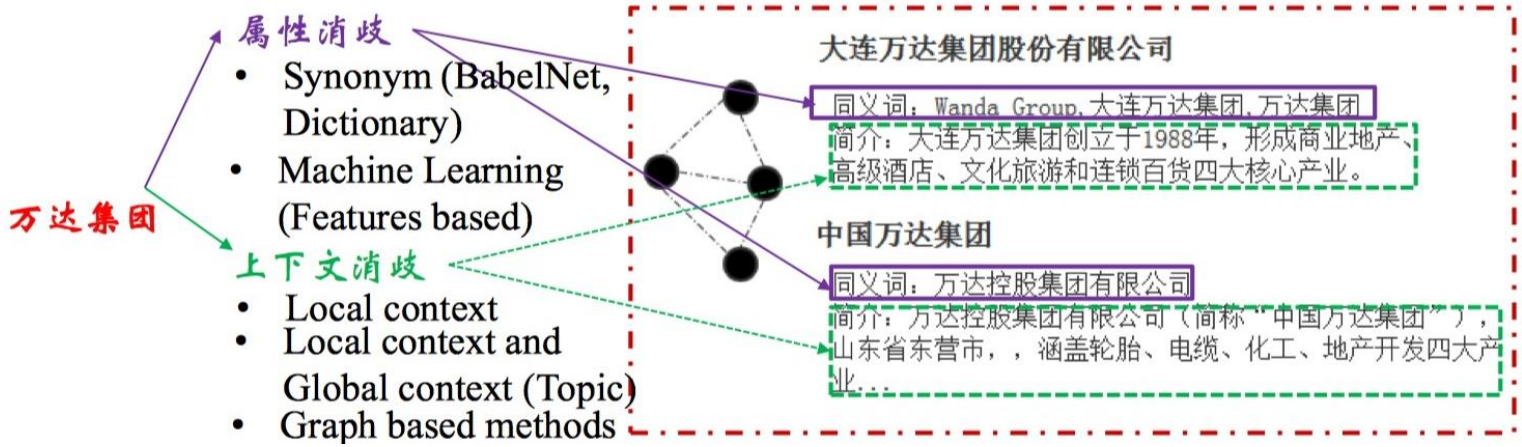
AI DISCOVERY

# 信息抽取——实体消歧

AI DISCOVERY

**实体消歧**是在非结构化文档中，由于书写风格和上下文的需要，同一个命名实体实体可能包含多种形式的表达（多对一），同时文档中的一个名词可能从字面意思上对应多种命名实体（一对多）。

中国证券网讯（记者 王雪青）中国证券网记者今日获悉，万达集团的文化产业版图将再添世界级新军——传奇影业，具体收购情况或于下周二正式发布。




**基本思路**：利用实体间的交互、及上下文信息进行消歧，可采用图、概率生成模型，主题模型，深度学习等方法。

# 信息抽取——关系抽取

AI DISCOVERY

**关系抽取**指检测和识别文本中实体之间的语义关系。关系抽取的输出通常是一个三元组(实体1,关系类别,实体2)，表示实体1 和实体2之间存在特定类别的语义关系。

北京是中国的首都、政治中心和文化中心 。
(中国, 首都, 北京)，(中国, 政治中心, 北京)和(中国, 文化中心, 北京)

- 基于模板的方法：基于触发词/字符串、基于依存句法
- 监督学习：机器学习、深度学习
- 半监督/无监督学习：Bootstrapping、Distant supervision、Unsupervised learning from the web


AI DISCOVERY


# 信息抽取——事件抽取

AI DISCOVERY

**事件抽取**指的是从非结构化文本中抽取事件信息，主要包括时间、地点、事件元素角色等，并将其以结构化形式呈现出来的任务。主要任务包括触发词和事件元素的提取等。

<Type:Life, Subtype:Be-Born>
Person: 毛泽东
Time: 1893年
Place: 湖南湘潭

毛泽东 1893年 出生 于 湖南 湘潭。

“出生”是该事件的触发词，所触发的事件类别（Type）为Life，子类别（Subtype）为Be-Born。事件的三个组成元素“毛泽东”“1893年”“湖南湘潭”，分别对应着该类（Life/Be-Born）事件模板中的三个元素标签，即：Person、Time、Place。

## 主要方法

- 模式匹配方法：词汇-语法、词汇-语义、模板匹配、触发词扩展
- 基于机器学习的方法：SVM、最大熵、CNN、RNN
- 混合方法：模式匹配+机器学习

（MUC/ACE会议标注语料）

AI DISCOVERY

# 自动文摘

AI DISCOVERY

**自动文摘**是利用计算机按照某类应用自动地将文本（或文本集合）转换生成简短摘要的一种信息压缩技术

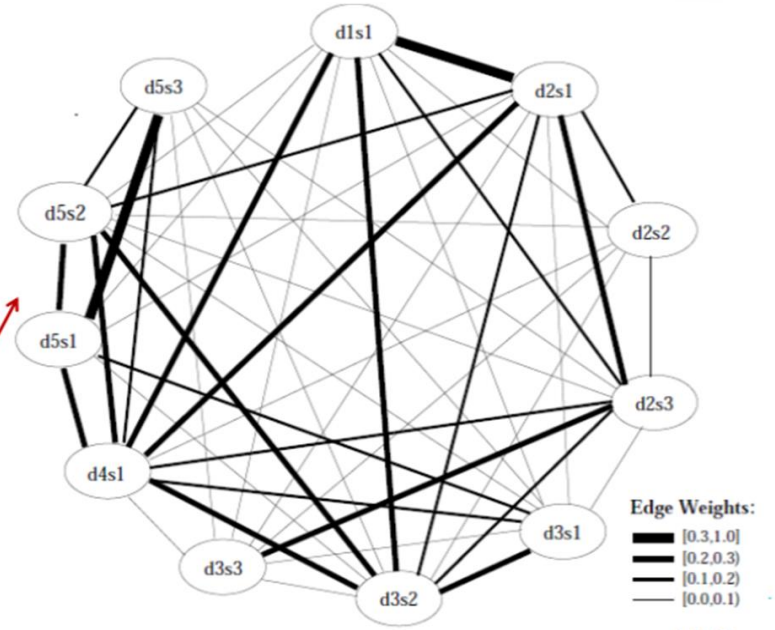
要求：信息量足、覆盖面广、冗余度低和可读性高。

- **抽取式摘要**
  - ✓ 从原文中抽取已有句子组成摘要（排序）
- **生成式摘要**
  - ✓ 改写或重新组织原文内容形成最终文摘


**理解式摘要抽取方法**

- G=(V, E)
- V：句子
- E：句间关系

PageRank算法：
计算每个句子
的重要性得分



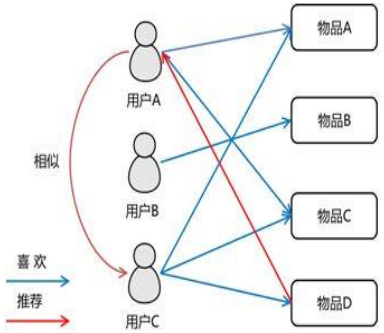

**基于图模型的方法**

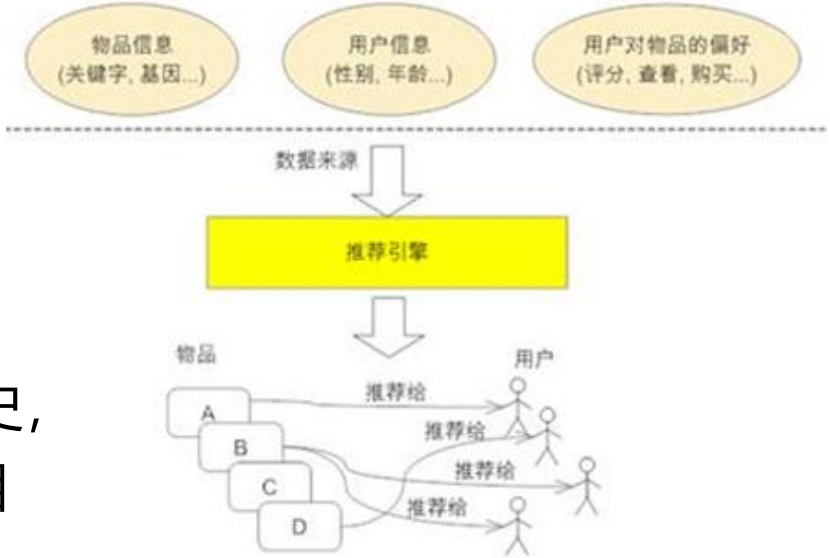

# 信息推荐

AI DISCOVERY

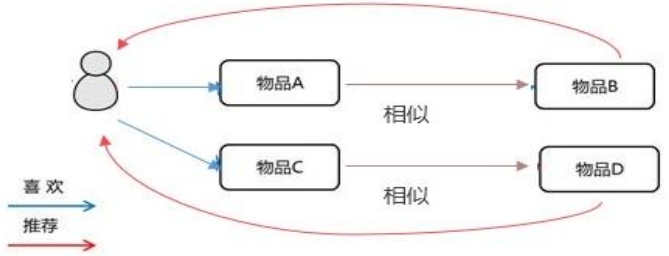
**信息推荐**是指根据用户的习惯、偏好或兴趣，从不断到来的大规模信息中识别满足用户兴趣的信息的过程。广泛应用在电子商务、电影和视频、音乐、社交网络、阅读、基于位置的服务、个性化邮件和广告等。

- **协同过滤推荐**
  - ✓ 基于用户的协同过滤
  - ✓ 基于物品的协同过滤




- **基于内容的推荐系统**
  - ✓ 使用元素的描述性属性进行推荐。如根据Sophie的听歌历史，推荐系统注意到她似乎喜欢乡村音乐，因此系统可以推荐相同或相似类型的歌曲
- **基于知识的推荐系统**
  - ✓ 基于顾客的需求和商品描述之间的相似度，或是对特定用户的需求使用约束来进行的，适用于物品购买频率很低的情况



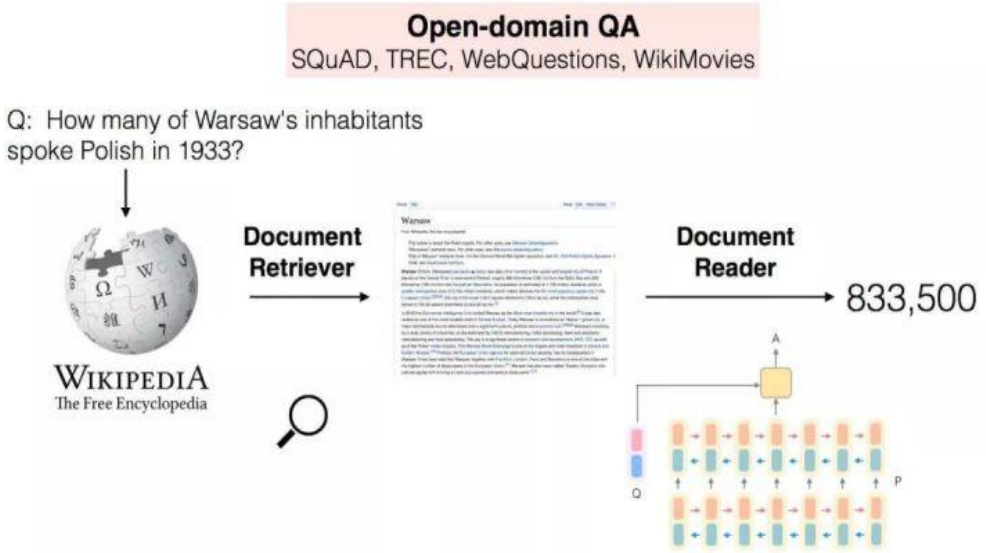
# 自动问答

AI DISCOVERY

**问答系统(question answering，简称QA)**是讨论如何从大规模真实文本中对指定的提问找出正确回答的技术，是集知识表示、信息检索、自然语言处理和智能推理等技术于一身的新一代搜索引擎。Web形式的问答网站、聊天机器人等。

按照答案的生成反馈机制，问答系统可划分为：

- 基于检索式的问答系统
- 基于生成式的问答系统





IBM Waston


微软小冰


Apple Siri


JD JIMI

聊天机器人

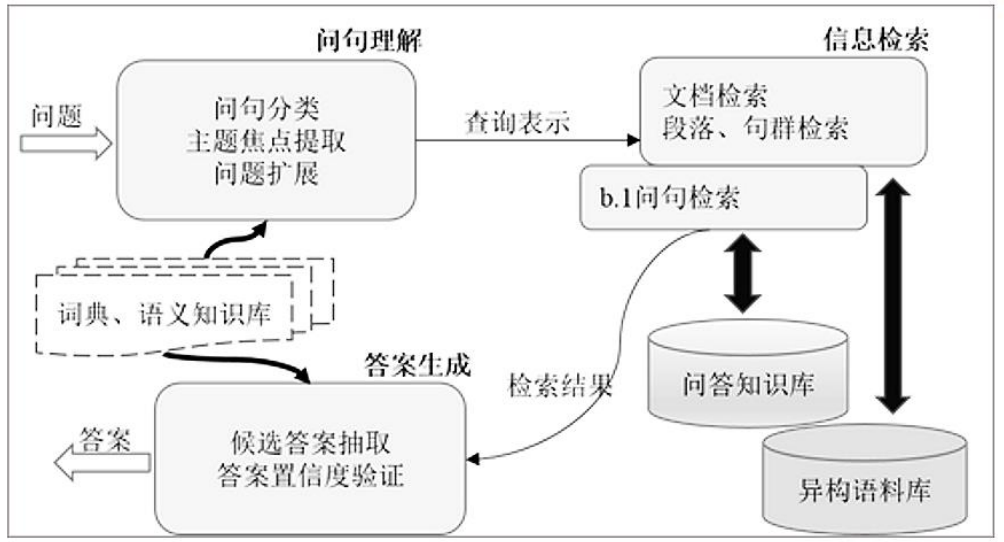
# 自动问答

PaddlePaddle

AI DISCOVERY

**检索式自动问答**一般包括问句理解、信息检索、答案抽取三个功能组成部分。

- **问句理解**
  - **问题分类**：对输入的问题进行处理分类并确定问题类型
  - **主题焦点提取**：实现用户问题的信息需求的精确定位。
  - **问题扩展**：分析问题中的潜在信息，提高答案的召回率




- **信息检索：**从用户的问题中得到的关键词，对于数据库中的文档与关键词的计算匹配程度，从而获取若干个可能包含答案的候选文章，并且根据它们的相似度进行排序。

- **答案抽取：**先从文章中提取出可能包含答案的段落，再对段落进行语义分析，抽取段落中所包含的答案。引入诸如语义词典（WordNet），知识库（Freebase）等外部语义资源。

# 自动问答

AI DISCOVERY

**检索式问答系统**预先定义好知识库，根据输入和上下文语境，从问句中提取关键词，使用启发式算法在知识库中检索并生成答案。

- 检索过程：段落或者句子级排序，利用不同类型关键词的加权组合
- 答案抽取过程：根据问答类型从排序后的段落或句子中抽取答案

代表系统：新加坡国立大学Hui Yang的系统（Yang TREC2002）。

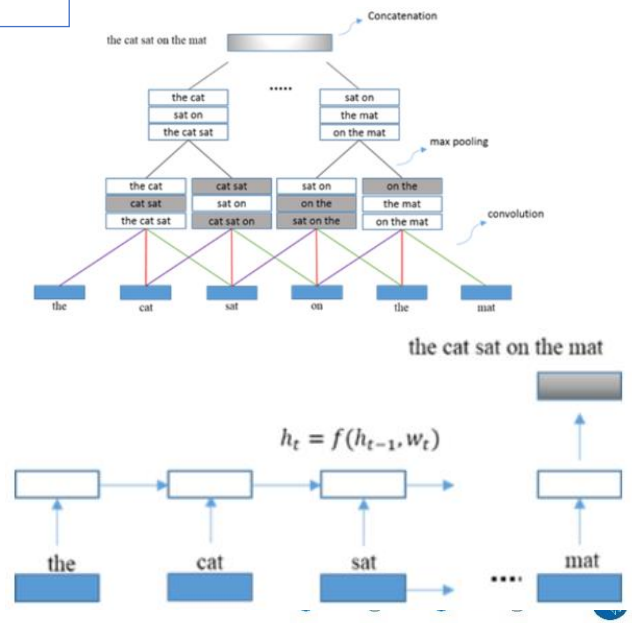

自然语言问句

我国有两个省接壤的省份最多，与它们接壤的省分别有哪些？

代表性系统:Wolfram Alpha

内蒙古：黑龙江、吉林、辽宁、河北、山西、陕西、宁夏、甘肃
陕西：内蒙古、宁夏、甘肃、四川、重庆、湖北、河南、山西

基于知识库的问答系统




基于深度学习的问答技术

# 自动问答

AI DISCOVERY

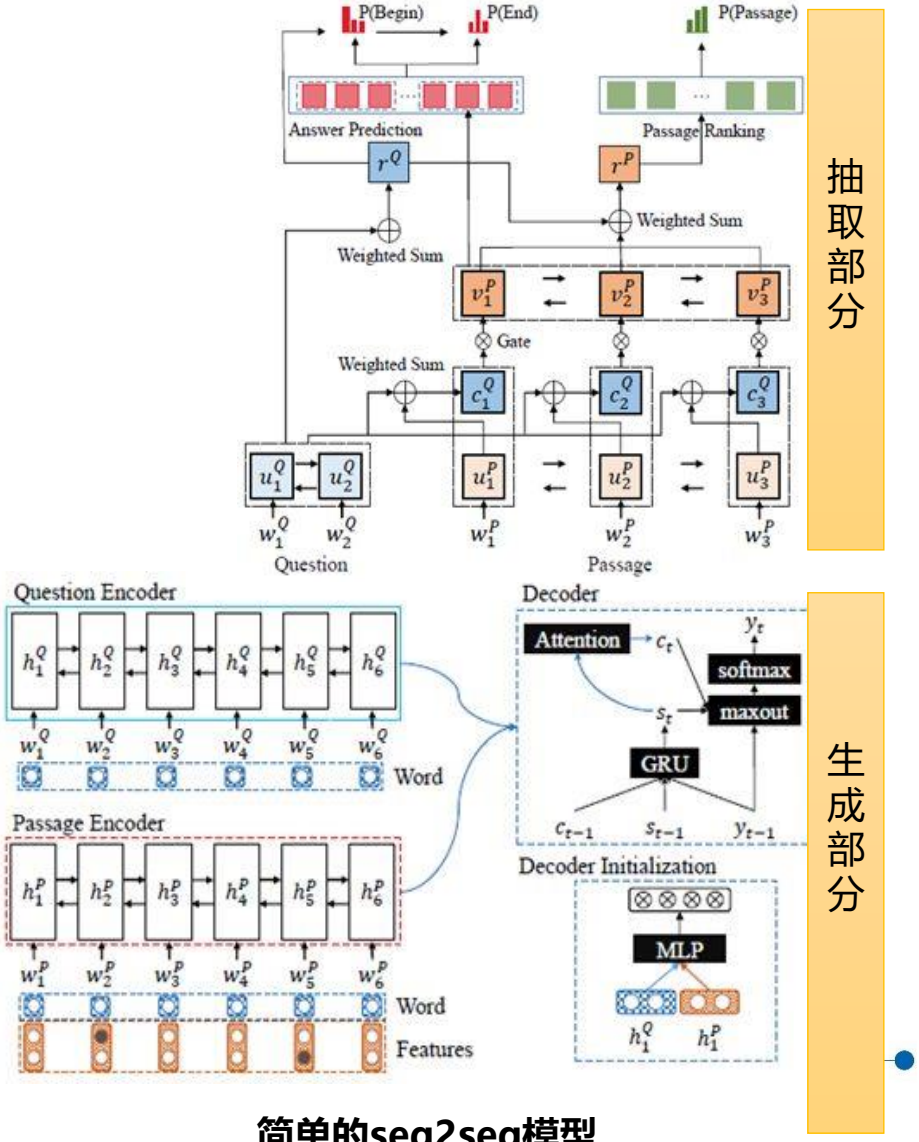
**生成式问答系统**的答案生成模式如下：

- ✓ 答案完全在某篇原文
- ✓ 答案分别出现在多篇文章中
- ✓ 答案一部分出现在原文，一部分出现在问题中
- ✓ 答案的一部分出现在原文，另一部分是生成的新词
- ✓ 答案完全不在原文出现（Yes / No 类型）

**根据训练数据，**比如MS MARCO数据集，运用**seq2seq模型**先抽取出答案，然后利用这个特征再对文章生成答案。

**（1）抽取部分的模型**同时做了两个任务，预测答案，并对文档进行排序；

**（2）生成部分的模型**将抽取部分标注的答案作为一个特征信息叠加到文章向量中，对这段文章和问题，重新通过Encoder建模，得到一个综合的语义向量，再输入Decoder中生成答案。




简单的seq2seq模型

# 机器翻译



**机器翻译 (machine translation, MT)** 是用计算机把一种语言(源语言，source language)翻译成另一种语言(目标语言，target language)的一门学科和技术。


源语言句子： $S = s_1^m = s_1 s_2 \cdots s_m$

目标语言句子： $T = t_1^l = t_1 t_2 \cdots t_l$

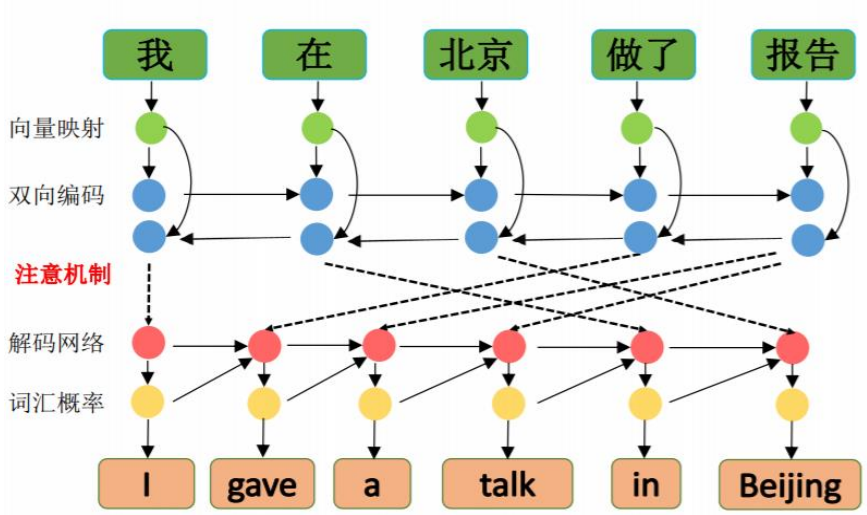
贝叶斯公式： $P(T|S) = \dfrac{P(T) \times P(S|T)}{P(S)}$

$$T' = \underset{T}{\operatorname{argmax}}\, P(T) \times P(S|T)$$

**语言模型**
Language model, LM

**翻译模型**
Translation model, TM

**基于统计模型的方法**




**基于神经网络模型的方法**


# 自然语言太难了（一）

AI DISCOVERY

## 中文分词（segmentation）困难

- ✓ 一行行行行行，一行不行行行不行
- ✓ 来到杨过曾经生活过的地方，小龙女说：“我也想过过过儿过过的生活”
- ✓ 另一个宿舍的人说你们宿舍的地得扫了
- ✓ 校长说衣服上除了校徽别别别的
- ✓ 来到儿子等校车的地方，邓超对孙俪说：“我也想等等等等等过的那辆车”
- ✓ 一位友好/的哥/谭市民

AI DISCOVERY


# 自然语言太难了（二）

AI DISCOVERY

## 歧义(ambiguity)现象

- **结构歧义**
  - 今天中午**吃馒头**　　今天中午**吃食堂**
- **指代歧义**
  - **他快抱不起儿子了，因为他太胖了**
- **语义歧义**

他说:“她这个人真有意思(funny)”。她说:“他这个人怪有意思的(funny)”。于是人们以为他们有了意思(wish)，并让他向她意思意思(express)。他火了:“我根本没有那个意思(thought)”！她也生气 了:“你们这么说是什么意思(intention)”？事后有人说:“真有意思(funny)”。也有人说:“真没意思 (nonsense)”。

AI DISCOVERY


# 自然语言太难了（三）

AI DISCOVERY

## 未知语言现象

- 新词
  - 不明觉厉　　累觉不爱　　十动然拒
- 旧词新义
  - 母鸡　　白骨精　　潜水
- 新用法新结构
  - 在口语中或部分网络语言中，不断出现一 些“非规范的”新的语句结构

    c位出道　　ORZ　　热skr人了

AI DISCOVERY


# 自然语言太难了（四）

AI DISCOVERY

## 不同语系的差异

- 屈折语
  - 用词的形态变化表示语法关系，如英语、法语等
- 黏着语
  - 词内有专门表示语法意义的附加成分，词根或词干与附加成分的结合不紧密，如日语、韩语、土耳其语等
- 孤立语
  - 形态变化少，语法关系靠词序和虚词表示，如汉语

AI DISCOVERY


# 目录

AI DISCOVERY



AI DISCOVERY


# 自然语言处理

AI DISCOVERY

词向量概念

词向量学习模型

词向量学习模型的优化

句子向量

AI DISCOVERY

# 词向量的概念

AI DISCOVERY

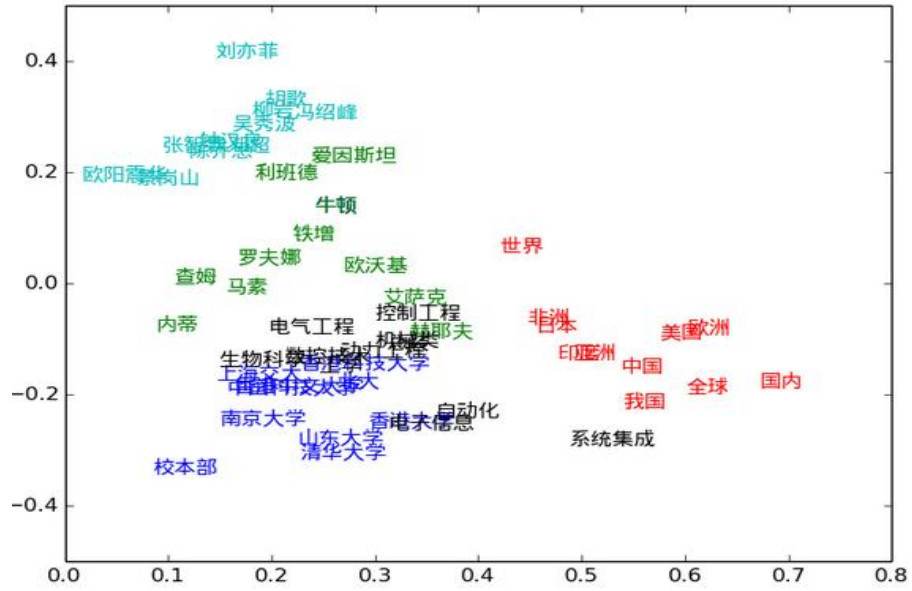
将自然语言转化成机器理解符号

深度学习应用于自然语言处理之前，传统的词表达通常采用one-hot方式表达

杭州 [0,0,0,0,0,0,0,1,0,……，0,0,0,0,0,0,0]
上海 [0,0,0,0,1,0,0,0,0,……，0,0,0,0,0,0,0]
宁波 [0,0,0,1,0,0,0,0,0,……，0,0,0,0,0,0,0]
北京 [0,0,0,0,0,0,0,0,0,……，1,0,0,0,0,0,0]

1.向量维度取决于语料库中词数，导致维数灾难
2.向量之间相互独立，看不出关联关系

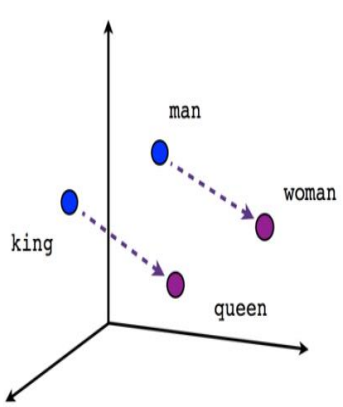
词向量可以将one-hot编码转化为低维度的连续值，也就是稠密向量



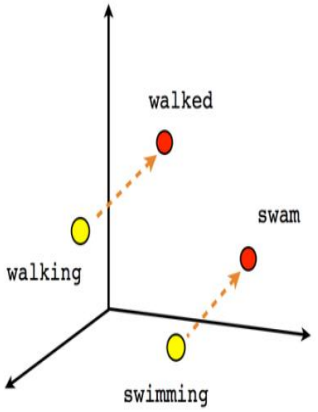

1.词向量中**语义上**相近的词距离接近
2. 词向量中**语法上**相近的词距离接近

AI DISCOVERY

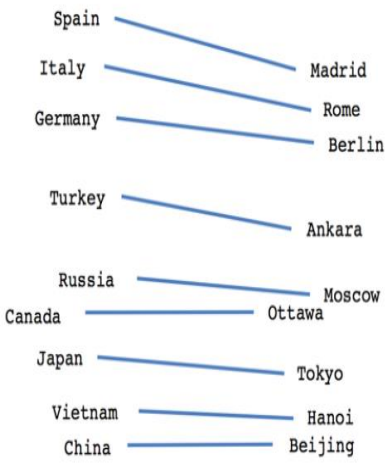
# 词向量的应用



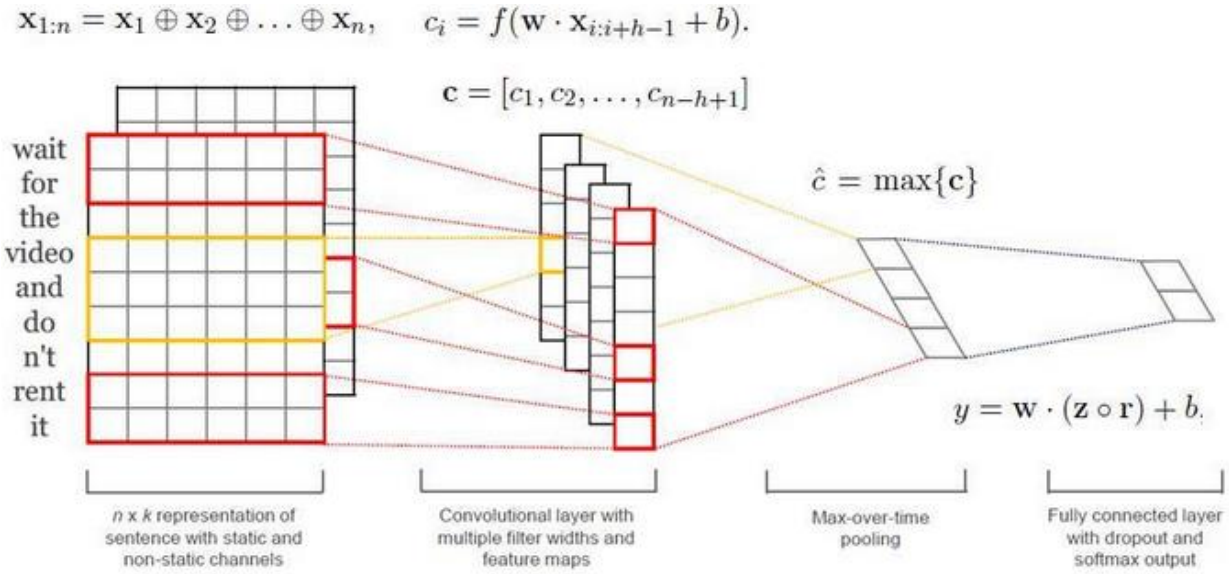

Male-Female




Verb tense




Country-Capital




## 计算相似度

- 寻找相似词
- 信息检索查询扩展
- 知识推演

## 作为神经网络的输入

- 文本分类
- 情感分析
- 文档主题判别

## 句子/文档表示

- 无监督句子/文档表示
- 有监督句子/文档表示


# 自然语言处理

AI DISCOVERY

词向量概念

词向量学习模型

词向量学习模型的优化

句子向量

AI DISCOVERY

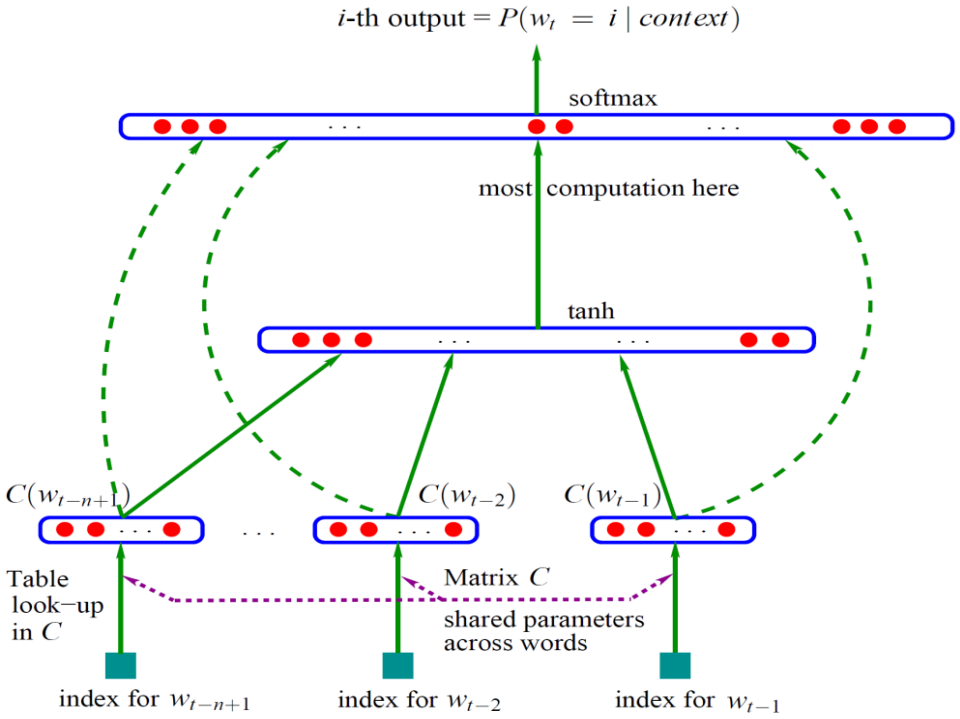
# 词向量学习模型——神经网络语言模型

AI DISCOVERY

词向量是在训练语言模型的同时获得的，而语言模型就是判断给定字符串为自然语言的概率$P(w_1, w_2, \ldots, w_n)$，其中$w_1, w_2, \ldots, w_n$依次表示字符串中的各个词。如果$P$大于某个阈值，就认为该字符串为自然语言。

**n-gram 语言模型**

$$P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)\ldots P(w_n \mid w_1, w_2, \ldots, w_{n-1})$$




从下往上依次为该模型的输入层、隐藏层和输出层。

（1）首先根据训练集生成词典$D$;

（2）对于语料中的任意词$w_t$，获取其前面$n-1$个词，输入层将这前$n-1$个词的词向量$C(w_{t-n+1})$，$C(w_{t-n+2})$，……，$C(w_{t-1})$拼接起来;

（3）隐藏层通过激励函数将输入信息进行转换;

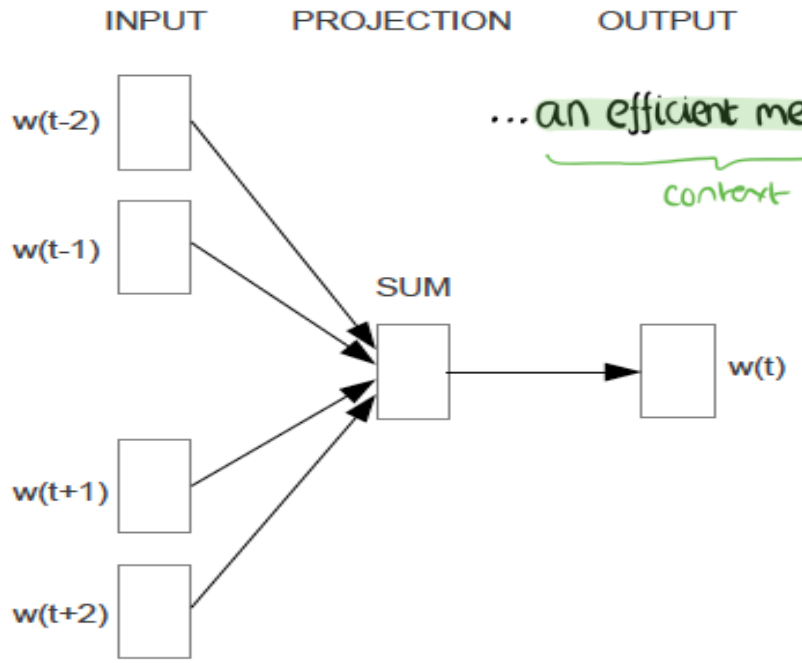
（4）输出层计算词$w_t$的概率。

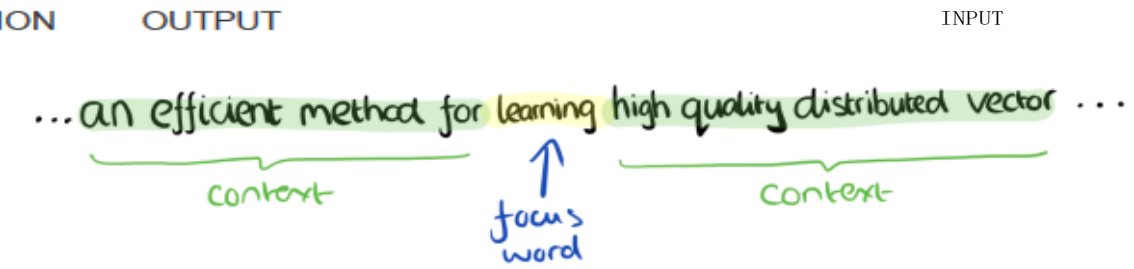
# 词向量学习模型——CBOW和skip-gram

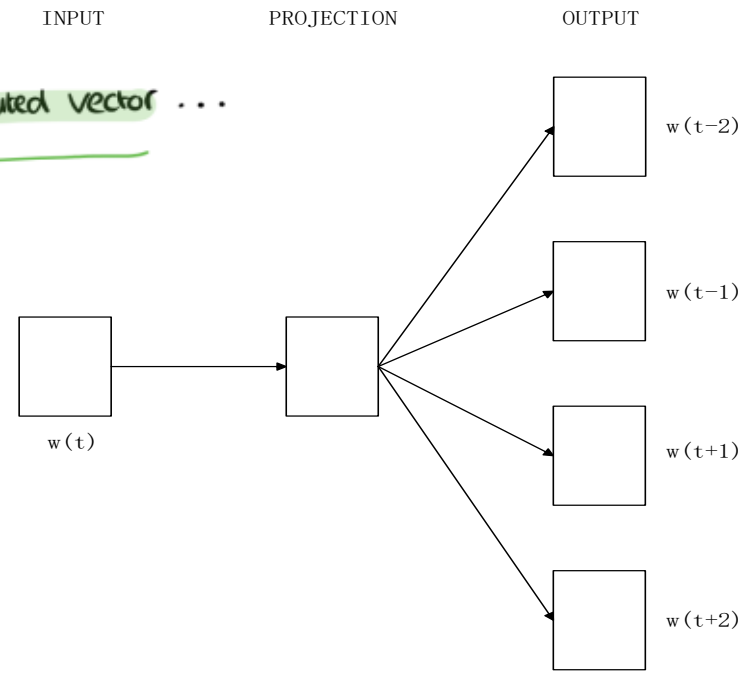
AI DISCOVERY

如果是用一个词的上下文作为输入，来预测这个词本身，则是 『CBOW 模型』

如果是用一个词语作为输入，来预测它的上下文词，则这个模型叫做『Skip-gram 模型』




CBOW与神经网络语言模型类似，不同之处在于隐藏层的计算方法不同，一个是向量串联，一个是按位累加。

Skip-gram中的隐藏层是输入的直接投影，共享给输出层的每一个神经元。


# 自然语言处理

AI DISCOVERY

词向量概念

词向量学习模型

词向量学习模型的优化

句子向量

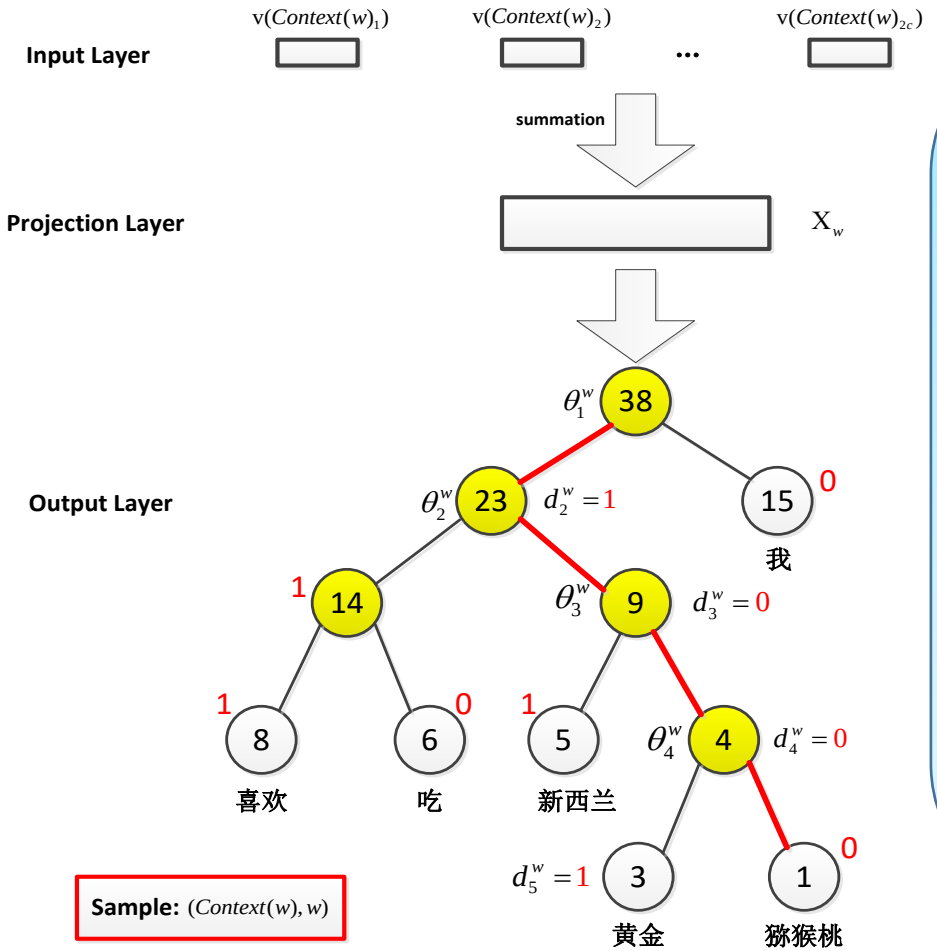
AI DISCOVERY

# 词向量学习模型——层次化softmax方法

AI DISCOVERY

CBOW/skip-gram模型的输出层使用的softmax函数，时间复杂度为O(|*D*|)，因此计算代价很大，对大规模的训练语料来说，训练非常耗时。

- 层次化softmax是一种对输出层进行优化的策略。
- 输出层从原始模型单层计算概率值改为利用Huffman树计算概率值。




以CBOW模型为例，输出层对应一颗Huffman树，它以语料中出现过的词作为叶子节点，以各词在语料中出现的次数当作权值。

对于词典*D*中的任意词，Huffman树中必存在一条从根节点到该词的唯一路径，路径上的每个分支都可以看成一个二分类问题，每一次分类都产生一个概率，将这些概率乘起来就能得到目标概率。

# 词向量学习模型——负采样方法

AI DISCOVERY

与层次化softmax方法相比，负采样不再使用复杂的Huffman树，而是采用随机负采样策略，优化目标改为：最大化正样本的概率，同时最小化负样本的概率。

在CBOW模型中，已知词$w$的上下文$Context(w)$，需要预测$w$。因此对于给定上下文，词$w$就是一个正样本，其他词就是负样本。但是负例样本太多了，我们怎么去选取呢？

在语料库中，各个词出现的频率是不一样的，我们采样的时候要求高频词选中的概率较大，而低频词选中的概率较小。这就是一个带权采样的问题。

$$g(w)=\sigma\left(\mathrm{x}_w^{\mathrm{T}}\theta^w\right)\prod_{u\in \mathrm{NEG}(w)}\left[\left(1-\sigma\left(\mathrm{x}_w^{\mathrm{T}}\theta^u\right)\right)\right]$$

$\sigma\left(\mathrm{x}_w^{\mathrm{T}}\theta^w\right)$

表示当上下文为$Context(w)$时，预测词为$w$（正样本）的概率

$\sigma\left(\mathrm{x}_w^{\mathrm{T}}\theta^u\right)$

表示当上下文为$Context(w)$时，预测词为$u$（负样本）的概率

AI DISCOVERY


# 自然语言处理

AI DISCOVERY

词向量概念

词向量学习模型

词向量学习模型的优化

句子向量

AI DISCOVERY


# 思考这样一个问题



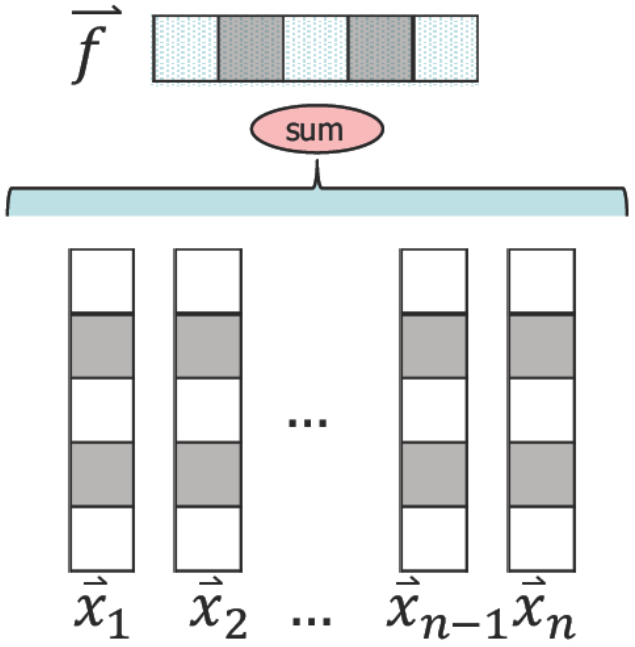
已知词向量，如何获得一个句子的向量？

一个好的句子向量要有哪些特性？

# 句子嵌入表示模型（1）

AI DISCOVERY

- Bag-of-words (Kalchbrenner et al., 2014)
  - Simply element-wise summing embedding
  - Learning embeddings by back-propagation




◆ 优点：操作简单；不需要额外的模型训练，得到的句子向量是定长的，句子向量的长度等于词向量的长度

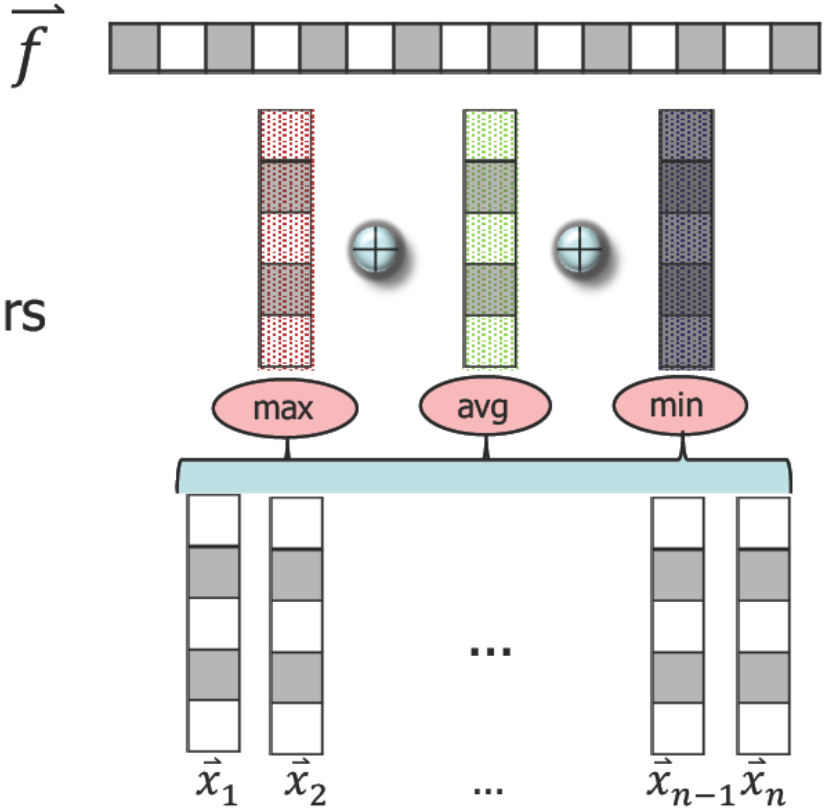
◆ 缺点：丢失了词汇之间的顺序信息，在自然语言里，词汇之间的顺序信息对整体语义的理解非常重要，有时候语序错了，句子的意思截然相反

AI DISCOVERY

# 句子嵌入表示模型（2）

AI DISCOVERY

- Pooling (Tang et. al.,2014; Vo and Zhang, 2015)
  - Make use of Pre-trained word embeddings
  - Extract salient features for traditional classifiers




句子向量长度是词向量的3倍，除了最大、最小和平均，当然也可以追加其他操作继续进行扩展

AI DISCOVERY

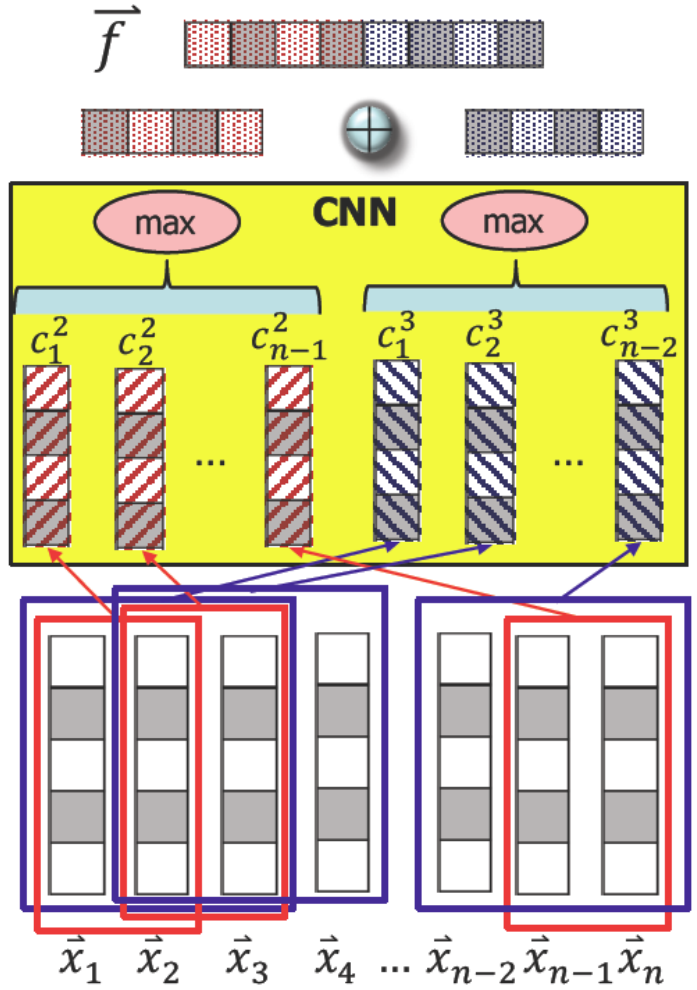
# 句子嵌入表示模型（3）

AI DISCOVERY

- CNN (Kim, 2014)
  - Feature combinations
  - Single CNN layer
  - Varied-window-size convolutional filters
  - Multichannel (1 static+ 1 nonstatic)

✓ 是一个有监督学习模型

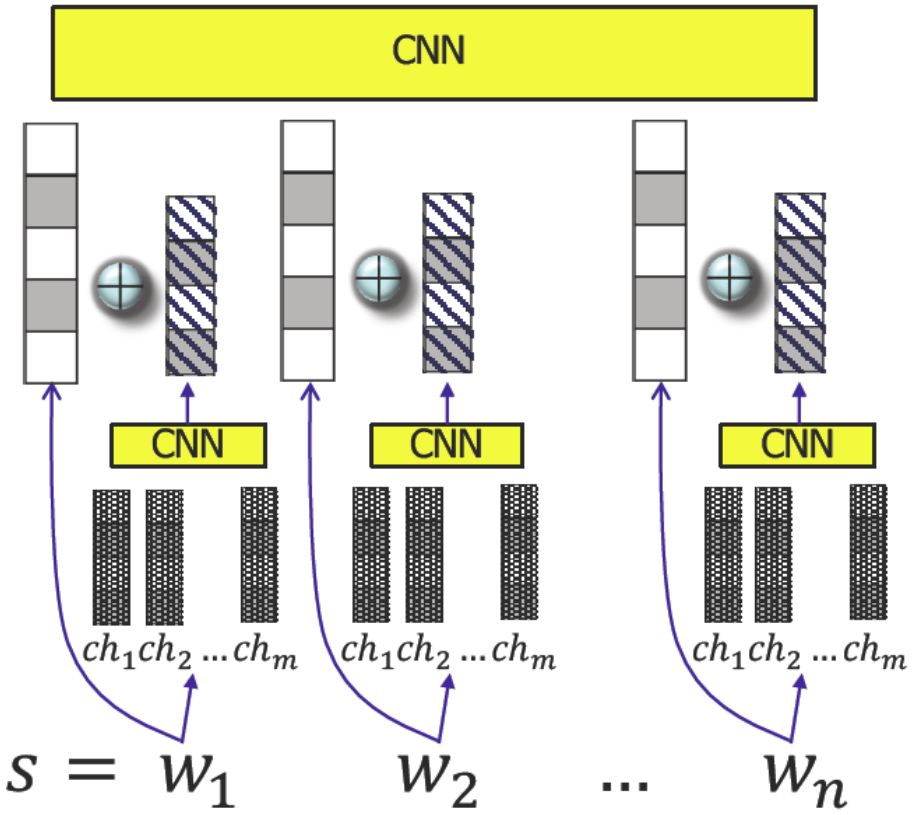
✓ 可以学习到词序信息



# 句子嵌入表示模型（4）

AI DISCOVERY

- Variations
  - dos Santos et al. (2014)
    - Add character information



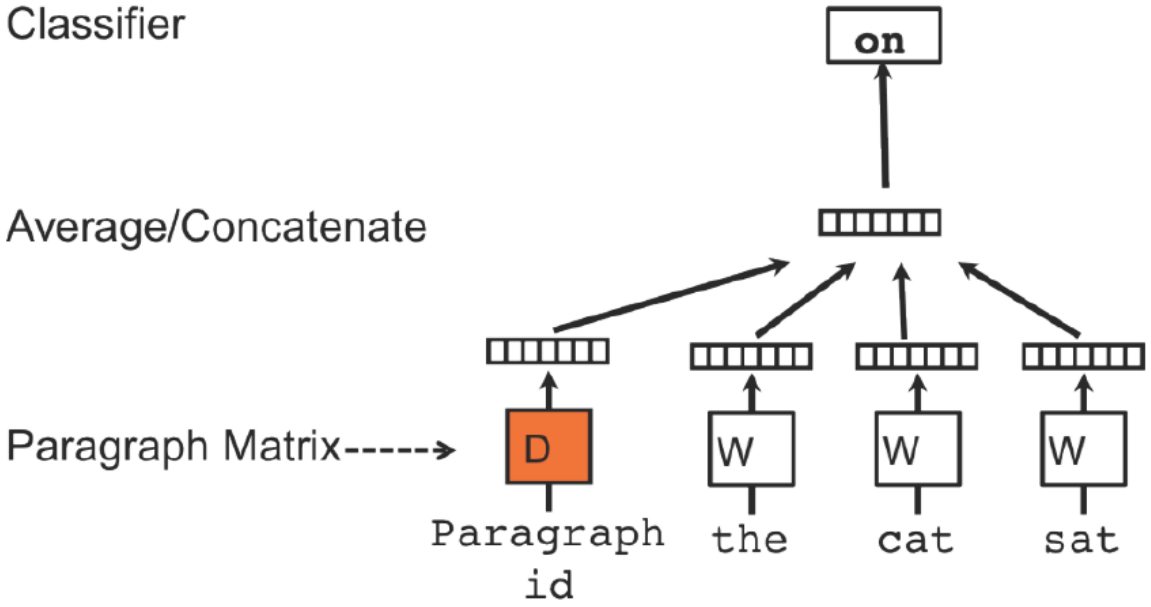

- ✓ 层级CNN，在输入端引入了字符向量
- ✓ 在文本分析的很多任务中，都证明引入字符向量可以提高分析性能

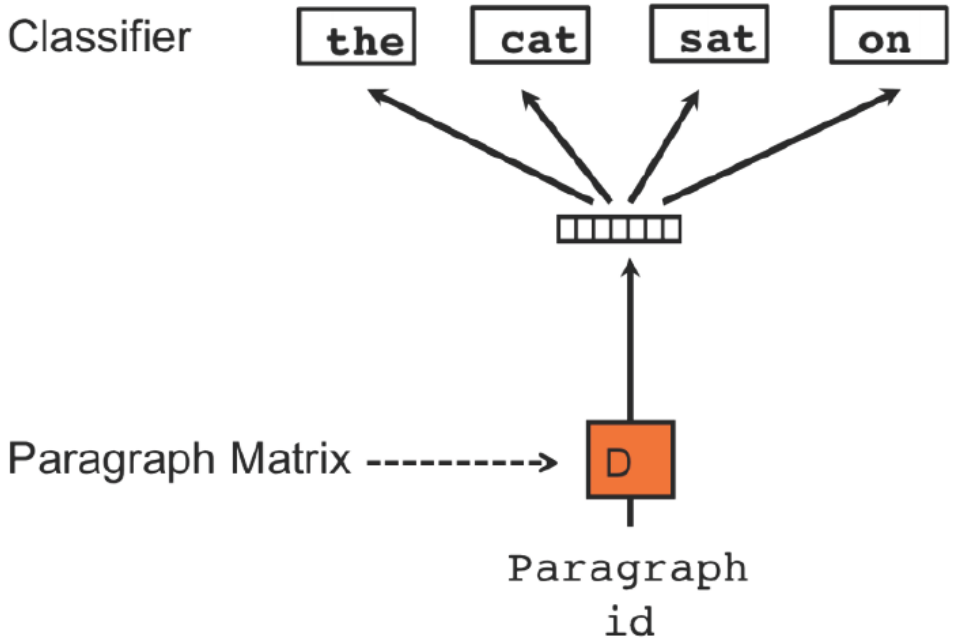
AI DISCOVERY

# 文档嵌入表示模型（1）

AI DISCOVERY

## Document Embedding



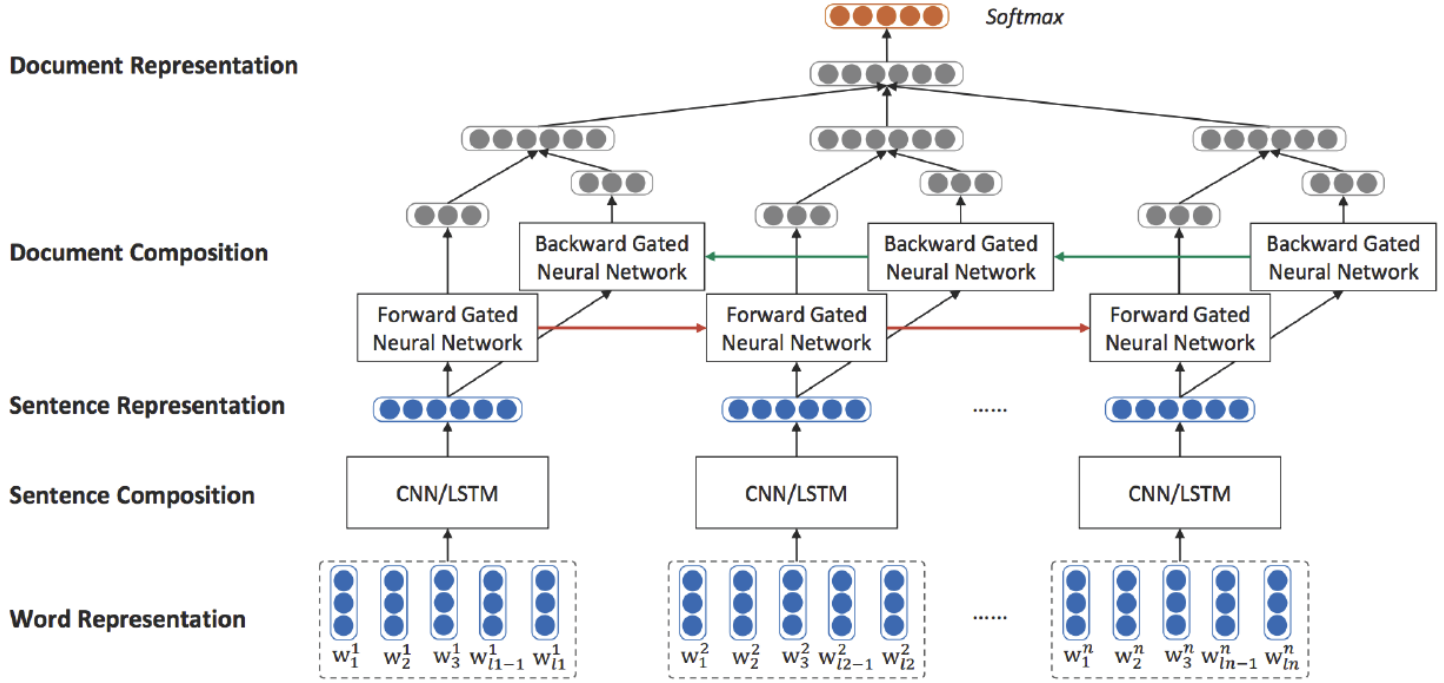




AI DISCOVERY

# 文档嵌入表示模型（2）

AI DISCOVERY

- Pooling (Tang et al., 2015a)
  - Average pooling sentence representations as document representation
- LSTM/CNN-GRU (Tang et al., 2015b)




# 目录

AI DISCOVERY



AI DISCOVERY


# 自然语言处理

AI DISCOVERY

RNN

LSTM、GRU

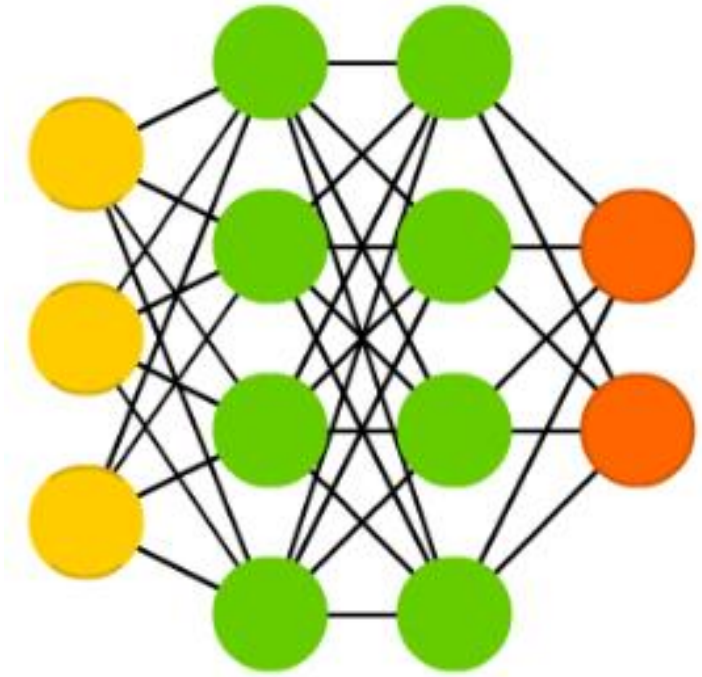
Attention

# 前馈神经网络与NLP



隐层表示/抽象

输入句子



输出目标

- ✓ 层与层之间有连接，每层的节点之间是无连接的。
- ✓ 输入和输出的维数都是固定的，不能任意改变。无法处理变长的序列数据。
- ✓ 词汇节点位置无关，无法对语言中的词序进行建模。

举个例子：
特朗普的儿子是谁？
谁的儿子是特朗普？


# 各种处理任务

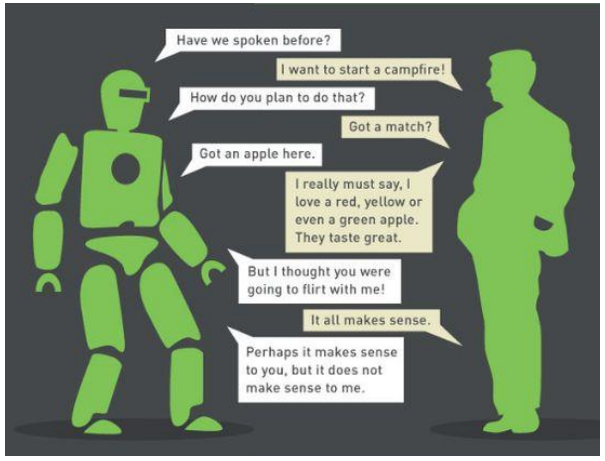
AI DISCOVERY

## ◆变长输入

- ✓不同大小的图片
- ✓时长不一的视频
- ✓长短不同的句子
- ✓序列长度不同的对话

## ◆相互依赖

- ✓视频由连续的图片组成
- ✓词义取决于上下文
- ✓情感取决于上下文

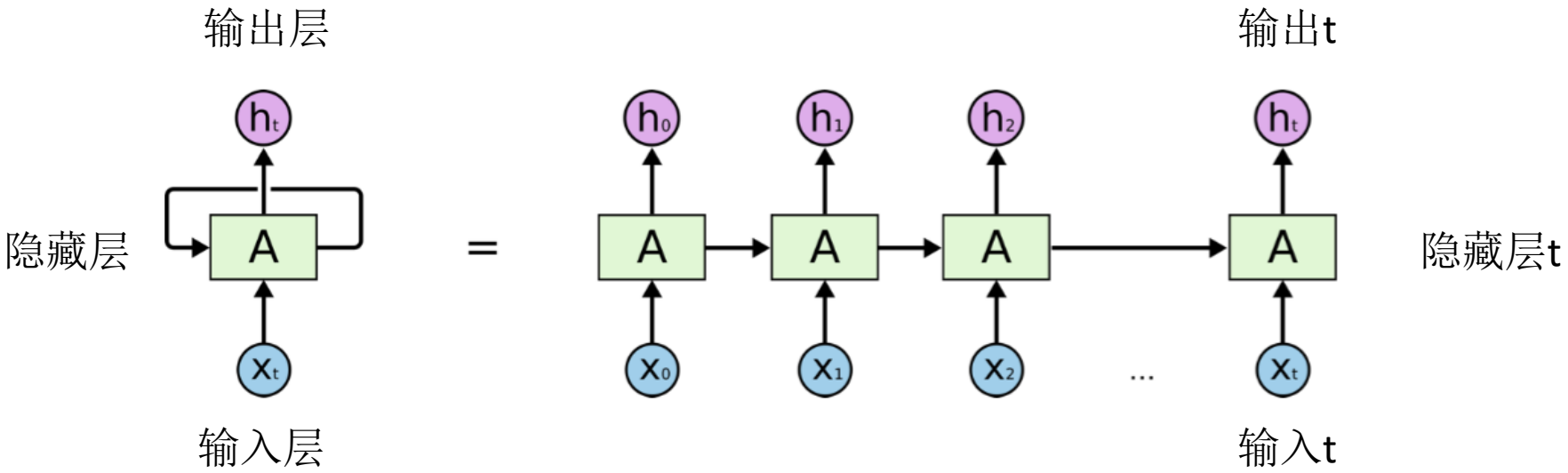



OR


AI DISCOVERY

# 循环神经网络

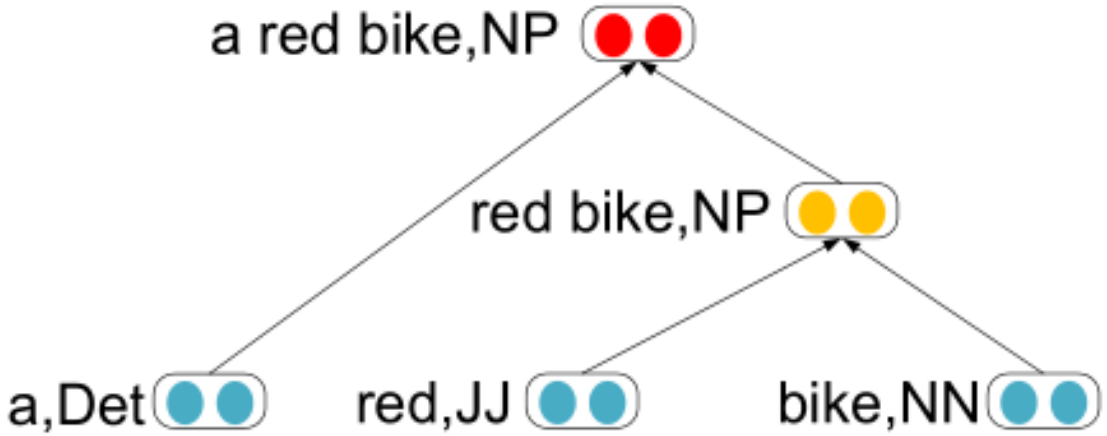
◆**循环神经网络（Recurrent Neural Network，RNN），**也叫递归神经网络。这里为了区别与另外一种**递归神经网络（Recursive Neural Network），**我们称为**循环神经网络。**




✓循环神经网络通过使用带自反馈（隐藏层）的神经元，能够处理任意长度的序列。循环神经网络比前馈神经网络更加符合生物神经网络的结构。已经被广泛应用在语音识别、图像处理、语言模型以及自然语言生成等任务上。

# 递归神经网络






根据一个给定的拓扑结构，
进行词的语义组合。

给定一个二叉句法树，
$((p_2 \rightarrow a\ p_1),(p_1 \rightarrow bc))$，
父节点由子节点计算得到。

$$\mathbf{p}_1 = f(\mathbf{W}\begin{bmatrix}\mathbf{b}\\ \mathbf{c}\end{bmatrix}),$$

$$\mathbf{p}_2 = f(\mathbf{W}\begin{bmatrix}\mathbf{a}\\ \mathbf{p}_1\end{bmatrix}).$$

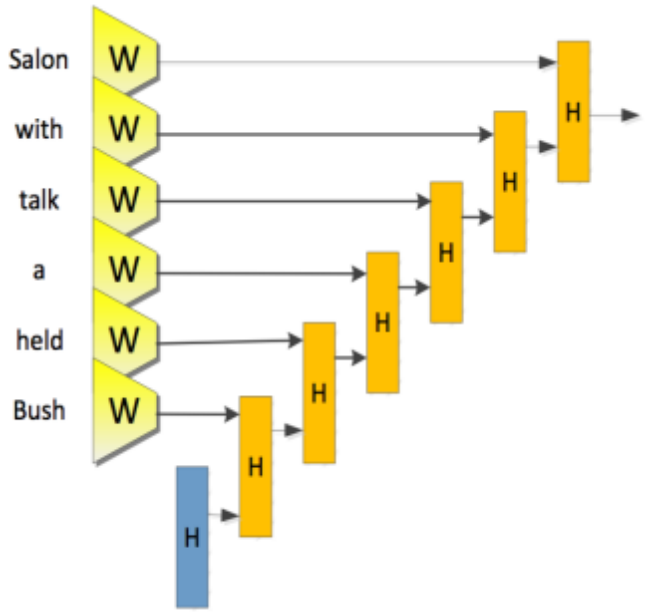
R. Socher et al. “Parsing with compositional vector grammars”. In: Proceedings of ACL. 2013.

# 循环神经网络

AI DISCOVERY

给定一个输入序列$\boldsymbol{x}^{(1:n)} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(t)}, \ldots, \boldsymbol{x}^{(n)})$，循环神经网络通过下面公式更新带反馈边的隐藏层的活性值 $\mathbf{h}(t)$，即抽象表示：

$$\mathbf{h}_t = \begin{cases} 0 & t = 0 \\ f(\mathbf{h}_{t-1}, \mathbf{x}_t) & \text{otherwise} \end{cases}$$




给定一个输入序列$\boldsymbol{x}^{(1:n)} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(t)}, \ldots, \boldsymbol{x}^{(n)})$，循环神经网络通过下面公式更新带反馈边的隐藏层的活性值 $\mathbf{h}(t)$，即抽象表示：

AI DISCOVERY

# 简单循环神经网络

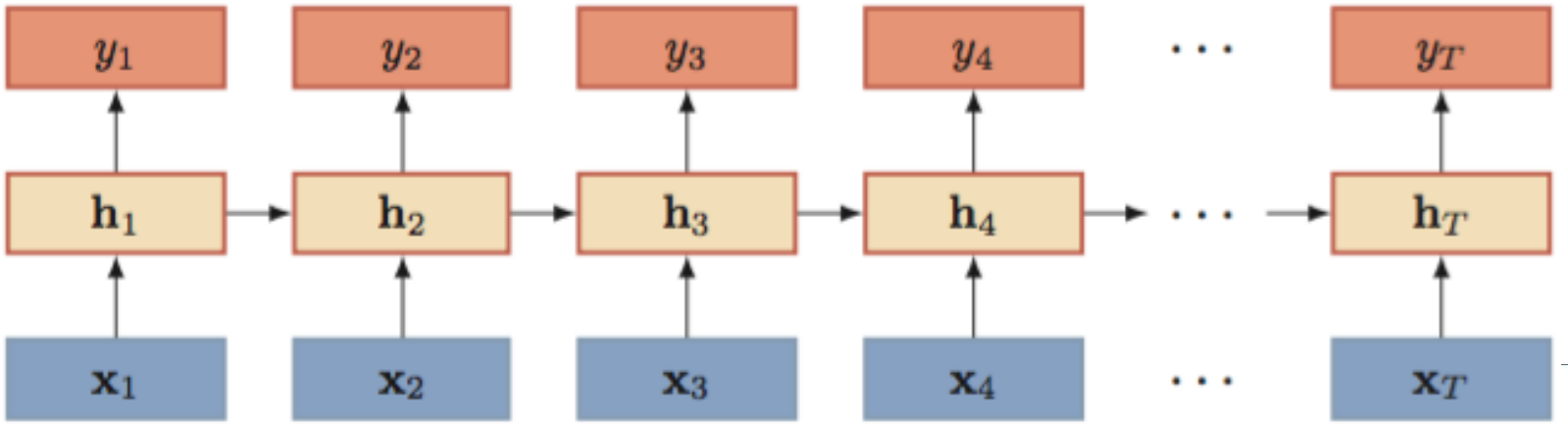
AI DISCOVERY

✓假设时刻 t 时，输入为 $\mathbf{x_t}$，隐层状态（隐层神经元活性）为 $\mathbf{h_t}$。 $\mathbf{h_t}$ 不仅和当前时刻的输入相关，也和上一个时刻的隐层状态相关。

✓一般我们使用如下函数：

$$\boldsymbol{h}_t = f(\boldsymbol{U}\boldsymbol{h}_{t-1} + \boldsymbol{W}\boldsymbol{x}_t + b) \qquad \boldsymbol{y}_t = softmax(\boldsymbol{W}^{(S)}\boldsymbol{h}_t)$$

✓ 这里，$f$ 是非线性函数，通常为 *sigmod*函数或 *tanh* 函数。




AI DISCOVERY


# 自然语言处理

AI DISCOVERY

RNN

LSTM、GRU

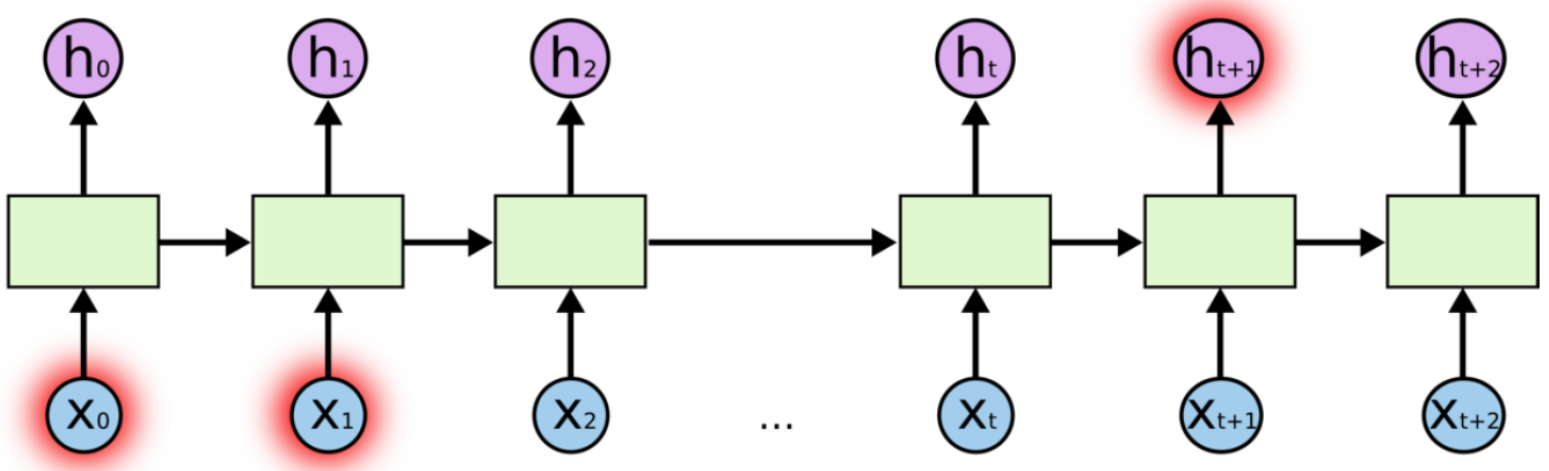
Attention

# 长期依赖的问题

AI DISCOVERY

- 很久以前的输入，对当前时刻的网络影响较小；反向传播的梯度，也很难影响很久以前的输入

- 例如：
  - The cat, which already ate a bunch of food, (was) full.
  - The cats, which already ate a bunch of food, (were) full.



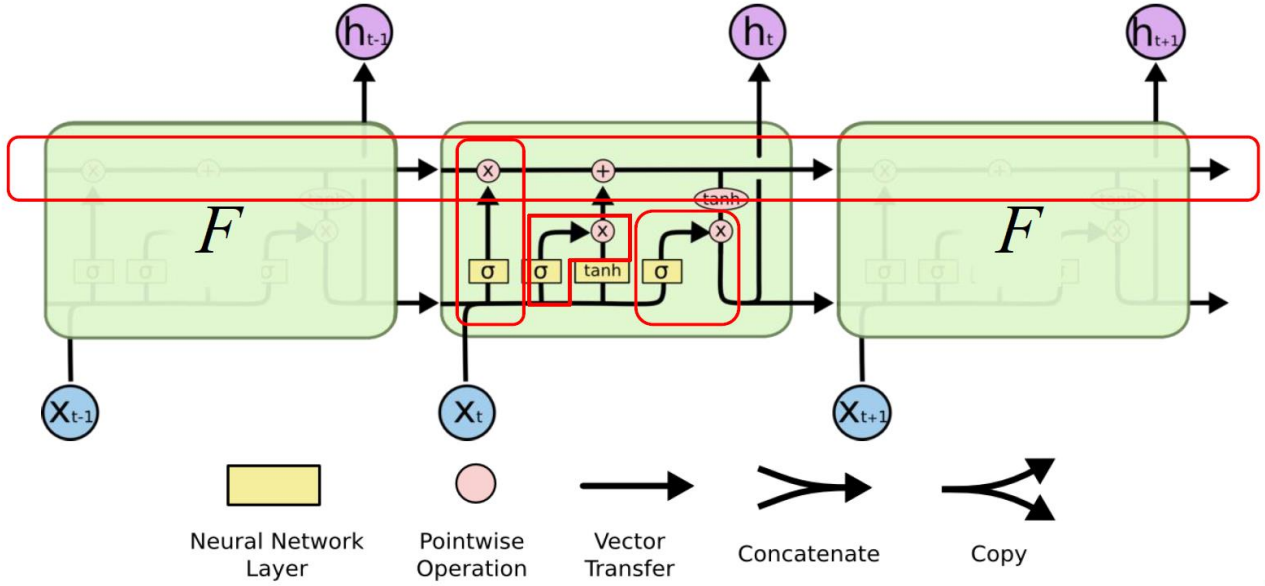

AI DISCOVERY

- 解决思路：采用ReLu函数，或采用其他模型来代替非线性激活函数

# 改进方案：长短时记忆神经网络LSTM

AI DISCOVERY

长短时记忆神经网络（Long Short-Term Memory Neural Network，LSTM）是循环神经网络的一个变体，可以有效地解决**长期依赖问题/梯度消失**问题。



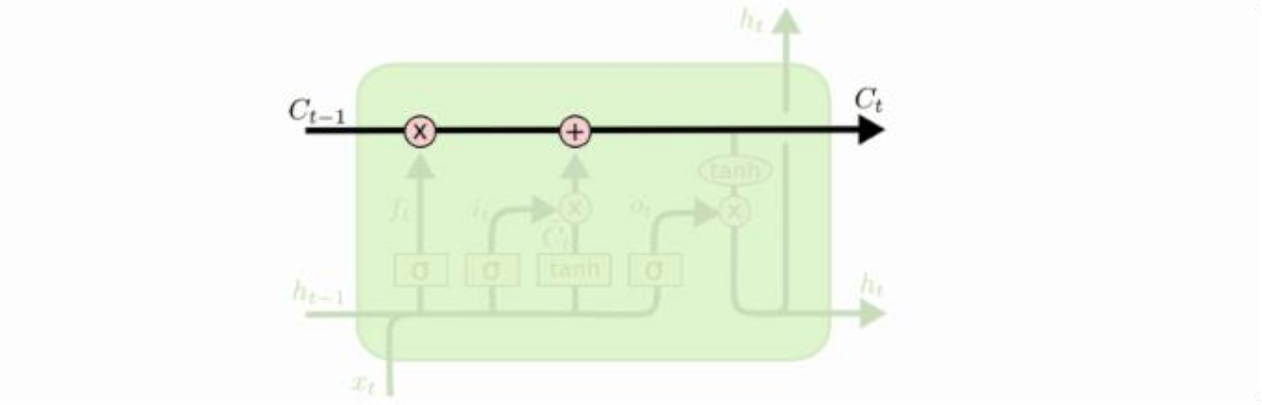

LSTM 模型的关键是引入了一组**记忆单元**（Memory Units），允许网络可以学习何时遗忘历史信息，何时用新信息更新记忆单元。在时刻 t 时，记忆单元 $\mathbf{c}_t$ 记录了到当前时刻为止的所有历史信息，并受**三个“门”控制**：输入门 $\mathbf{i}_t$，遗忘门$\mathbf{f}_t$和输出门$\mathbf{o}_t$。三个门的元素的值在 [0, 1] 之间。

AI DISCOVERY

# LSTM

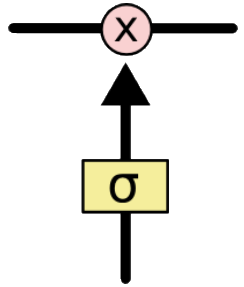
AI DISCOVERY

## ◆核心：记忆（细胞状态）和门机制



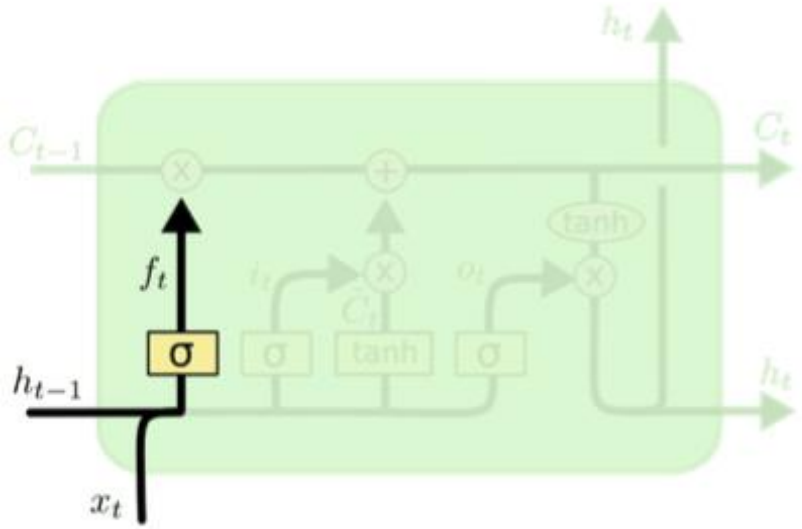

细胞的状态在整条链上运行，只有一些小的线性操作作用其上，信息很容易保持不变的流过整条链。

门(Gate)是一种可选地让信息通过的方式。 它由一个Sigmoid神经网络层和一个点乘法运算组成。




Sigmoid神经网络层输出0和1之间的数字，这个数字描述每个组件有多少信息可以通过， 0表示不通过任何信息，1表示全部通过

# Forget Gate




以语言模型为例，细胞状态可能包括当前主语的性别，从而决定使用正确的代词（它/他/她），当看到一个新主语时，需要忘记旧主语的性别。

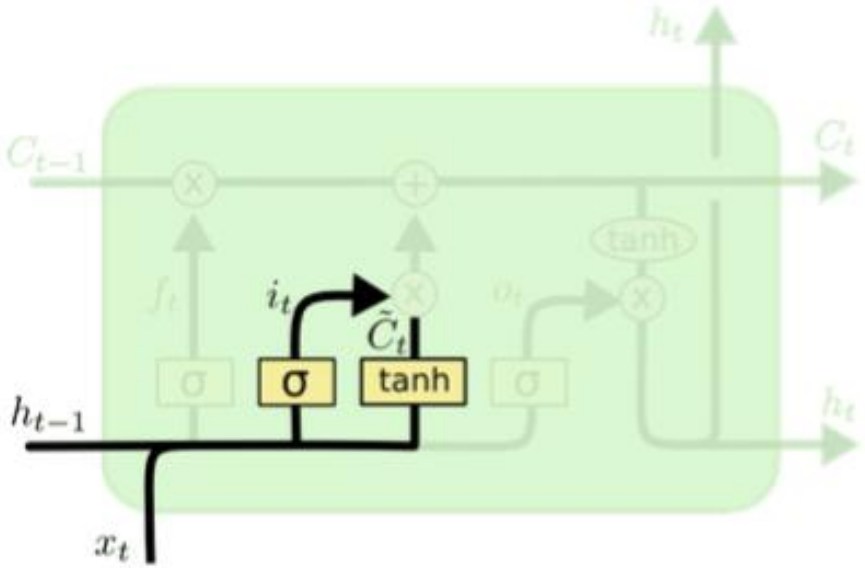
$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

- ✓ 遗忘门决定我们要从细胞状态中丢弃什么信息
- ✓ 它查看$h_{t-1}$(前一个隐藏状态)和$x_t$(当前输入)，并为状态$C_{t-1}$(上一个状态)中的每个数字输出0和1之间的数字
- ✓ 1代表完全保留，而0代表彻底删除

# Input Gate

AI DISCOVERY

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



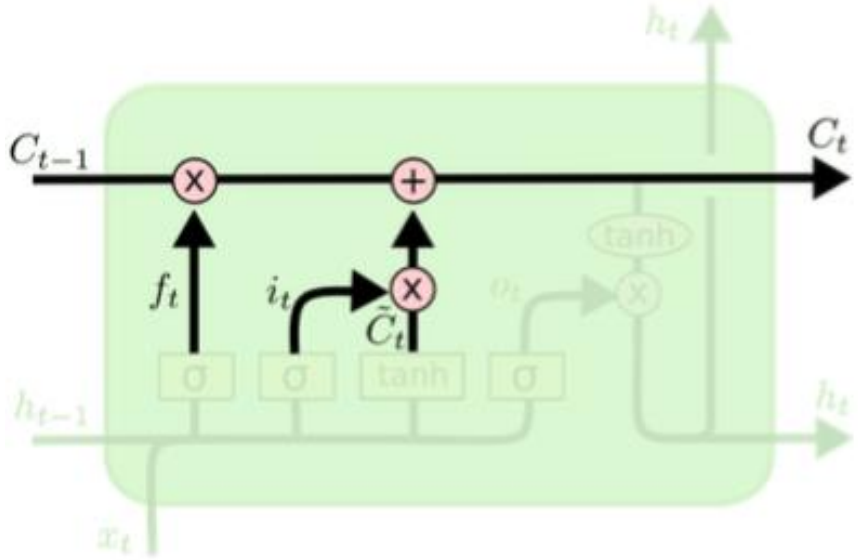

输入门决定我们要在细胞状态中存储什么信息

- ✓ 首先，输入门的Sigmoid层决定了我们将更新哪些值
- ✓ 然后，一个tanh层创建候选向量$\tilde{C}_t$,该向量将会被加到细胞的状态中
- ✓ 最后，结合这两个向量来创建更新值

AI DISCOVERY


# Update Memory





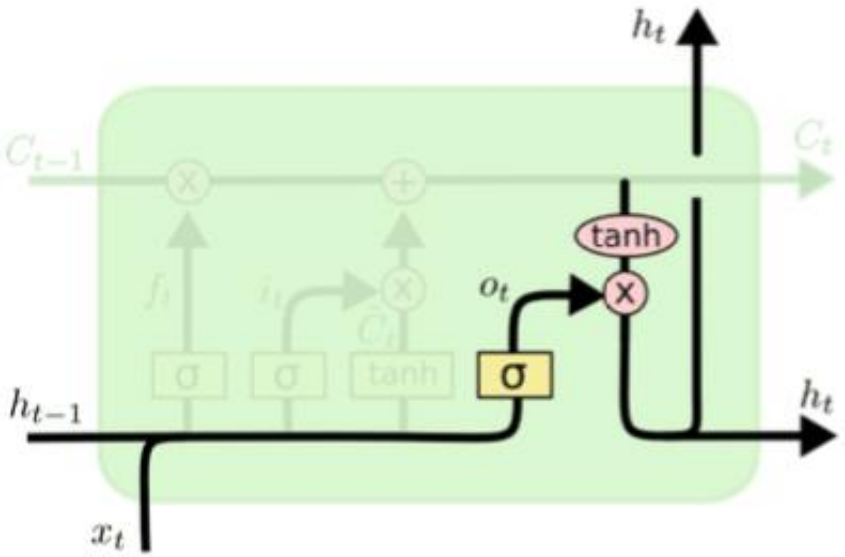

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

- 现在是时候去更新上一个状态值$C_{t-1}$了，将其更新为$C_t$
- 将上一个状态值$C_{t-1}$乘以$f_t$，以此表达期待忘记的部分。之后将得到的值加上 $i_t * \tilde{C}_t$。这个得到的是新的状态值

# Output Gate

AI DISCOVERY

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




- ✓ 输出门决定我们要输出什么， 此输出将基于当前的细胞状态
- ✓ 首先，通过一个sigmoid层，决定了我们要输出细胞状态的哪些部分。
- ✓ 然后，将细胞状态通过tanh（将值规范化到-1和1之间），并将其乘以Sigmoid门的输出，至此完成了输出门决定的那些部分信息的输出。

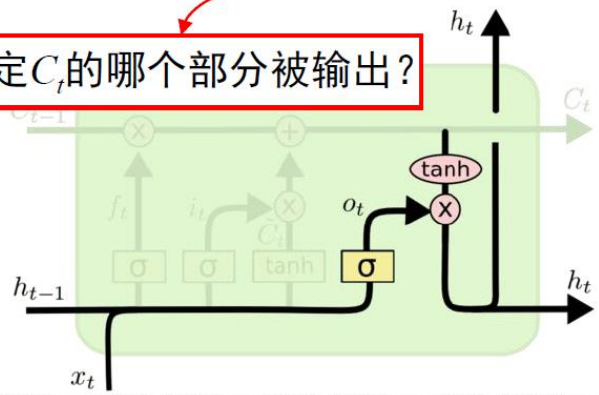
举个语言模型的例子，当看到一个主题词，考虑到后面可能出现的词，可能需要输出与动词相关的信息，比如单数还是复数，需要根据主题信息来决定具体输出什么。

AI DISCOVERY

# Long Short Term Memory (LSTM)

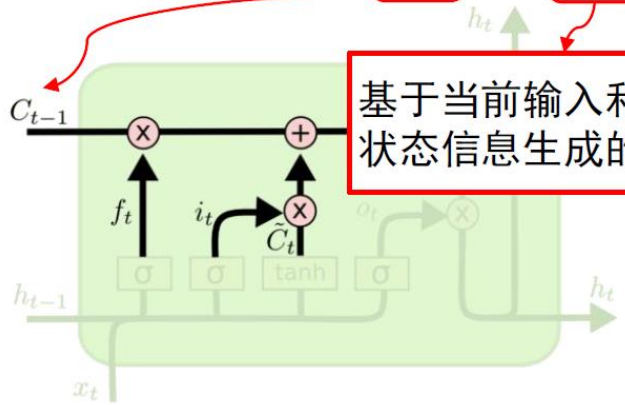
- 输出： $h_t = o_t * \tanh(C_t)$
- 输出门： $o_t = \sigma(W_o\,[h_{t-1}, x_t] + b_o)$

决定$C_t$的哪个部分被输出？



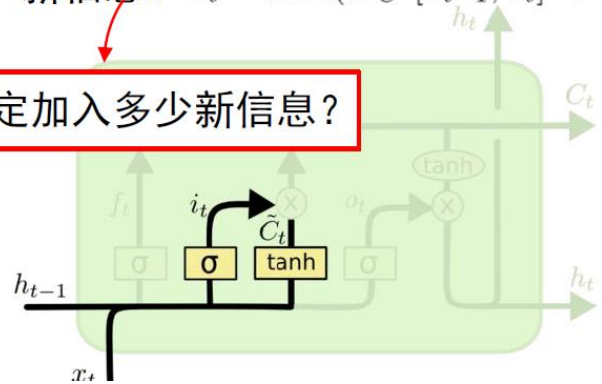

- 细胞状态： $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$

基于当前输入和上个隐状态信息生成的新信息



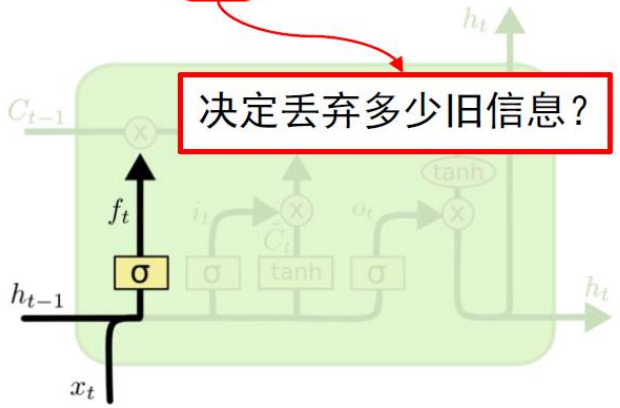

- 输入门： $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$
- 新信息： $\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$

决定加入多少新信息？




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

决定丢弃多少旧信息？

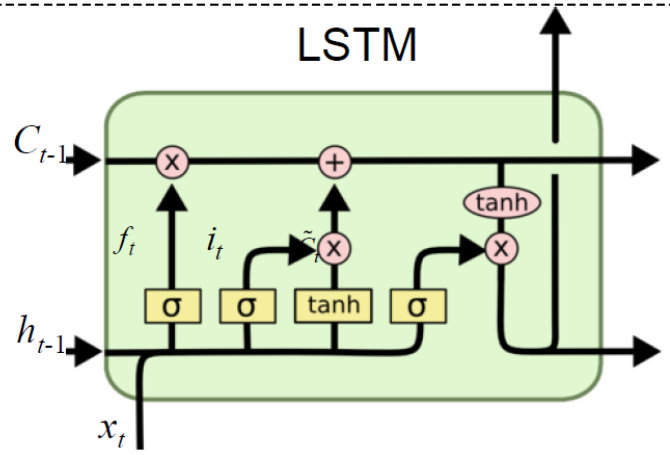


# 门限循环单元：GRU

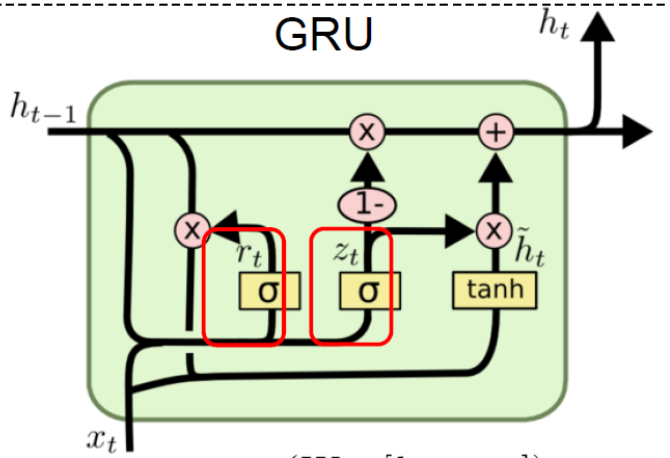
AI DISCOVERY

- 有单独的细胞状态
- 用输入门和遗忘门决定保留或放弃
- 新信息 $\tilde{C_t}$ 来源于 $h_{t-1}$ 和 $x_t$
- 输出门控制细胞状态的输出

- 没有单独的细胞状态
- 用更新门决定保留或放弃
- $\tilde{h_t}$ 由重置门决定来自 $h_{t-1}$ 的信息
- 直接输出隐状态




- 遗忘门 $f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$
- 输入门 $i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)$
- 新信息 $\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$
- 细胞状态 $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$
- 输出门 $o_t = \sigma\left(W_o\ [h_{t-1}, x_t] + b_o\right)$
- 隐状态 $h_t = o_t * \tanh\left(C_t\right)$




- 更新门 $z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$
- 重置门 $r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$
- 新信息 $\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$
- 隐状态 $h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$

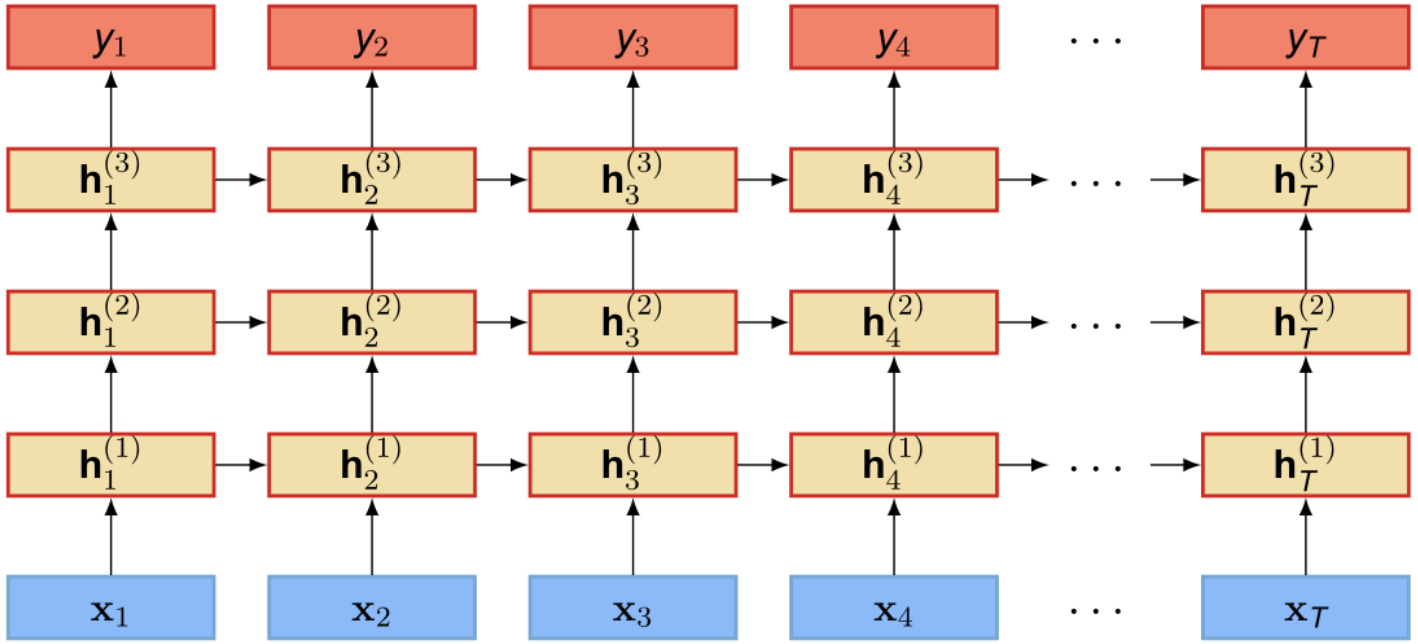
保留哪些旧状态

接收哪些新状态

门限循环单元（Gated Recurrent Unit，GRU）是一种比 LSTM 更加简化的版本。在 LSTM 中，输入门和遗忘门是互补关系，因为同时用两个门比较冗余。GRU 将输入门与和遗忘门合并成一个门：更新门（Update Gate），同时还合并了记忆单元和隐藏神经元。

AI DISCOVERY

# 堆叠(Stack)循环神经网络




◆循环神经网络的深度是一个有争议的话题

✓一方面来说，如果把循环网络按时间展开，不同时刻的状态之间存在非线性连接，循环网络已经是一个非常深的网络了。

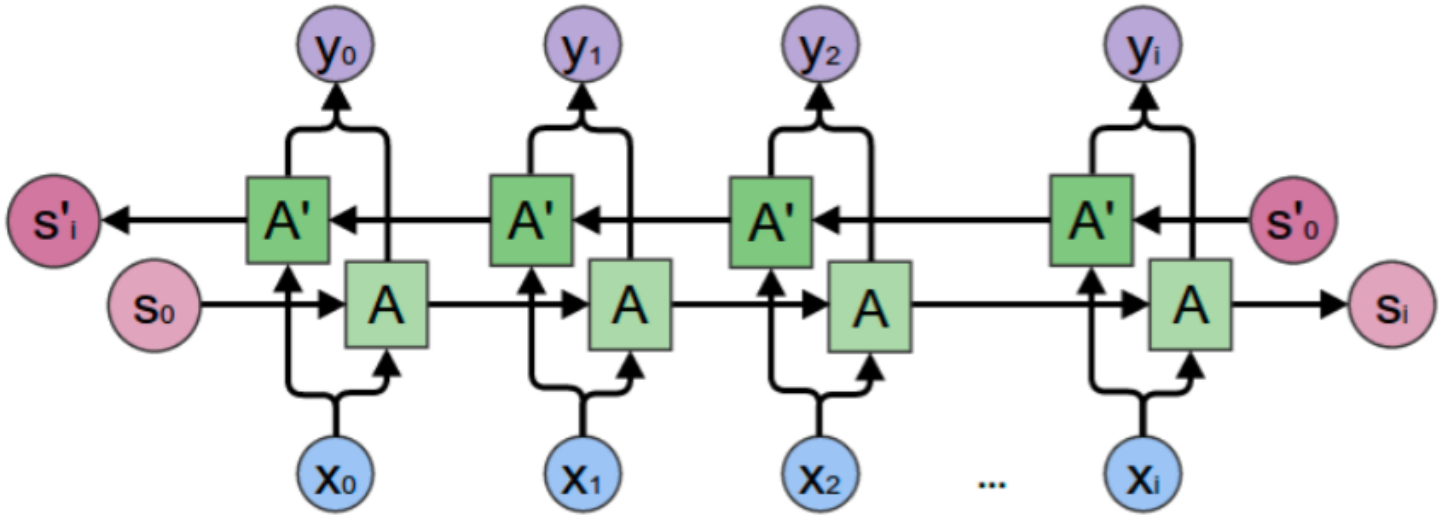
✓另一方面来说，这个网络是非常浅的。隐藏状态到输出（$\boldsymbol{h_{t-1} \rightarrow y_t}$），以及输入到隐藏状态之间（$\boldsymbol{x_t \rightarrow h_t}$）之间的转换只有一个非线性函数。

# 双向循环神经网络

AI DISCOVERY

我今天不舒服，我打算___一天。

只根据‘不舒服‘，可能推出我打算‘去医院‘，‘睡觉‘，‘请假‘等等，但如果加上后面的‘一天‘，能选择的范围就变小了，‘去医院‘这种就不能选了，而‘请假‘‘休息‘之类的被选择概率就会更大




在Forward层**从0时刻到i时刻**正向计算一遍，得到并保存每个时刻**向前**隐含层的输出；
在Backward层**从i时刻到0时刻**反向计算一遍，得到并保存每个时刻**向后**隐含层的输出；
最后在每个时刻结合Forward层和Backward层的相应时刻输出的结果得到最终的输出。

AI DISCOVERY


# 自然语言处理



RNN

LSTM、GRU

Attention


# Seq2Seq模型存在的问题

AI DISCOVERY

## 两个问题

- **定长的中间向量$c$限制了模型性能**

Zhuge Liang--*Northern Expedition Memorial*

Permit me to observe: the late Emperor was taken from us before he could finish his life's work, the restoration of the Han. Today, the empire is still divided in three, and our very survival is threatened. Yet still, the officials at court and the soldiers throughout the realm remain loyal to you, your majesty. Because they remember the late emperor, all of them, and they wish to repay his kindness in service to you. This is the moment to extend your divine influence, to honor the memory of the late Emperor and strengthen the morale of your officers. It is not the time to listen to bad advice or close your ears to the suggestions of loyal men. The emperors of the Western Han chose their courtiers wisely, and their dynasty flourished.

- **输入序列的不同部分对于输出序列的重要性不同**

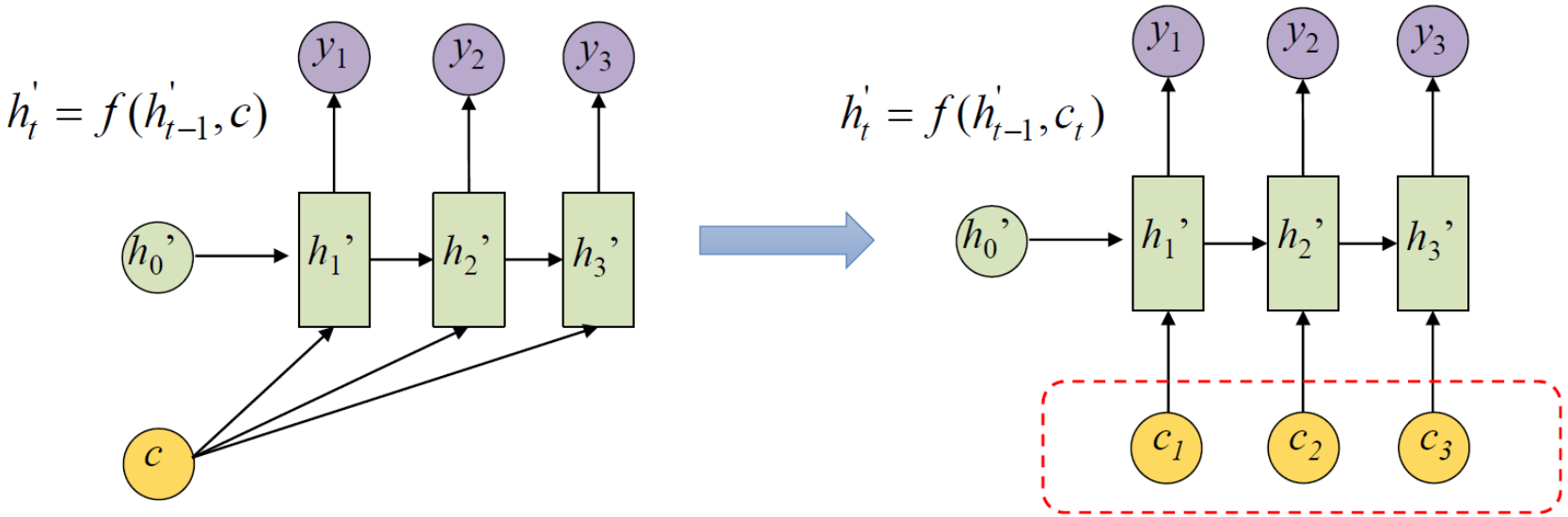
先帝创业未半而中道崩殂，今天下三分，益州疲弊，此诚危急存亡之秋也。然侍卫之臣不懈于内，忠志之士忘身于外者，盖追先帝之殊遇，欲报之于陛下也。诚宜开张圣听，以光先帝遗德，恢弘志士之气，不宜妄自菲薄，引喻失义，以塞忠谏之路也。

AI DISCOVERY

# 注意力机制

AI DISCOVERY

- 解决方案
  - 解码器中的每个时刻不是输入固定的$c$，而是输入不同的$c_i$
  - 每个时刻的$c$自动选取与当前输出最相关的上下文

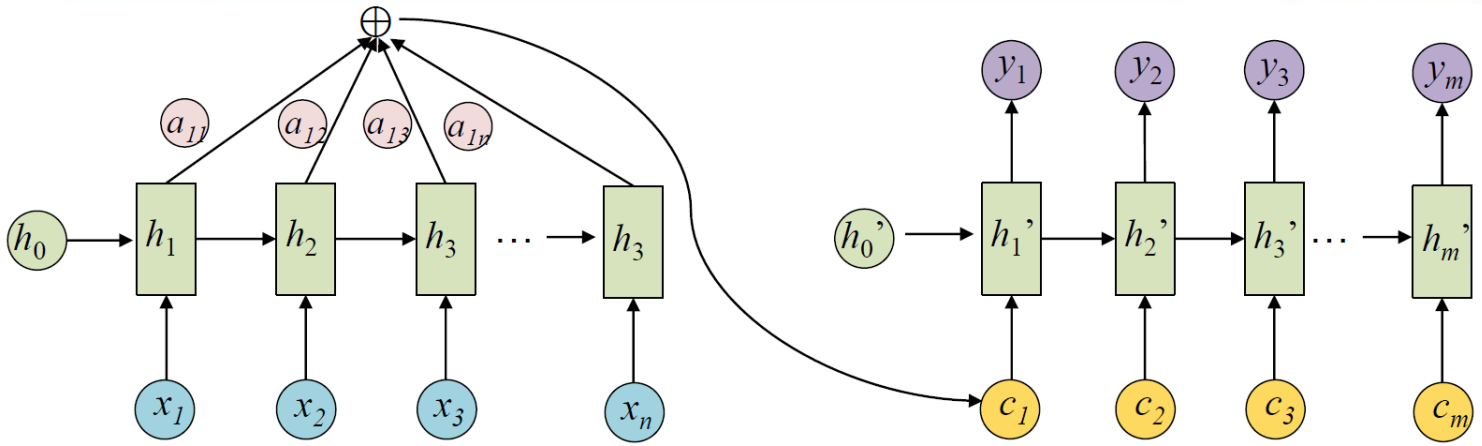



如何计算？

# 注意力机制

AI DISCOVERY

## $C_i$如何计算?




$$c_i = \sum_{j=1}^{T_x} a_{ij} h_j$$

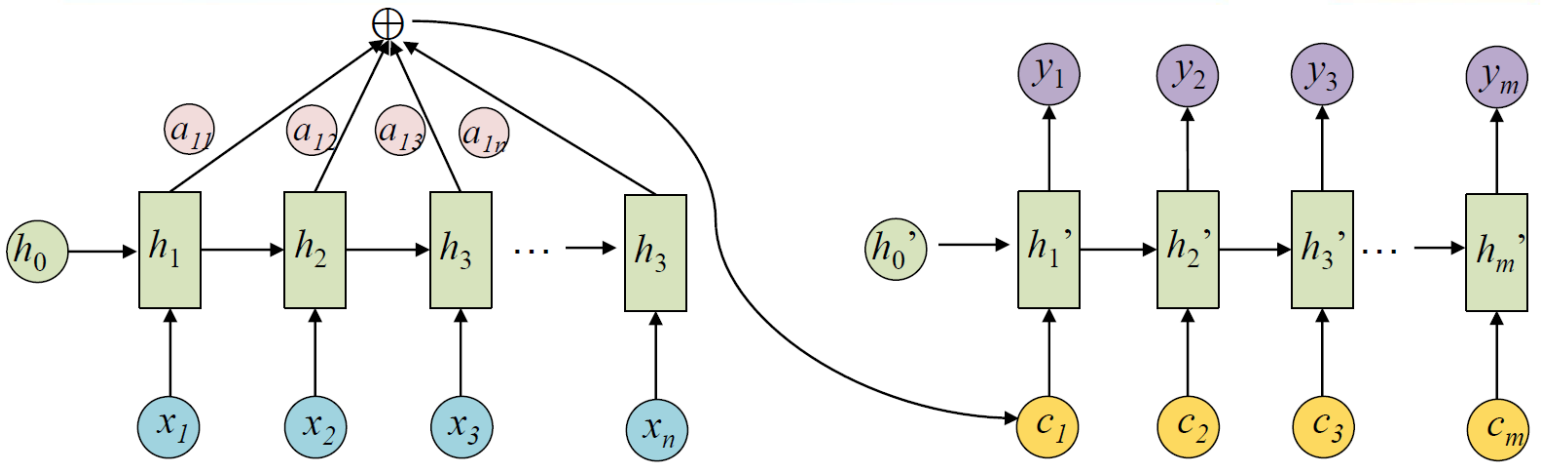
- $c_i$是编码器中隐状态的加权和
- $a_{ij}$是目标词$y_i$与源词$x_j$对齐的概率

AI DISCOVERY


# 注意力机制

AI DISCOVERY

$C_i$如何计算?




Today empire is divided in three

$h_1 * a_{11} + h_2 * a_{12} + h_3 * a_{13} + h_4 * a_{14} + h_5 * a_{15} + h_6 * a_{16} = c_1 \rightarrow$ 今

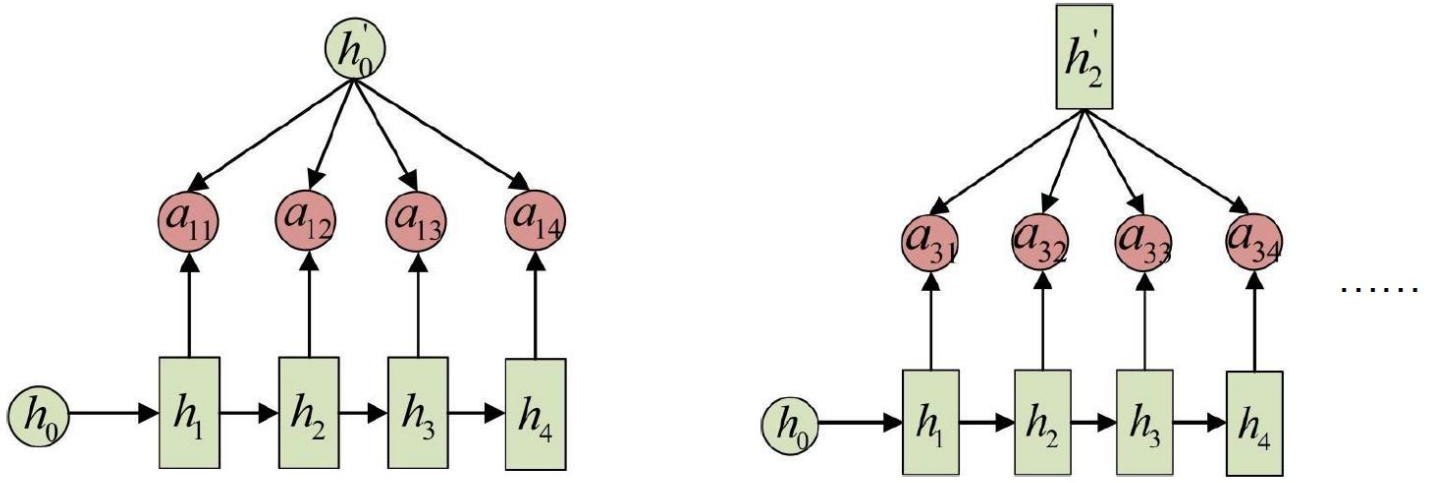
$h_1 * a_{21} + h_2 * a_{22} + h_3 * a_{23} + h_4 * a_{24} + h_5 * a_{25} + h_6 * a_{26} = c_2 \rightarrow$ 天下

$h_1 * a_{31} + h_2 * a_{32} + h_3 * a_{33} + h_4 * a_{34} + h_5 * a_{15} + h_6 * a_{36} = c_1 \rightarrow$ 三分

# 注意力机制

AI DISCOVERY

$a_{ij}$如何计算?




$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}, \quad e_{ij} = \varphi(h'_{i-1}, h_j) = V^T \tanh(Wh'_{i-1} + Uh_j)$$

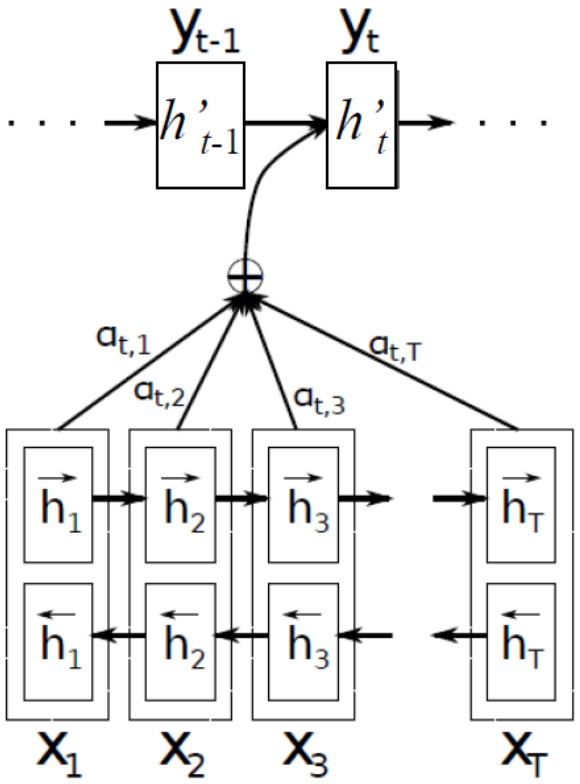
- $a_{ij}$ 是目标词$y_i$与源词$x_j$对齐的概率
- $e_{ij}$ 是$a_{ij}$对应的能量函数
- $\varphi$ 是一个对齐模型，用于衡量$j$位置输入与$i$位置输出的匹配程度

通过**输出端**的前一个状态$\boldsymbol{h'_{i-1}}$，计算**输入端**每一个$\boldsymbol{h_j}$的attention，即$\boldsymbol{a_{ij}}$

AI DISCOVERY

# 机器翻译中的注意力机制





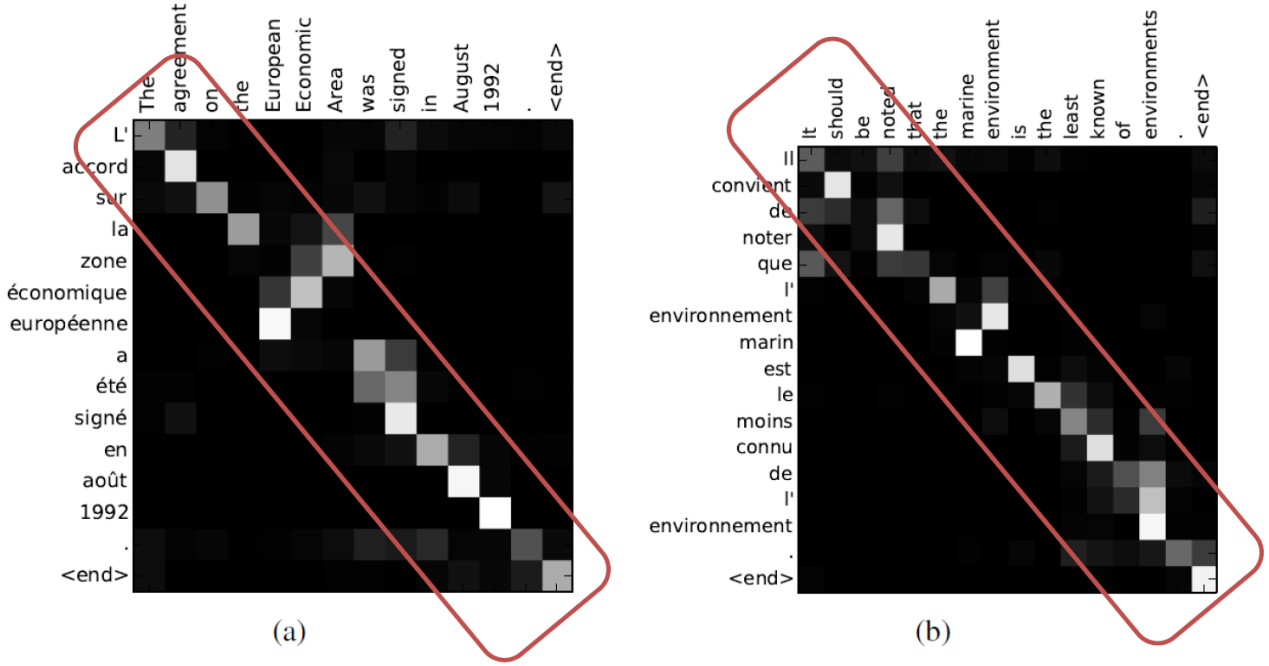




图. 展示注意力机制的双语对齐（英语→法语）效果，每个像素的灰度表示第$j$个源词对于生成第$i$个目标词的权重$a_{ij}$（0:黑，1:白）

<源语言词，目标语言词>的注意力权重基本与两个词互为翻译的情况一致


# 机器阅读中的注意力机制

AI DISCOVERY

给文章中每句话一个attention权重，根据问题选出最有可能包含答案的句子

- **任务：给定一个问题，从候选句子集合中选出答案**

**例　子**

Question1：When did Michael Jordan retired form NBA?

Question2：Which sports does Michael Jordan participates after his retirement from NBA?

Answer：Michael Jordan abruptly retired from Chicago Bulls before the beginning of the 1993-94 NBA season to pursue a career in baseball.


# 目录

AI DISCOVERY

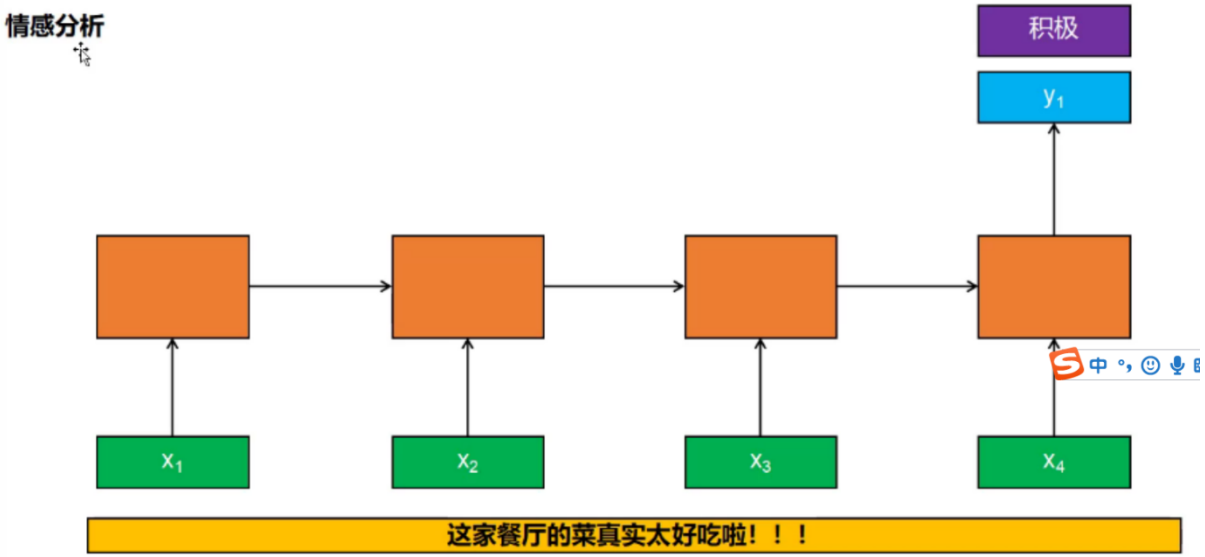


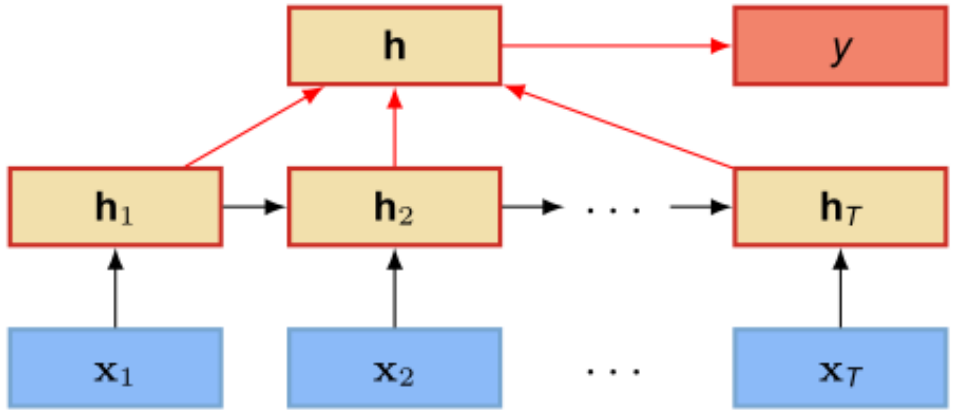
AI DISCOVERY

# RNN应用：序列到类别

AI DISCOVERY

✓输入为序列，输出为类别。比如在文本分类中，**输入数据为单词的序列，输出为该文本的类别。**



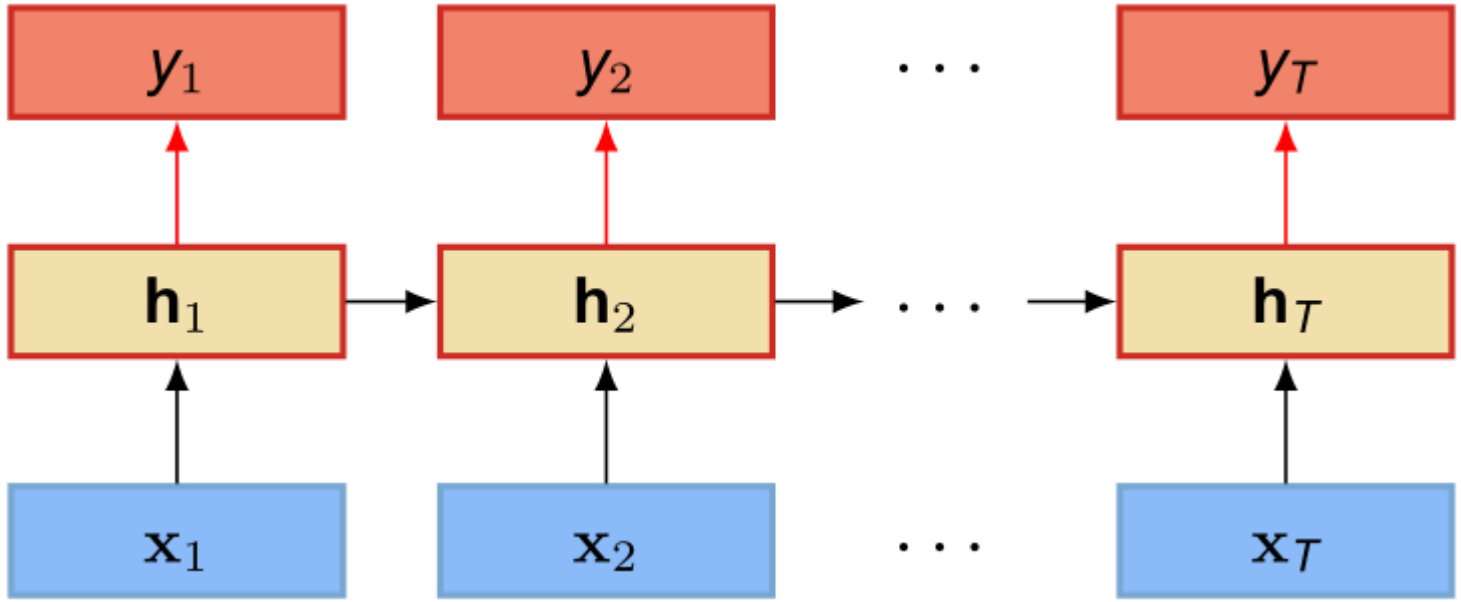




AI DISCOVERY


# 应用：同步序列到序列



✓**输入和输出同步，即每一时刻都有输入和输出**。比如在序列标注问题，每个时刻的输入都需要有一个输出。**输入序列和输出序列的长度相同**。



# 同步序列到序列——中文分词

AI DISCOVERY

中文分词：

任何网购退款均无需提供

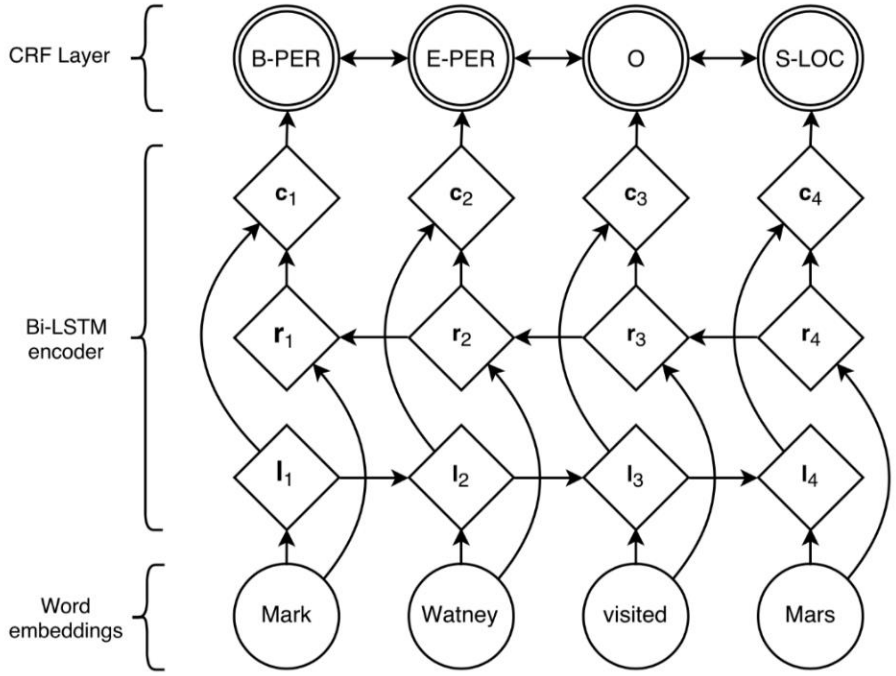
任何 | 网购退款 | 均 | 无需 | 提供

任(B)何(E) 网(B)购(I)退(I)款(E)均(O)无(B)需(E)提(B)供(E)

字级别的序列标注

输入：汉字序列

输出：标签序列



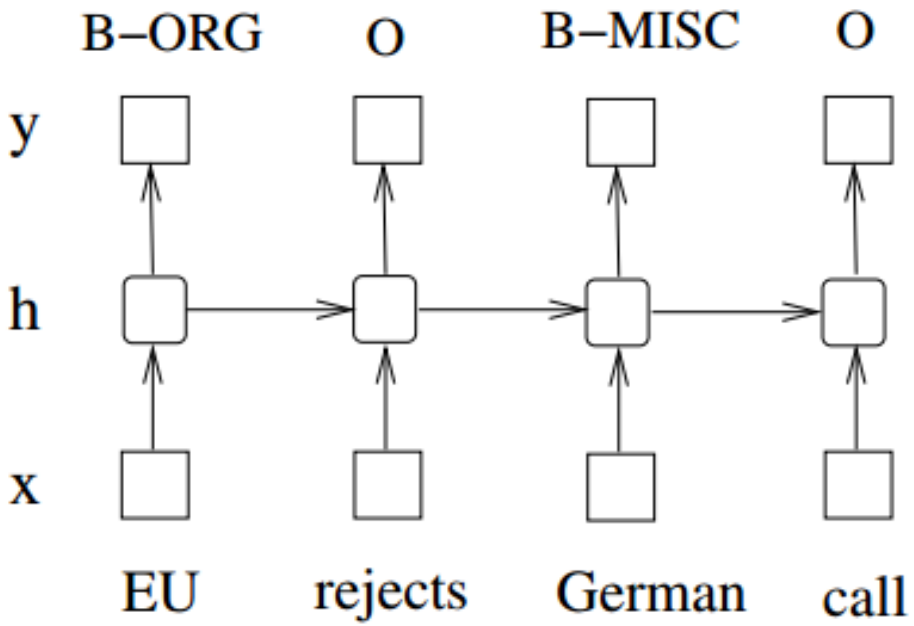

AI DISCOVERY

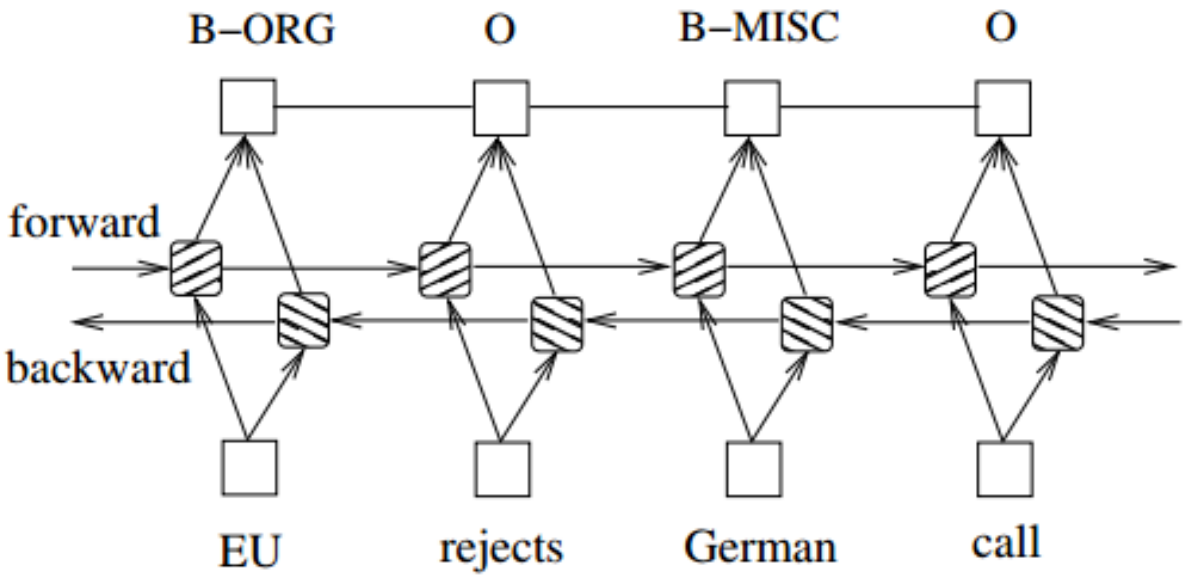
# 同步序列到序列——命名实体识别

AI DISCOVERY

输入：单词序列，每个时刻的输入状态是一个单词

输出：为每个单词输出一个标签，标识每个单词是否是特定命名实体的起始位置、中间位置、结束位置等

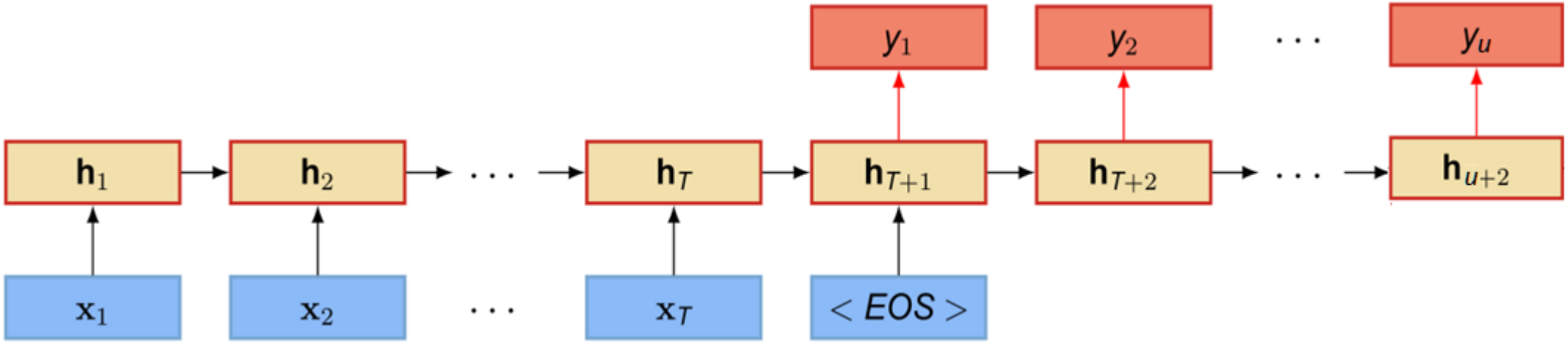


# 应用：异步序列到序列

AI DISCOVERY

✓**输入和输出不需要有严格的对应关系。**比如在机器翻译中，输入为源语言的单词序列，输出为目标语言的单词序列。输入和输出序列并不需要保持相同的长度。



# 传统统计机器翻译

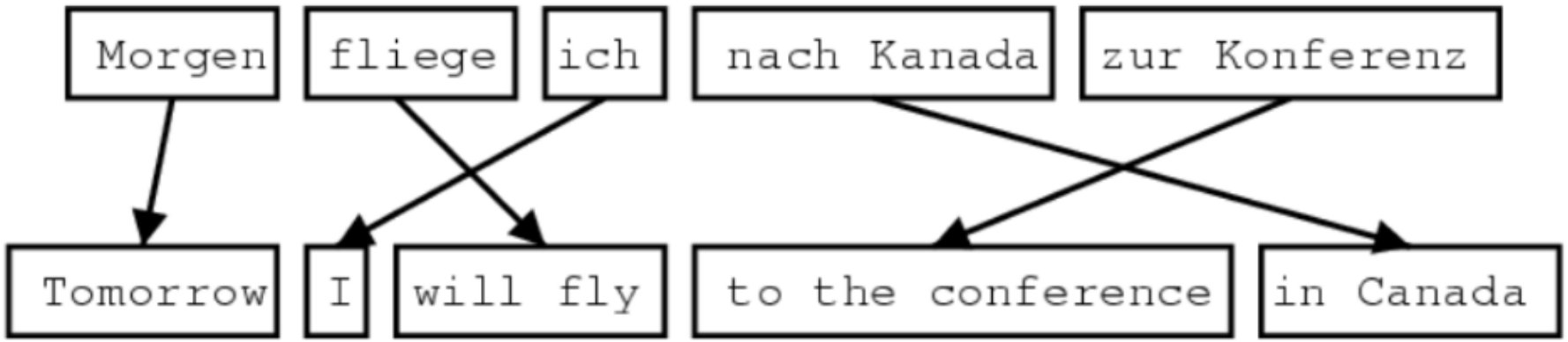
AI DISCOVERY

✓源语言：f

✓目标语言：e

✓模型：$\hat{e} = \text{argmax}_e p(e|f) = \text{argmax}_e\ p(f|e)\ p(e)$

- **p(f|e): 翻译模型**
- **p(e): 语言模型**




AI DISCOVERY

# 异步序列到序列——机器翻译

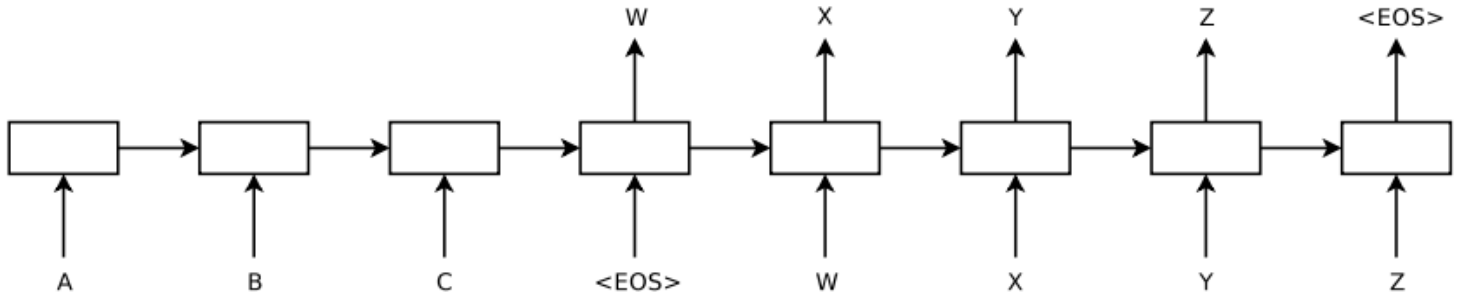
AI DISCOVERY

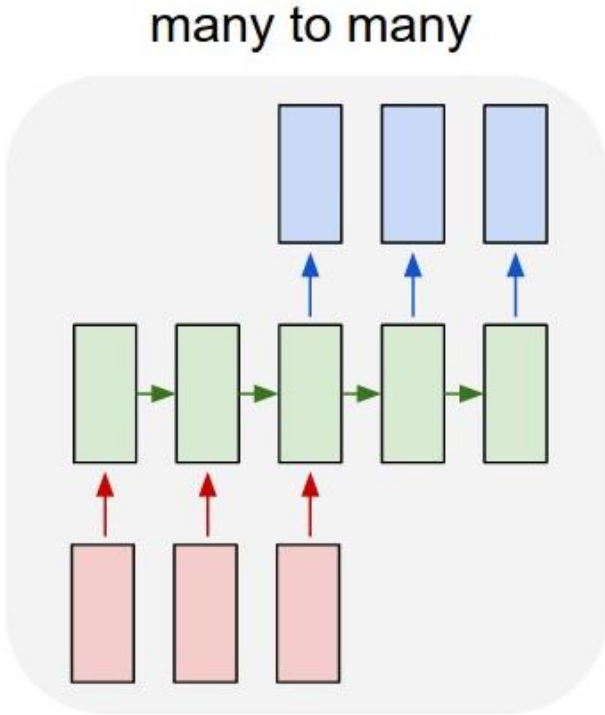
## ◆基于神经网络的机器翻译模型

- ✓一个 RNN 用来编码
- ✓另一个 RNN 用来解码

seq2seq

encoder-decoder




编码：读入源句子（变长向量），转换成一个固定的上下文向量$c$

解码：给定上下文和之前预测的词，预测下一个翻译的词

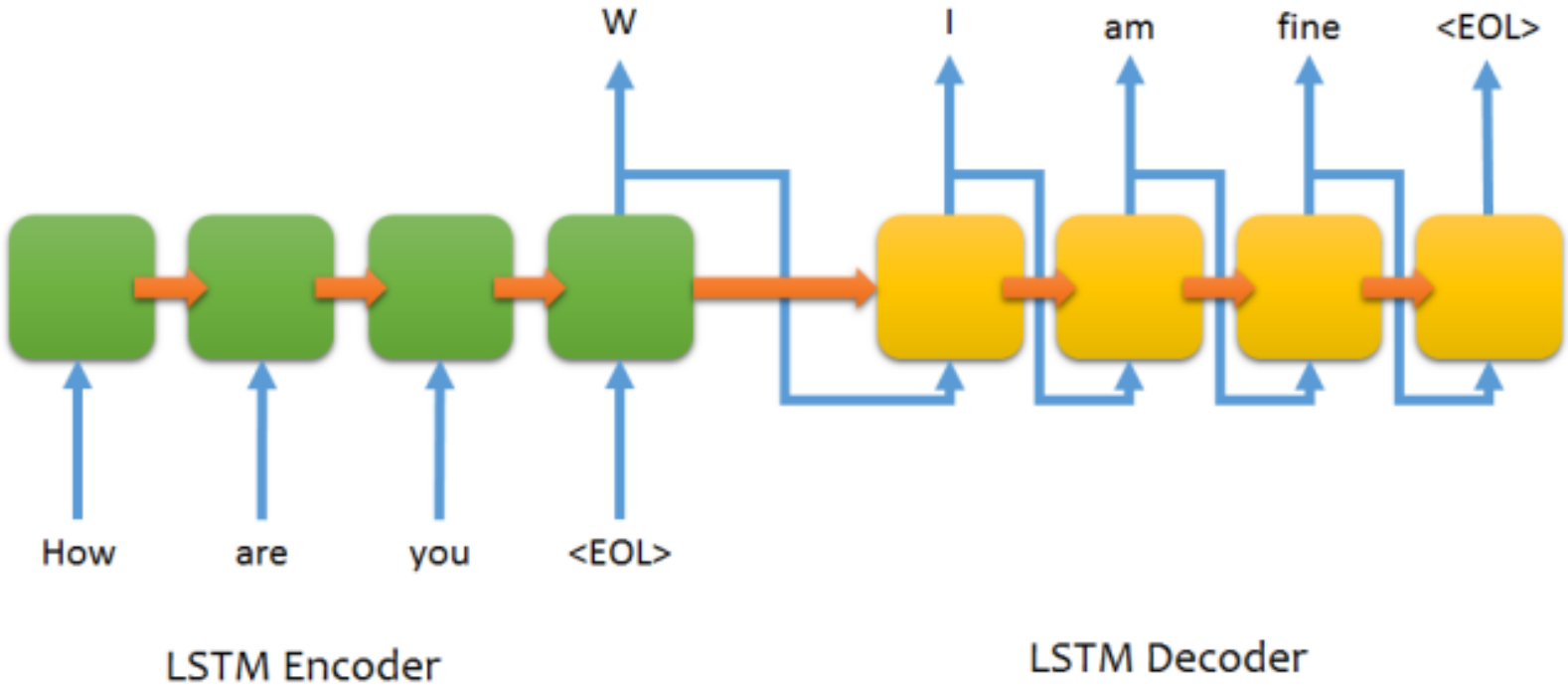
AI DISCOVERY

# 异步序列到序列——对话系统

AI DISCOVERY

Encoder端：对话中的上文（问句），例如 How are you

Decoder端：对话中的下句（回复句），例如 I am fine




AI DISCOVERY


# 其他应用——看图说话

AI DISCOVERY

✓ 输入是一张图片，输出是一句对图片进行描述的文本

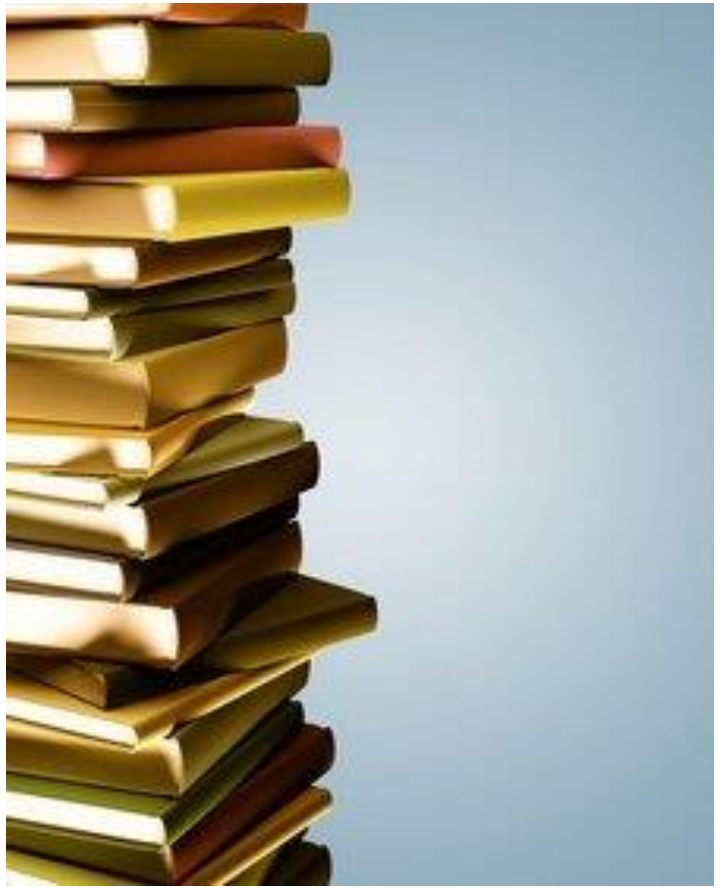
*"Two people are walking down at river in a wooded area"*

AI DISCOVERY


# 其他应用——自动摘要

AI DISCOVERY

✓ 输入是一篇文章的原文（长文本），输出是对原文的精简概述（短文本）

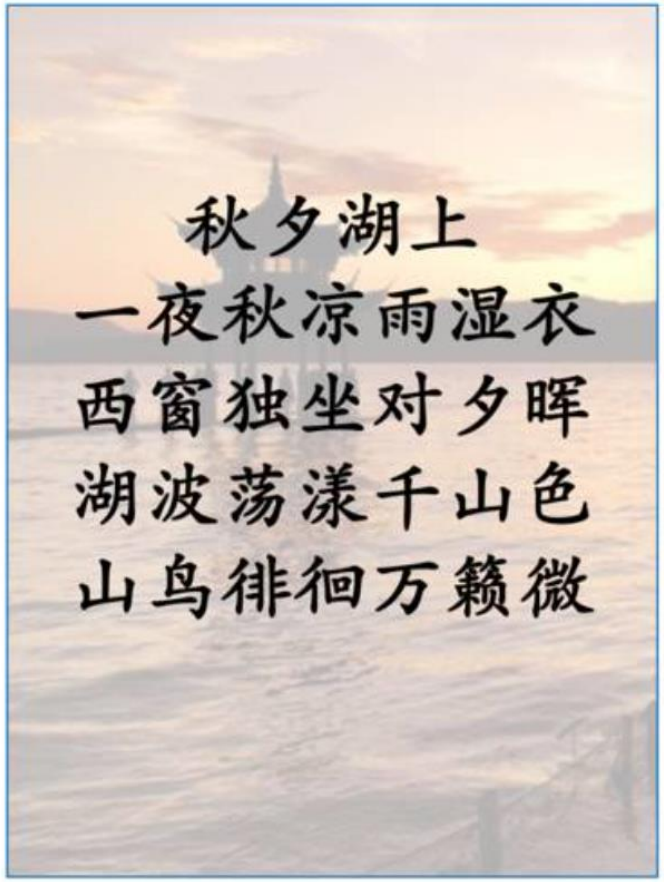


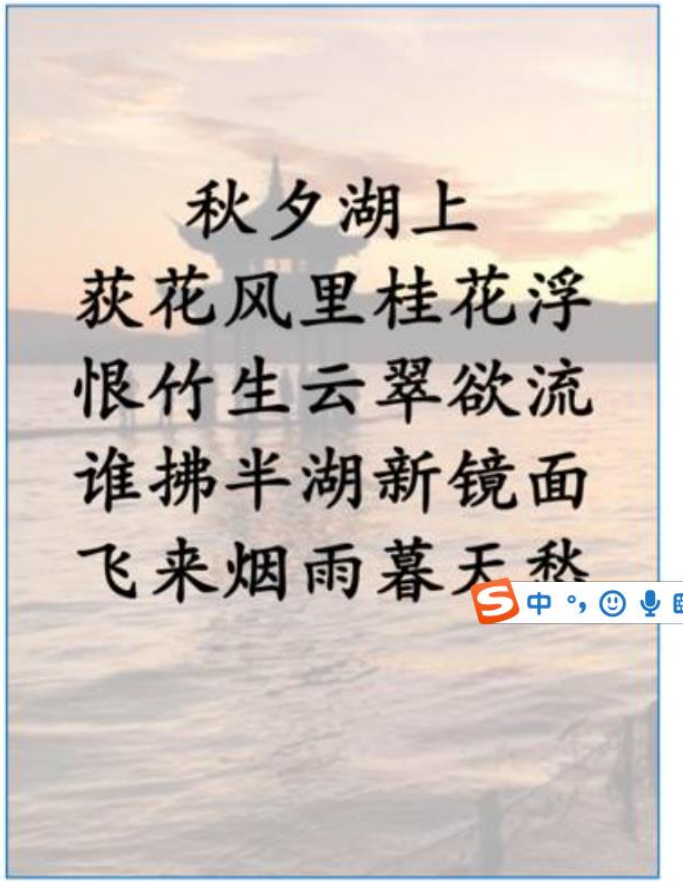

AI DISCOVERY


# 其他应用——自动写诗

AI DISCOVERY

✓ 输入的是诗歌第一句/关键词/标题等信息，输出的是完整的一首诗。




两首诗中有一首是计算机生成的，另一首是宋代诗人葛绍体所作，请读者猜一下哪首是计算机写的诗?


# 其他应用——自动作曲

AI DISCOVERY

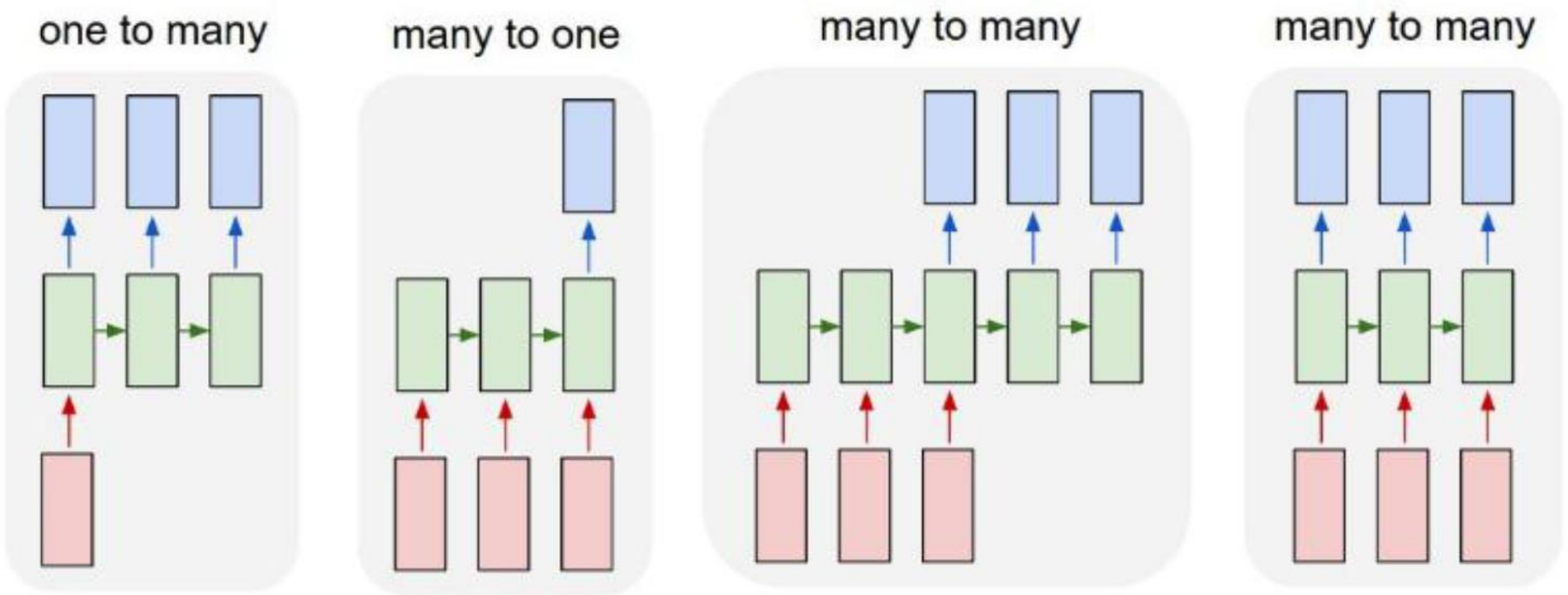
✓ 输入的是前几个旋律，输出的是完整的一首曲子。

哈杰里斯和帕切特开发了一个名为“DeepBach”(深度巴赫)的神经网络。经过训练，DeepBach能够创作出与巴赫风格高度相近的作品，几乎到了“以假乱真”的地步。

AI DISCOVERY

# RNN在NLP中的常见任务和应用

AI DISCOVERY




图像生成文字　　文本分类　　自动文摘、机器翻译等　　序列标注

AI DISCOVERY


# 课程实践

AI DISCOVERY

实践：文本分类、电影评论情感分析