

Natural Language Processing

Dong Xu

*EECS Department
C. S. Bond Life Sciences Center
Informatics Institute
University of Missouri, Columbia
<http://digbio.missouri.edu>*



Coverage

- Introduction to NLP
- NLP components and basic methods
- Structure of RNN
- Long Short-term Memory (LSTM)
- Basic language models
- BERT
- NLP applications

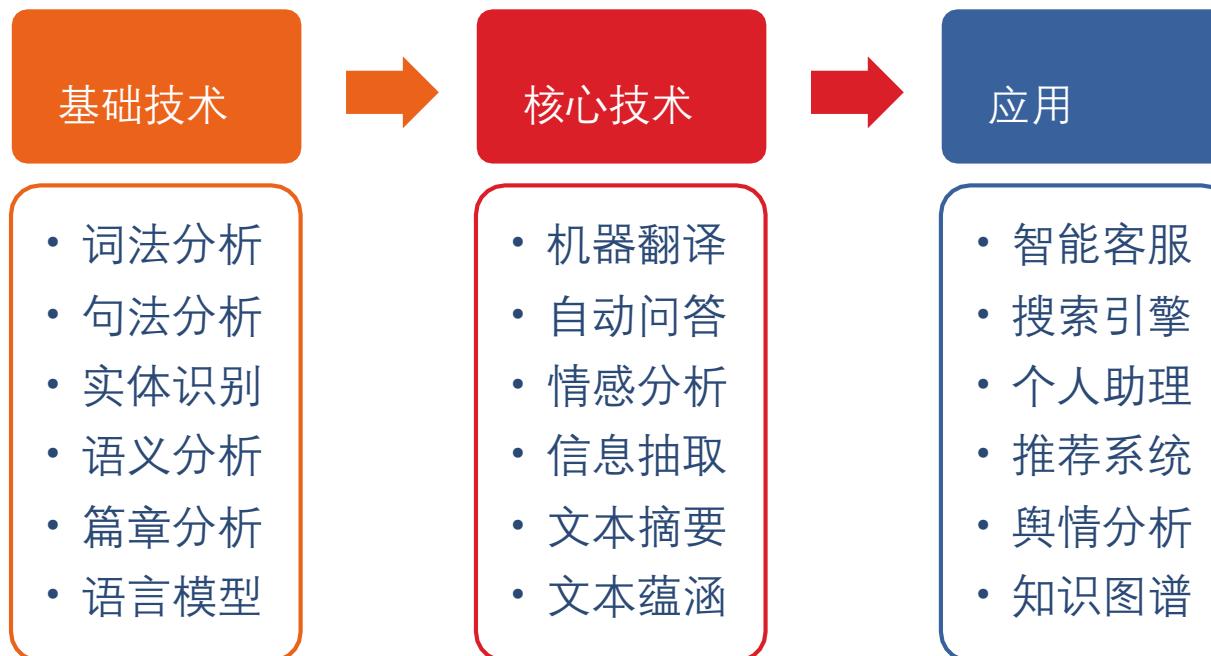
什么是自然语言处理 (NLP)

► 自然语言≈人类语言

□ 区别于人工语言 (比如程序语言)

□ 自然语言处理包括语音识别、自然语言理解、自然语言生成、人机交互以及所涉及的中间阶段。

□ 是人工智能和计算机科学的子学科。



自然语言处理的难点：歧义性

▶ 以中文分词为例

▶ 不同的语言环境中的同形异构现象，按照具体语言环境的语义进行切法。

▶ 交叉歧义

▶ 他 / 说 / 的 / 确实 / 在理

▶ 组合歧义

▶ 两个/人/一起/过去、个人/问题

从马/上/下来、马上/就/来

▶ 句子级歧义

白天鹅在水里游泳

该研究所获得的成果

伪歧义

语义歧义

灵魂八问

配钥匙师傅：你配吗？

食堂阿姨：你要饭吗？

算命先生：你算什么东西？

快递小哥：你是什么东西？

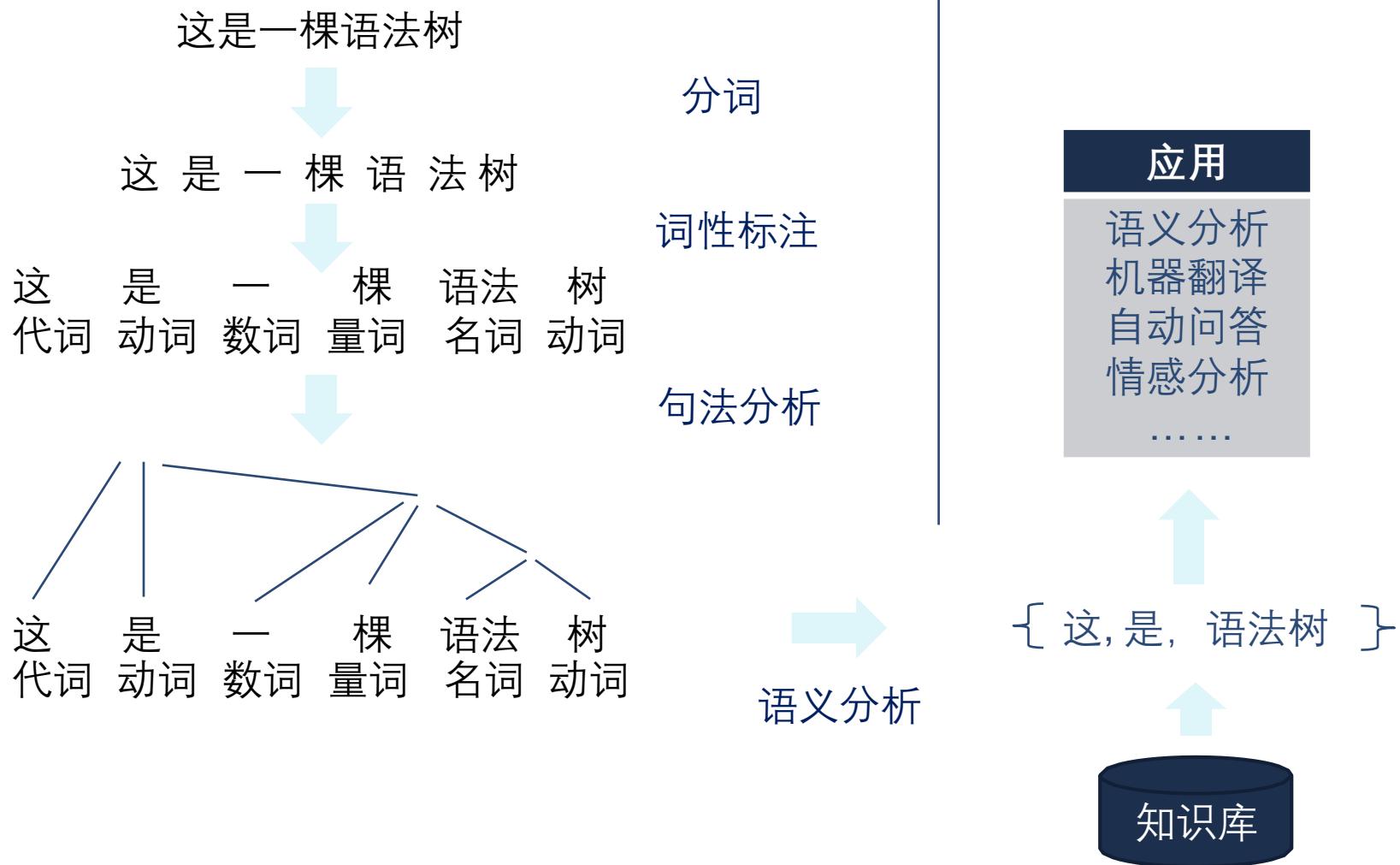
上海垃圾分类阿姨：你是什么垃圾？

滴滴司机：你搞清楚你自己的定位了么？

理发师傅：你自己照照镜子看看你自己，觉得还行么？

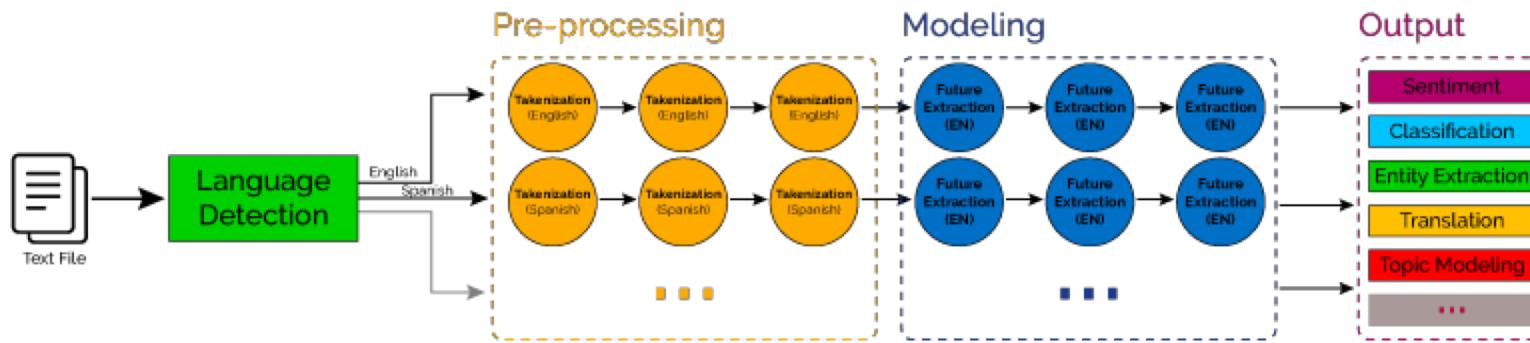
小区保安：你是谁？你从哪里来？要到哪里去？

理想中的NLP技术路线

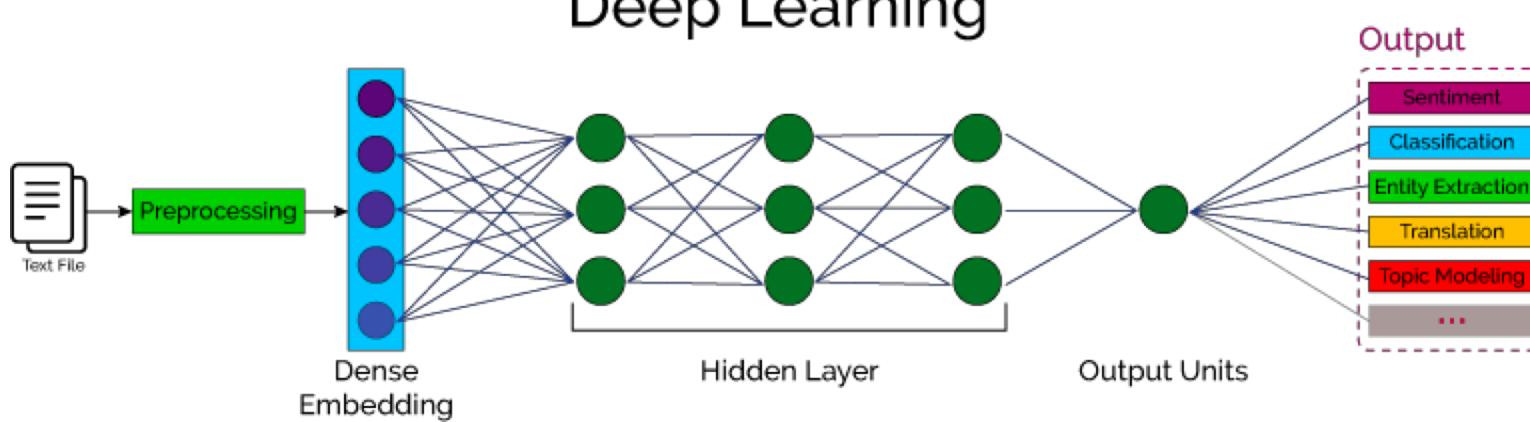


实际的NLP技术路线

Classical NLP



Deep Learning



自然语言处理中的典型任务

分类

文本分类

情感分类

文本匹配

文本蕴涵

序列标注

中文分词

词性标注

信息抽取

生成

机器翻译

文本摘要

风格迁移

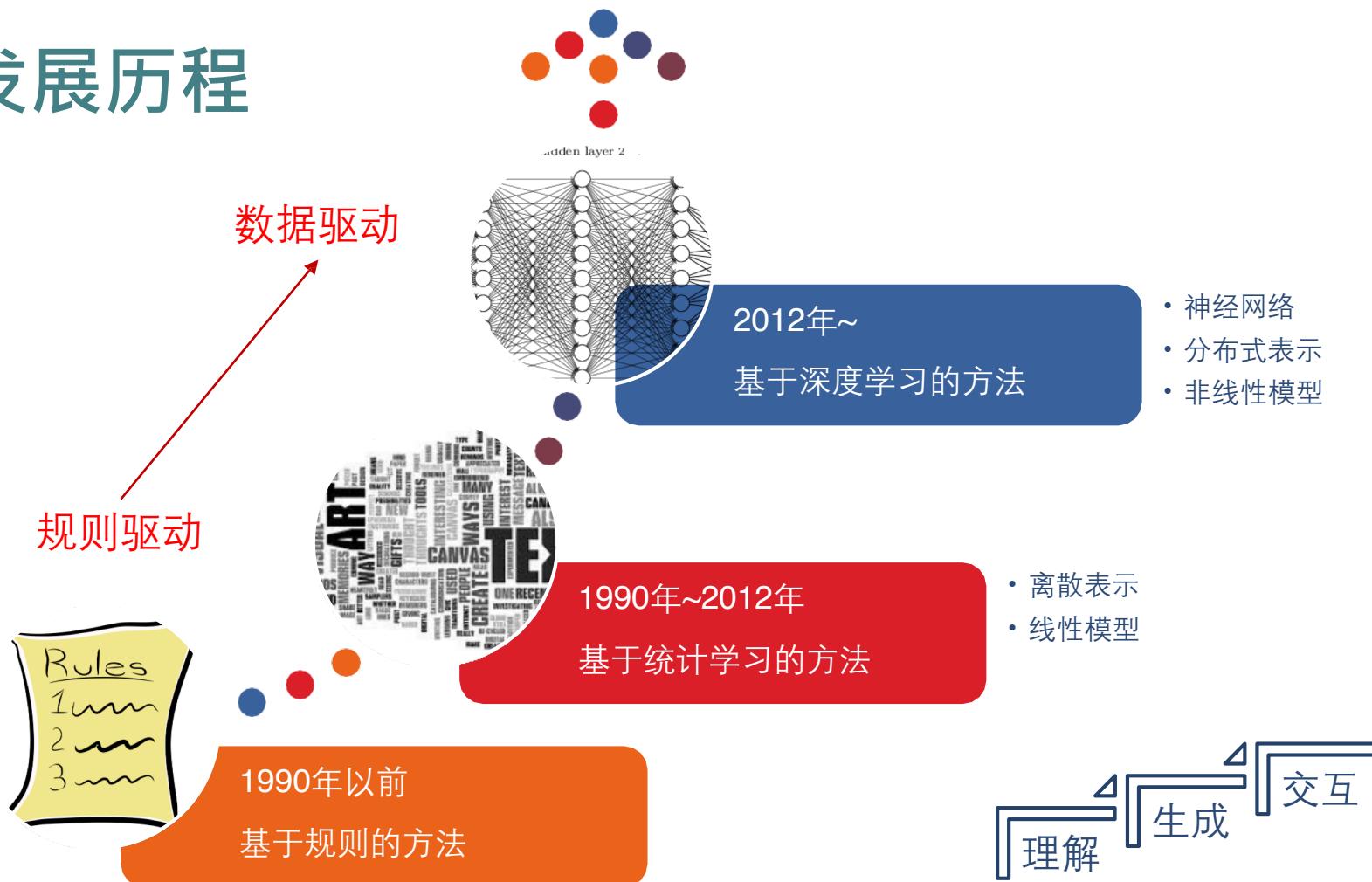
自动问答

对话系统

中文处理任务



发展历程

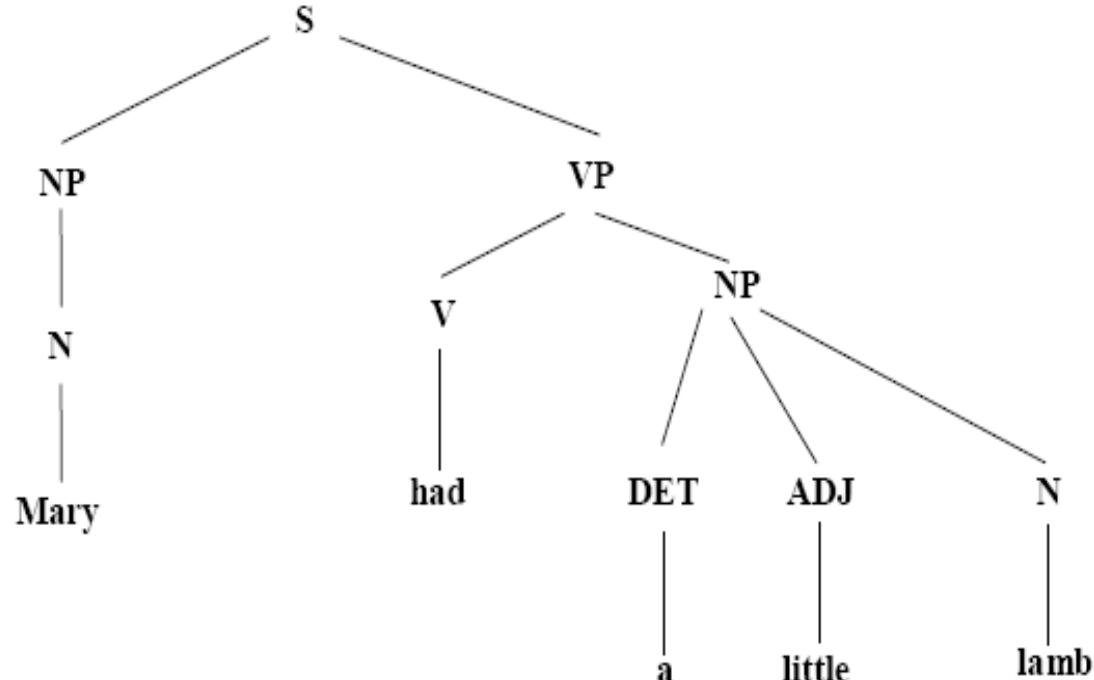


Coverage

- Introduction to NLP
- NLP components and basic methods
- Structure of RNN
- Long Short-term Memory (LSTM)
- Basic language models
- BERT
- NLP applications

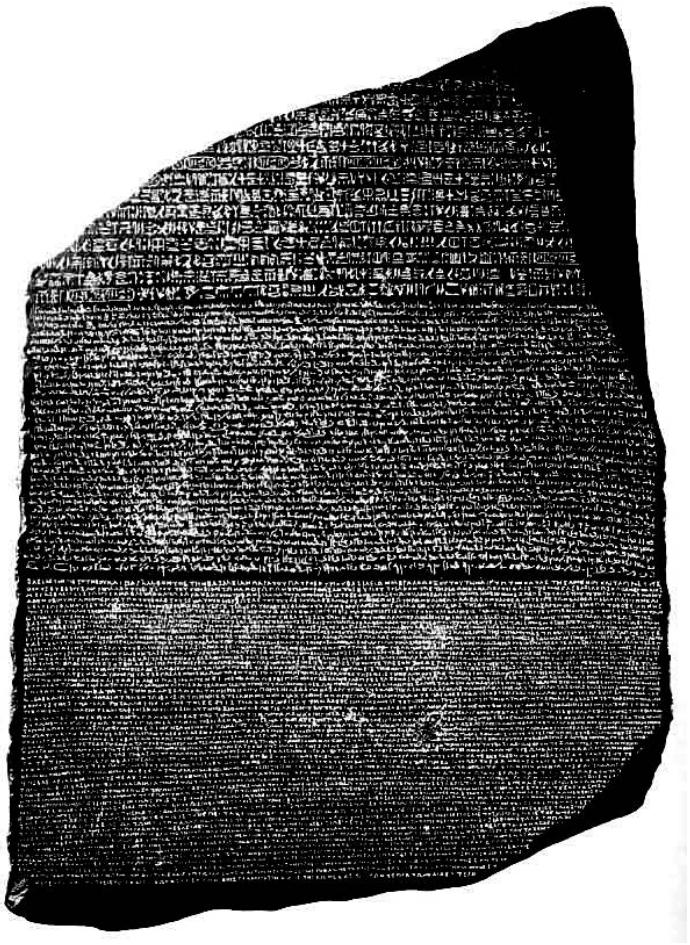
词表(Lexicon)

- 一个内部词典，用来进行句法和语义的分析
- 包含单词或此字符串的语义和语法信息（举例：part-of-speech）



“Mary had a little lamb”的分析树

语料库(Corpora)



- 一个语料库是一堆文本的集合，一般都会被标注，一些情况下只是大量文本。
 - 平衡语料库 vs. 统一语料库
 - 新闻专集: 500M+ words

一个关于“平行文本”的例子：一些文本用两种或多种语言组成：

罗塞塔石碑：古埃及象形文（圣书体），埃及草书（世俗体），古希腊文

平行文本可以用来理解象形文字书写系统

语言学基础知识

- 会一门语言意味着什么?
 - 掌握单词（词表lexicon）
 - 发音，格式，动词的变化
 - 知道单词是如何组成句子的
 - 句式结构，各成分的意义
 - 知道怎么解释句子的意义
 - 陈述，疑问...
 - 知道怎么将句子分组
 - 叙事连贯，对话

语言学基础知识

- **音系学 (Phonology)** – 关注单词与声音之间的关系以及意识到这些关系。
- **形态学 (Morphology)** – 关注如何从中构造单词，更基本的意义单位称为语素。语素是语言中意义的原始单位。
- **句法 (Syntax)** – 关注如何将单词放在一起以形成正确的句子并确定每个单词在句子里的结构作用以及哪些短语是其他句子的子部分。
- **语义 (Semantics)** – 关注单词的含义以及这些含义如何在句子中组合以形成句子含义。对上下文无关意义的研究。

语言学基础知识

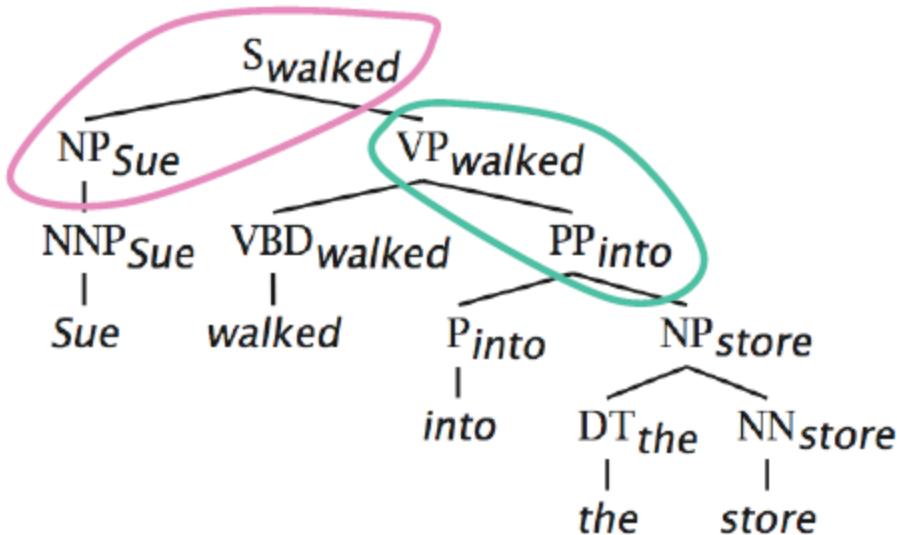
- **语用 (Pragmatics)** – 在不同情况下如何使用句子以及使用如何影响句子的解释。
- **语篇 (Discourse)** – 涉及紧接在前的句子如何影响下一个句子的解释。例如，解释代词和解释信息的时间方面。
- **知识图谱 (Knowledge graph)** – 包含关于世界的一般知识。

句法分析

- 句法（语法）规则规定了句子中单词可以组合的方式，并允许我们确定句子的结构
 - “John saw Mary with a telescope”
 - John saw (Mary with a telescope)
 - John (saw Mary with a telescope)
- 以与我们对世界认知一致的方式将单词和结构映射到特定的领域对象。
- 解析：给出一个句子和一个语法
 - 根据语法检查句子是否正确，如果是，则返回表示句子结构的分析树

分析树

- 句法分析



解析(parsing)与语义分析 (semantic analysis)

- 规则：句法规则或语义规则
 - 哪部分可以与哪部分结合？
 - 组合出的是什么？
- 类别
 - 句法类别：动词，名词，…
 - 语义类别：人，水果，苹果，…
- 分析
 - 识别出一个元素的类别
 - 知道如何将不同的元素组合成一个句子
 - 一个问题：选项通常不唯一

语义分析语法

$S(\text{pred}(\text{obj})) \rightarrow NP(\text{obj})\ VP(\text{pred})$

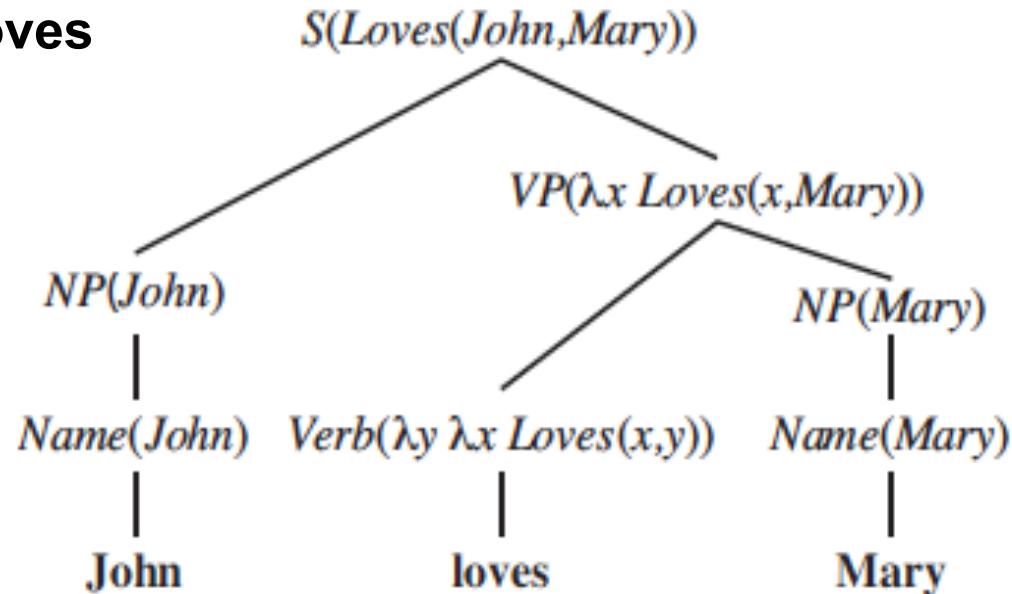
$VP(\text{pred}(\text{obj})) \rightarrow \text{Verb}(\text{pred})\ NP(\text{obj})$

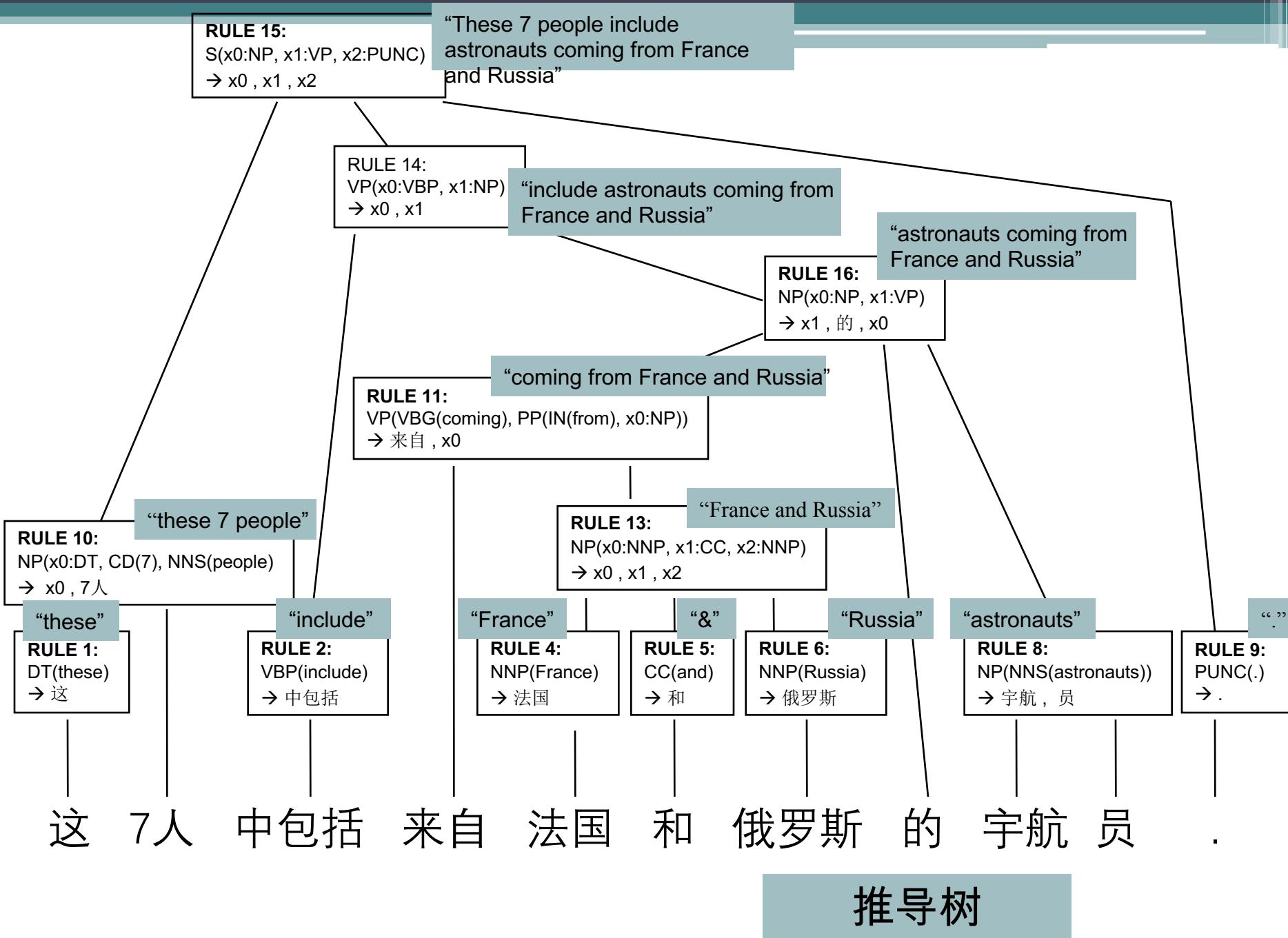
$NP(\text{obj}) \rightarrow \text{Name}(\text{obj})$

$\text{Name}(\text{John}) \rightarrow \textbf{John}$

$\text{Name}(\text{Mary}) \rightarrow \textbf{Mary}$

$\text{Verb}(\lambda y \lambda x \text{ Loves}(x,y)) \rightarrow \textbf{loves}$





语用学 (Pragmatics)

- 语言的实际使用：句子在实践中的含义（意图）
 - 你有空吗？
 - 你好吗？
- 使用上下文
 - 何时，何人，向谁，为什么，何时说
 - 意图：告知，要求，承诺，批评，…
- 处理代词
 - “Mary eats apples. She likes them.”
 - She= “Mary” , them= “apples” .
- 处理歧义
 - 现实中的歧义：“你晚了”，这是批评还是告知？

语篇 (Discourse)

- 语篇是连贯语句（不是任意语句集）的集合
- 语篇也具有层次结构（类似于语句）
- 复指分辨率 **anaphora resolution** -- 解决引用表达式
 - Mary bought a book for Kelly. She didn't like it.
 - She 指 Mary 或 Kelly. – 可能是 Kelly
 - It 指 book.
 - Mary had to lie for Kelly. She didn't like it.
- 语篇结构可能会根据应用改变
 - 独白
 - 对话
 - 人机交互

主题模型

“Arts”	“Budgets”	“Children”	“Education”	法国	全国	教育	产品	卫生
NEW	MILLION	CHILDREN	SCHOOL	欧洲	人大	学生	生产	下乡
FILM	TAX	WOMEN	STUDENTS	德国	常委会	学校	质量	药
SHOW	PROGRAM	PEOPLE	SCHOOLS	欧盟	人民	教师	企业	医疗
MUSIC	BUDGET	CHILD	EDUCATION	法	乔石	大学	工业	健康
MOVIE	BILLION	YEARS	TEACHERS	德	委员长	学	技术	药品
PLAY	FEDERAL	FAMILIES	HIGH	巴黎	届	教学	名牌	农村
MUSICAL	YEAR	WORK	PUBLIC	国	代表大会	高校	服装	医药
BEST	SPENDING	PARENTS	TEACHER	希拉克	委员会	大学生	开发	医院
ACTOR	NEW	SAYS	BENNETT	瑞典	审议	学习	国内	保健
FIRST	STATE	FAMILY	MANIGAT	主题 1	主题 2	主题 3	主题 4	主题 5
YORK	PLAN	WELFARE	NAMPHY					
OPERA	MONEY	MEN	STATE					
THEATER	PROGRAMS	PERCENT	PRESIDENT					
ACTRESS	GOVERNMENT	CARE	ELEMENTARY					
LOVE	CONGRESS	LIFE	HAITI					

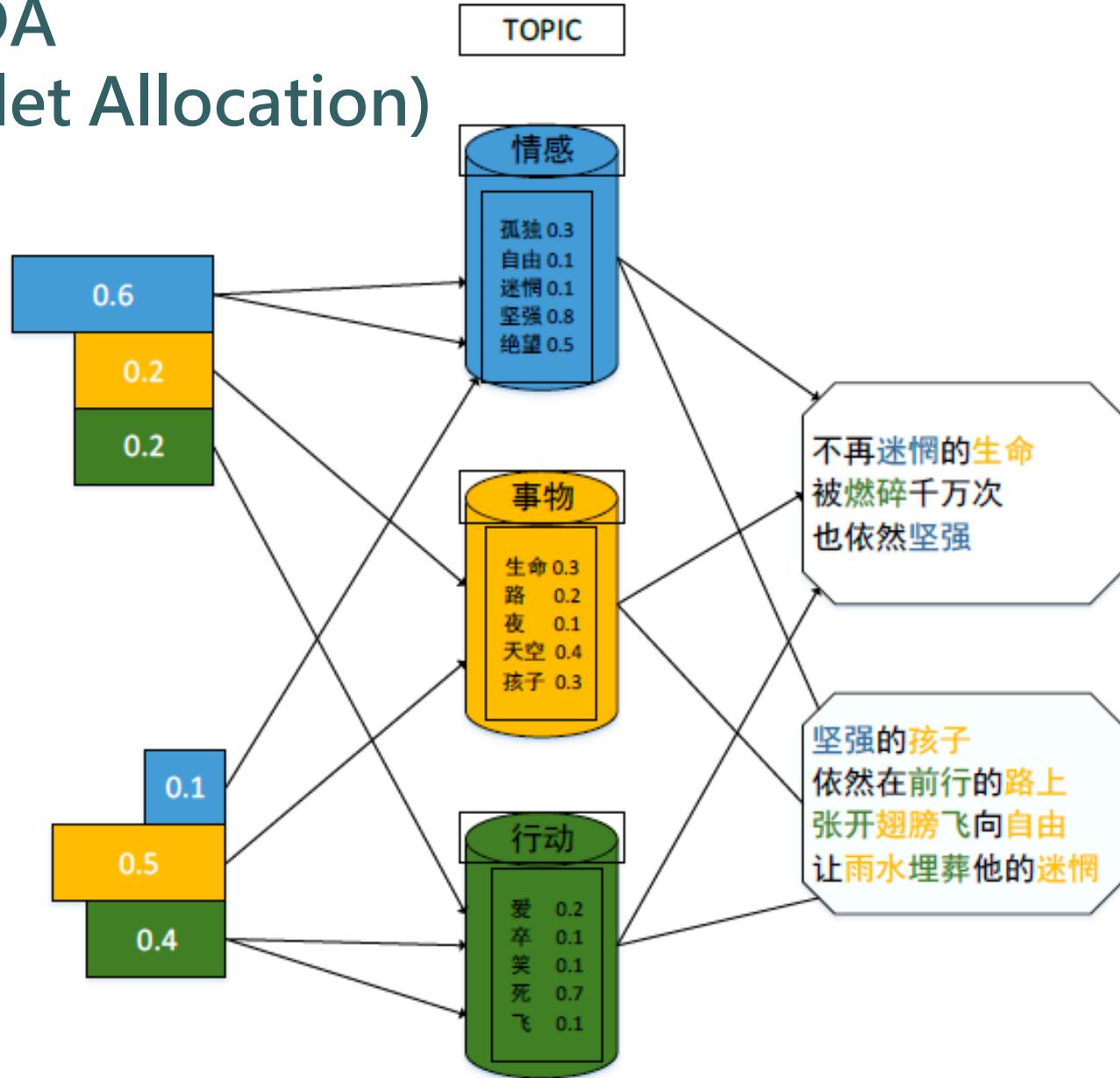
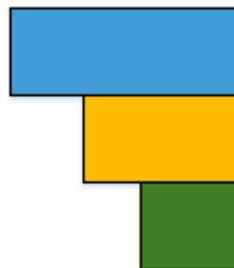
图 1 人民日报语料在 LDA 模型上的训练结果(部分)

主题模型

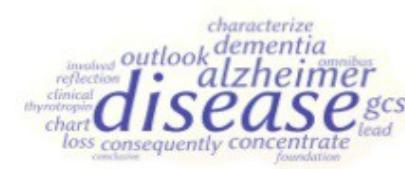
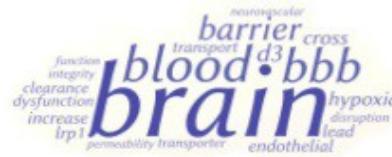
“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

LDA (Latent Dirichlet Allocation)



LDA Example



Natural Language Toolkit (NLTK)

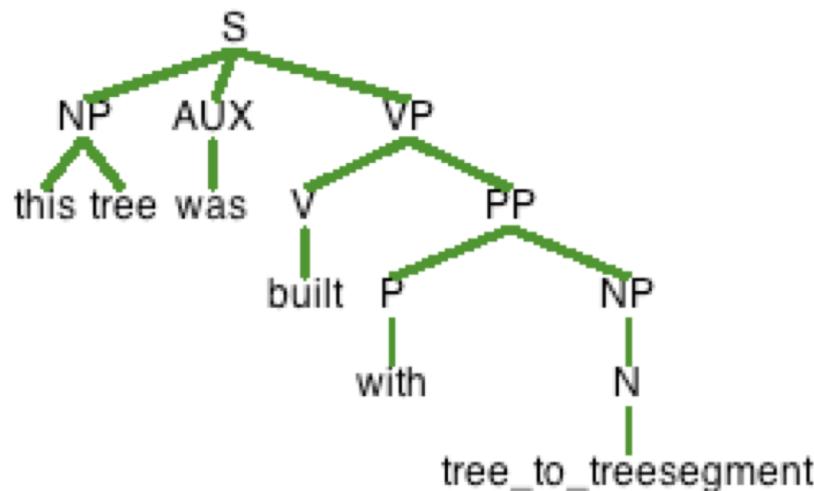
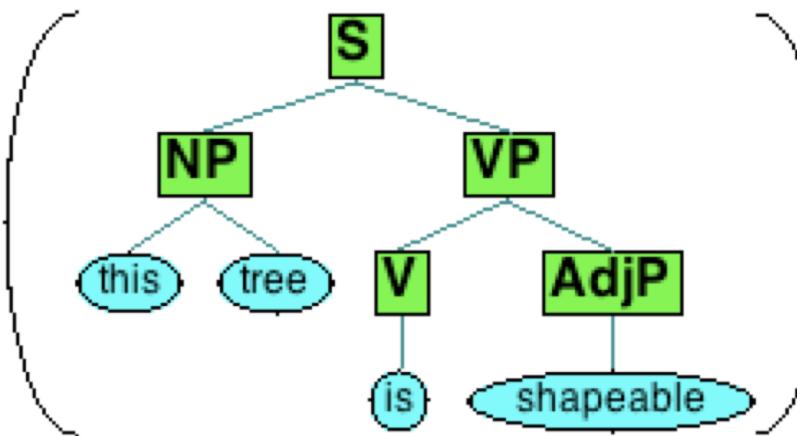
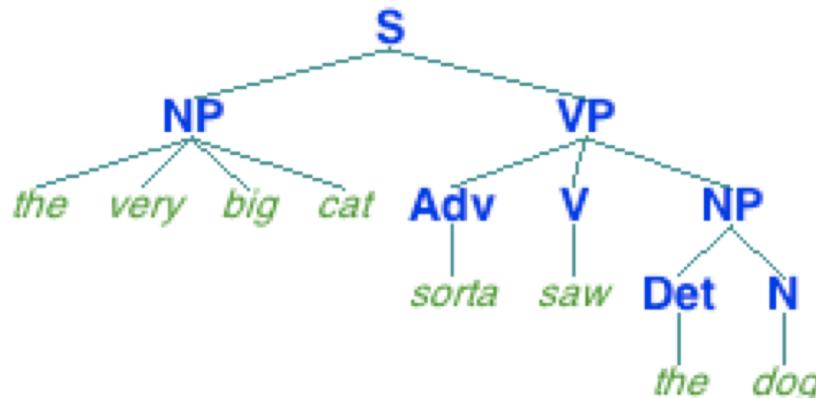
- NLTK是一套开源Python模块， 数据集和教程
- 用来支持自然语言处理的研究与开发
- 从nltk.org下载NLTK
- 所有代码， 数据， 文档完全免费

NLTK内容

1. **代码**: corpus readers (语料库阅读) , tokenizer (标记解析) , stemmers (词干生成) , taggers (标记) , chukers (分块) , parsers (解析器) , wordnet, ... 共有五万多行代码
2. **语料库**: 超过30个标注数据集, 被广泛应用在自然语言处理领域 (超过300Mb数据)
3. **文档**: 超过400页的资料, 包含文章, 评价, 和API文档

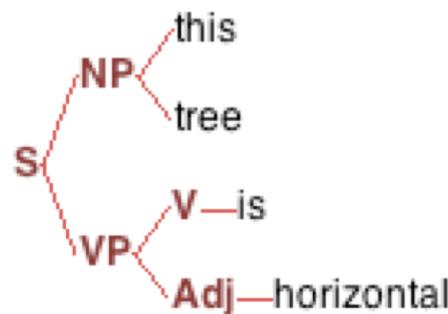
代码内容

- corpus readers 语料库阅读
- tokenizers 标记解析
- stemmers 词干生成
- taggers 标注
- parsers 语法分析
- wordnet 词关系词典
- semantic interpretation 语义解释
- clusterers 聚类
- evaluation metrics 评估指标
- ...



tree_to_treesegment

Try clicking, right clicking, and dragging different elements of each of the trees. The top-left tree is a TreeWidget built from a Tree. The top-right is a TreeWidget built from a Tree, using non-default widget constructors for the nodes & leaves (BoxWidget and OvalWidget). The bottom-left tree is built from tree_to_treesegment.



Coverage

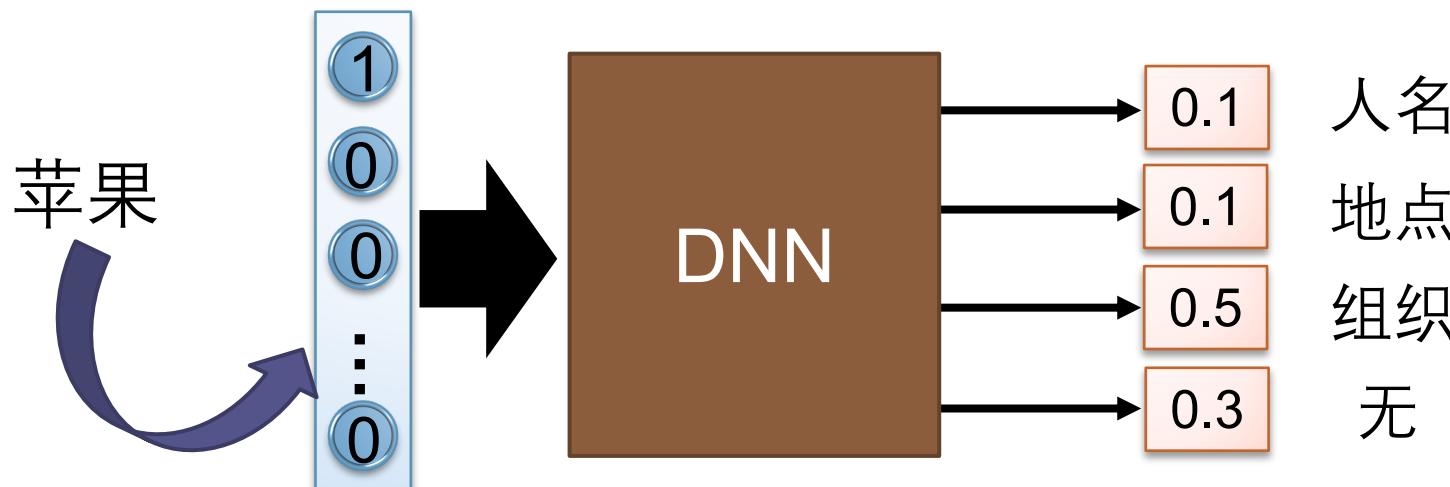
- Introduction to NLP
- NLP components and basic methods
- **Structure of RNN**
- Long Short-term Memory (LSTM)
- Basic language models
- BERT
- NLP applications

DNN/CNN的局限

- 标准的神经网络（以及CNN）应用受限于：
 - 它们只接受固定大小的向量作为输入（例如图像），并产生固定大小的向量作为输出（例如不同类别的概率）。
 - 这些模型使用固定数量的计算步骤（例如，模型中的层数）。
 - 网络中没有存储记忆。

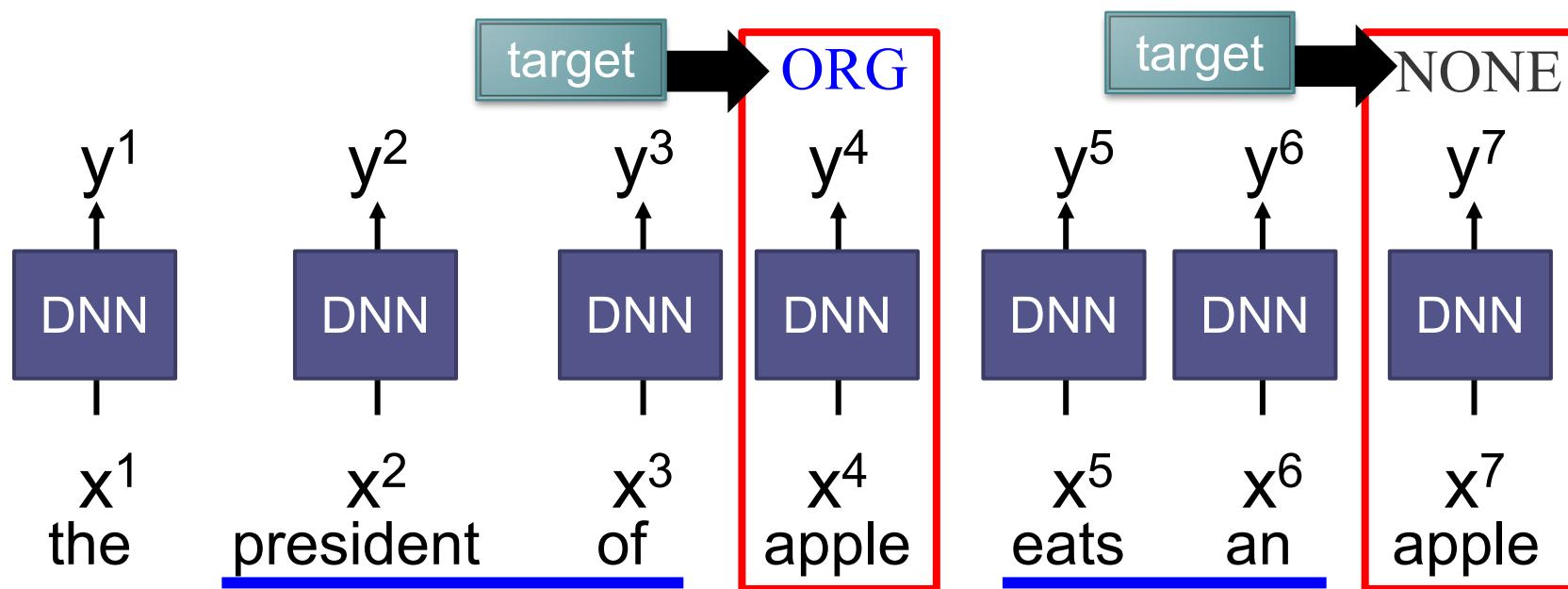
需要存储记录的神经网络

- 名称实体识别
 - 在句子中检查名称实体例如人名，地点，组织等



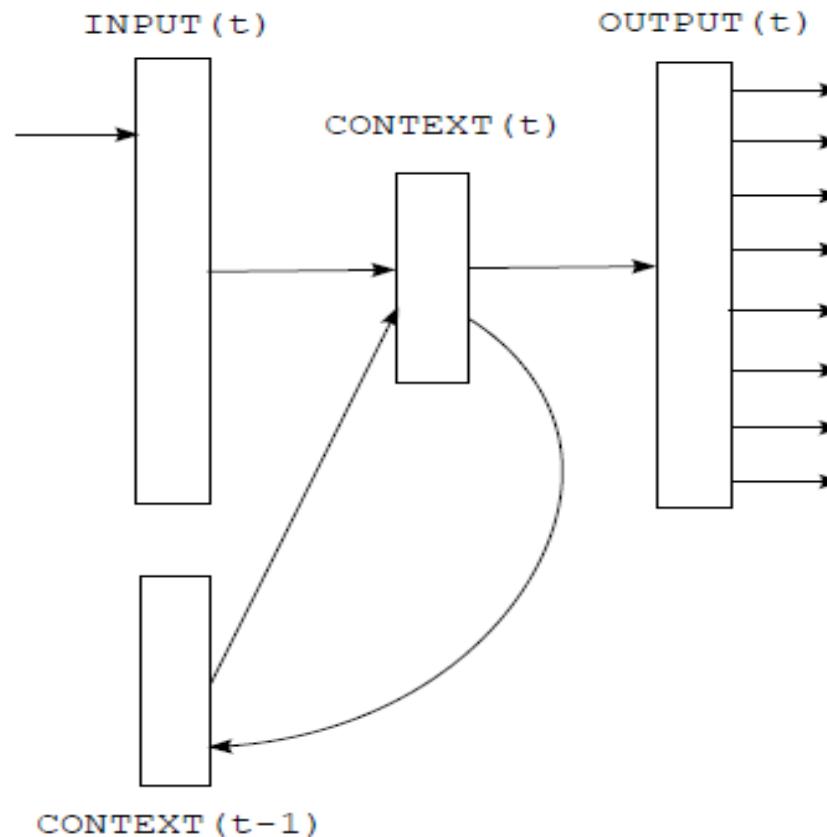
需要存储记录的神经网络

- 名称实体识别
 - 在句子中检查名称实体例如人名，地点，组织等。



循环神经网络 (RNN)

1. 在时间节点t，输入由向量w和前一个时间节点context层的输出连接形成

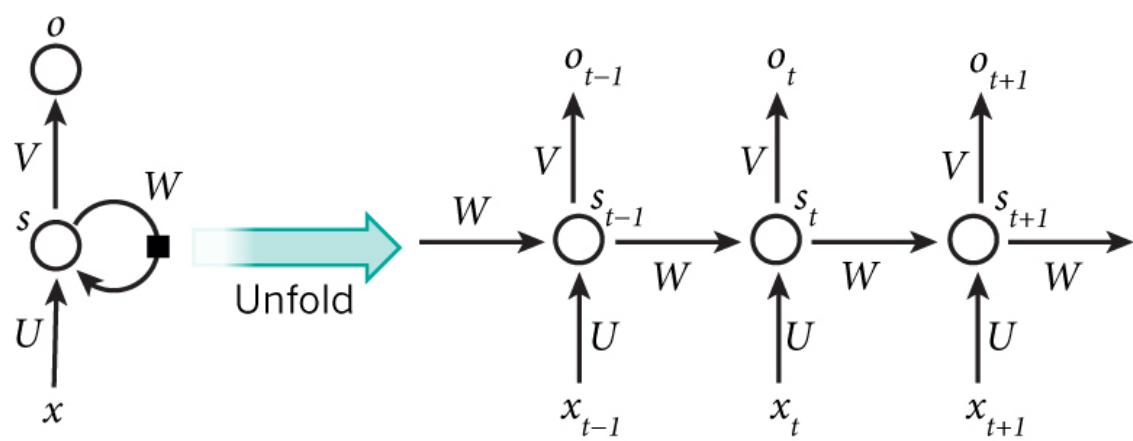


2. 输出为下一个单词的概率

Figure: Recurrent neural network based LM

RNN 的结构

在时间t处的计算基于在t之前收集的所有信息.



$$x(t) = w(t) + s(t-1)$$

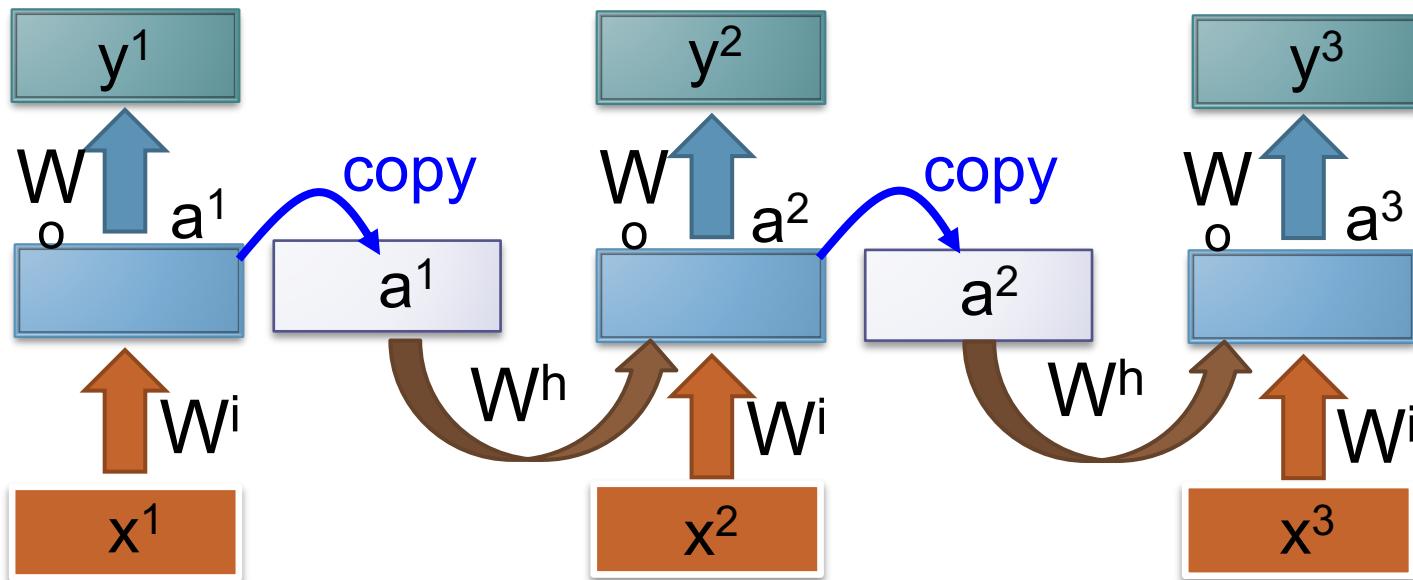
$$s_j(t) = f \left(\sum_i x_i(t) u_{ji} \right)$$

$$y_k(t) = g \left(\sum_j s_j(t) v_{kj} \right)$$

共享权重

每个时间给一个单词，每个epoch算作一次迭代。

RNN 的实现

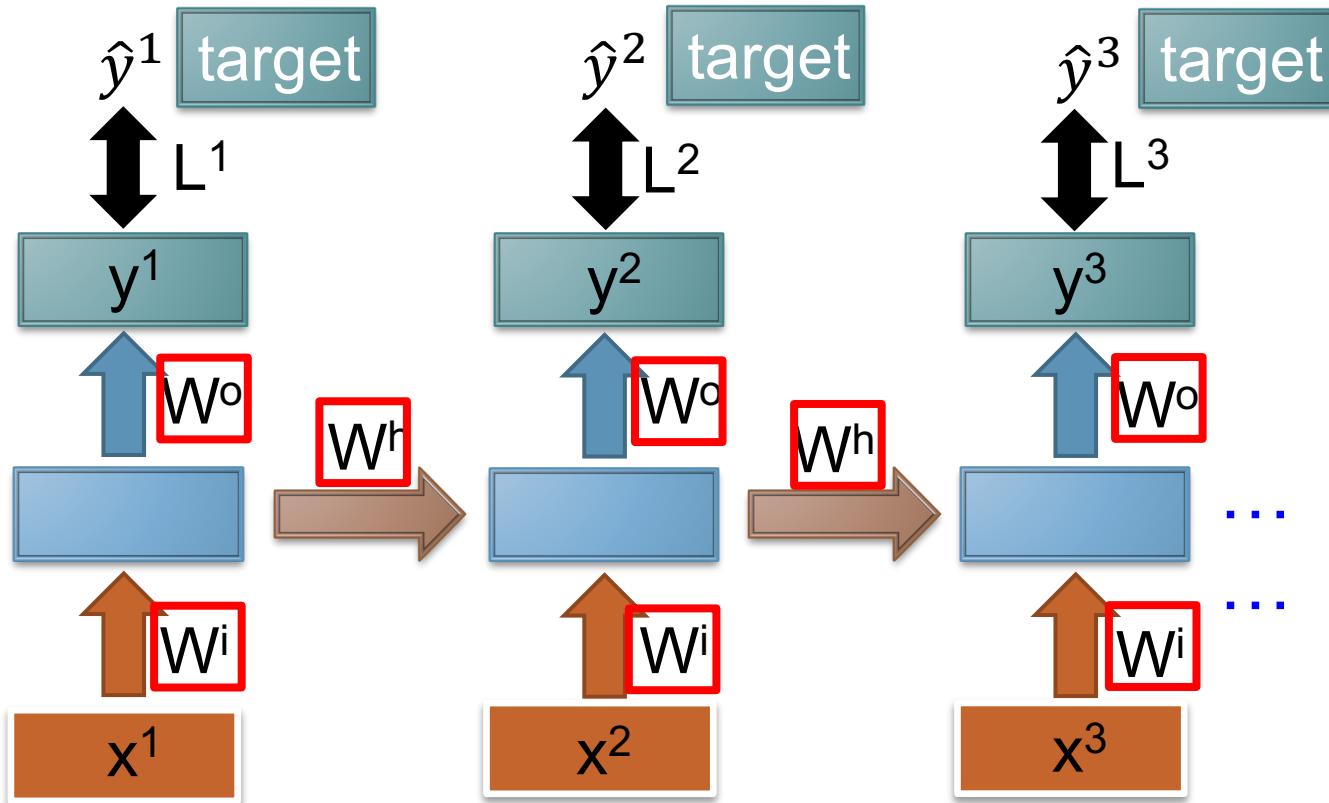


同样的网络结构重复使用.

输出 y^i 基于 x^1, x^2, \dots, x^i

如何训练

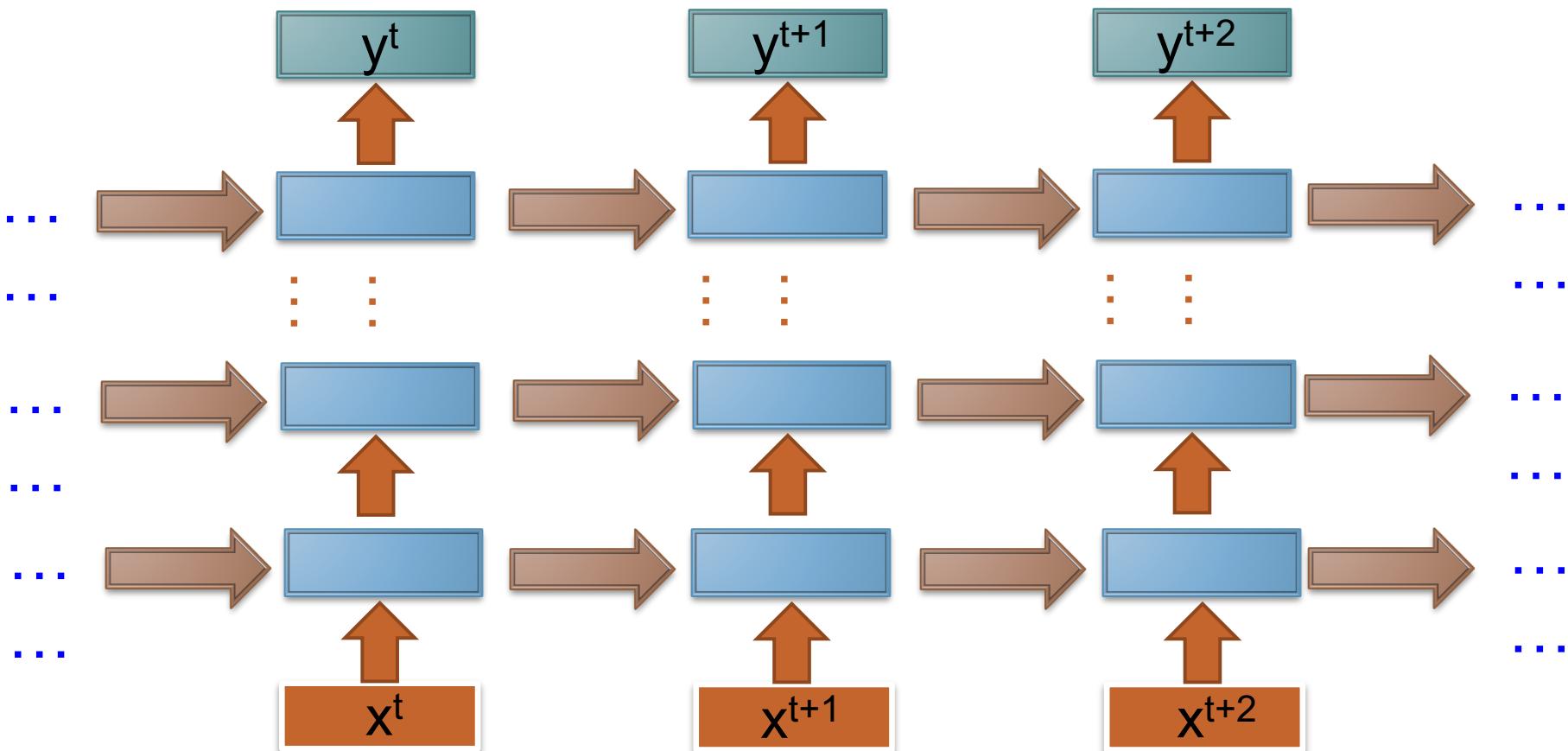
RNN



找到网络参数来最小化总花销：

Backpropagation through time (BPTT)

可以继续加深层数...



RNN 应用

- RNN特殊在它们使我们可以操作向量序列.
 - 一般情况下， 输入和输出向量都是序列
- 我们需要对数据建模， 使用时间或顺序的模型结构， 可变长度的输入和输出
 - 视频逐帧
 - 音频逐段
 - 语句逐词

举例：写一个少儿读物

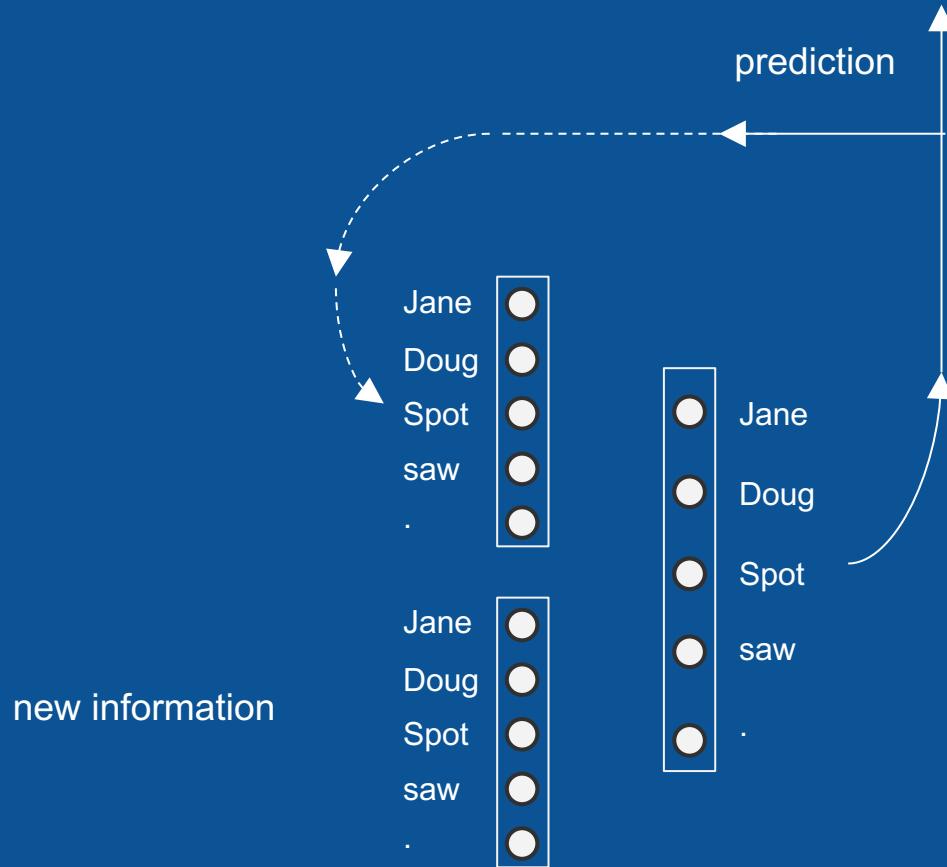
Doug saw Jane.

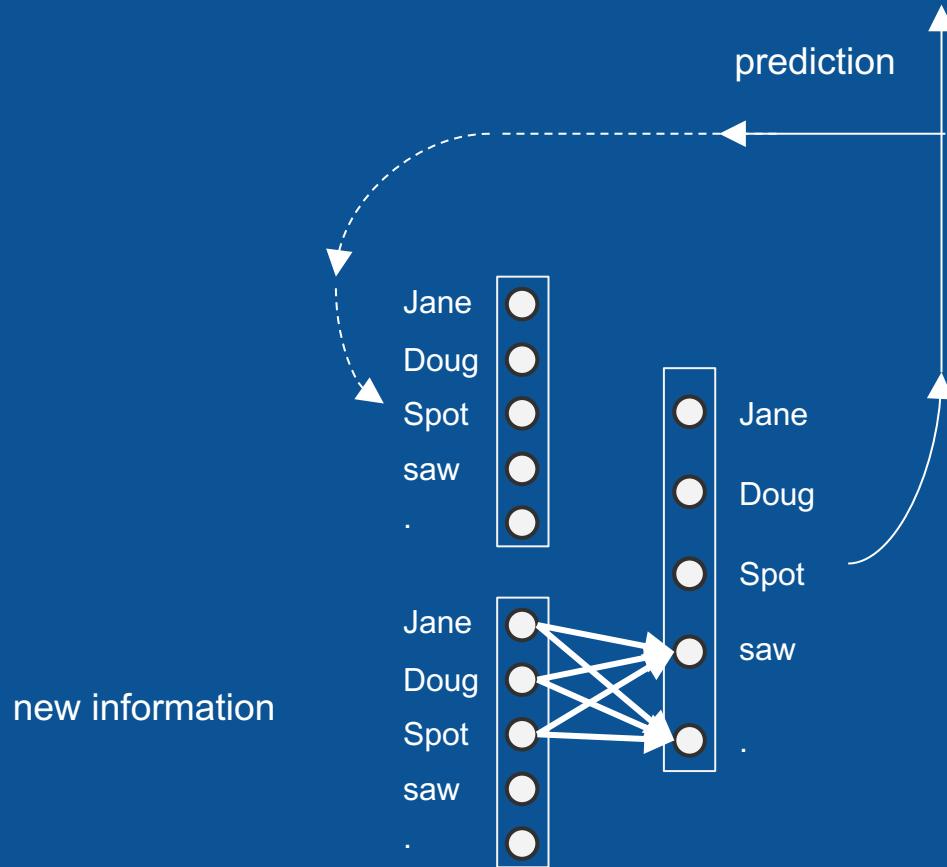
Jane saw Spot.

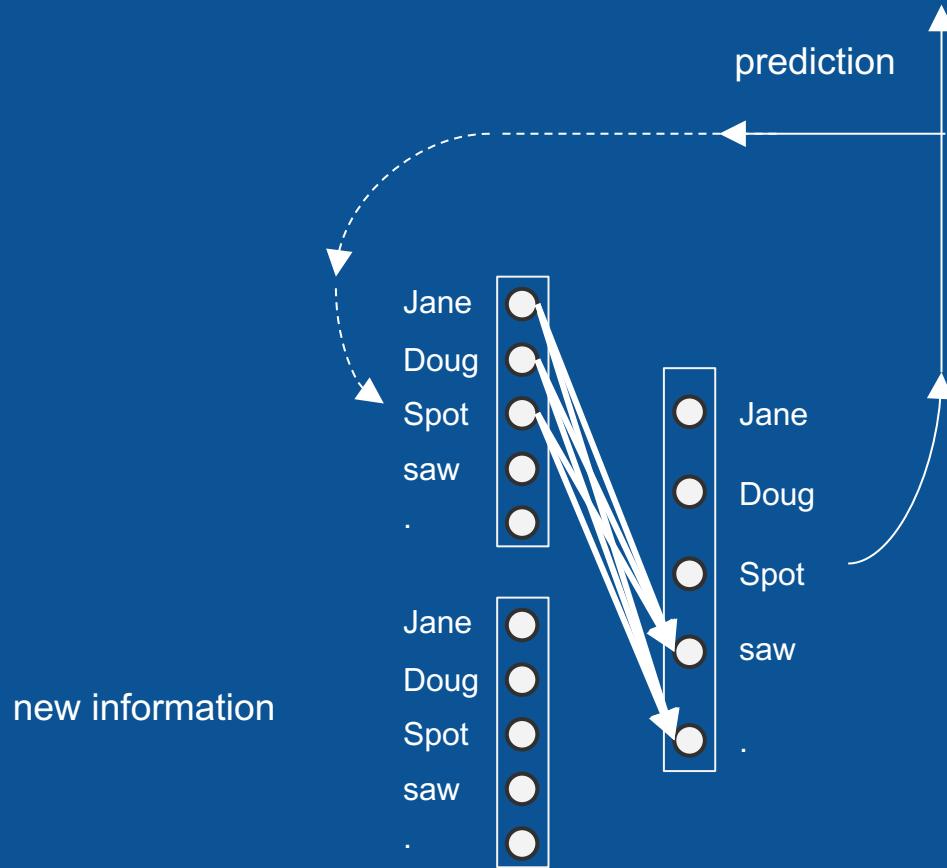
Spot saw Doug.

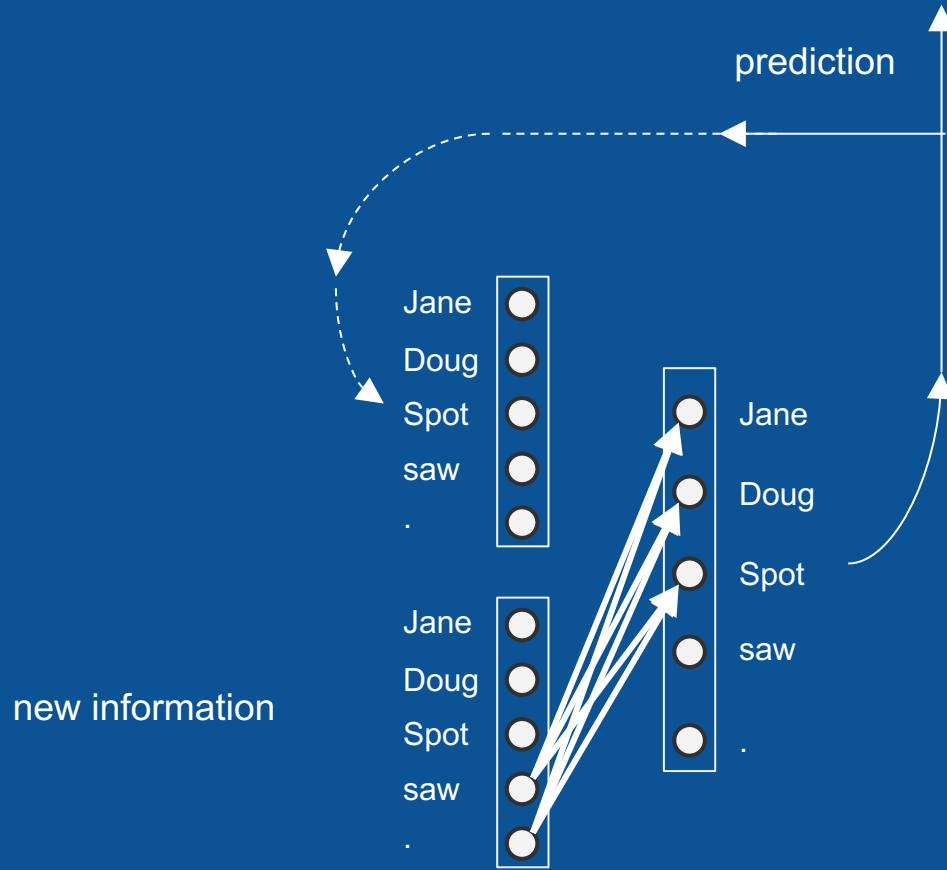
...

Your dictionary is small: {Doug, Jane, Spot, saw, ...}









RNN可能会犯的错误

Doug saw Doug.

Jane saw Spot saw Doug saw ...

Spot. Doug. Jane.

Coverage

- Introduction to NLP
- NLP components and basic methods
- Structure of RNN
- Long Short-term Memory (LSTM)
- Basic language models
- BERT
- NLP applications

LSTM (Long Short Term Memory)长短期记忆模型

- Hochreiter & Schmidhuber(1997) 解决了使用RNN记忆过去状态的问题，使得RNN能记忆大约100步时间状态



Sepp Hochreiter

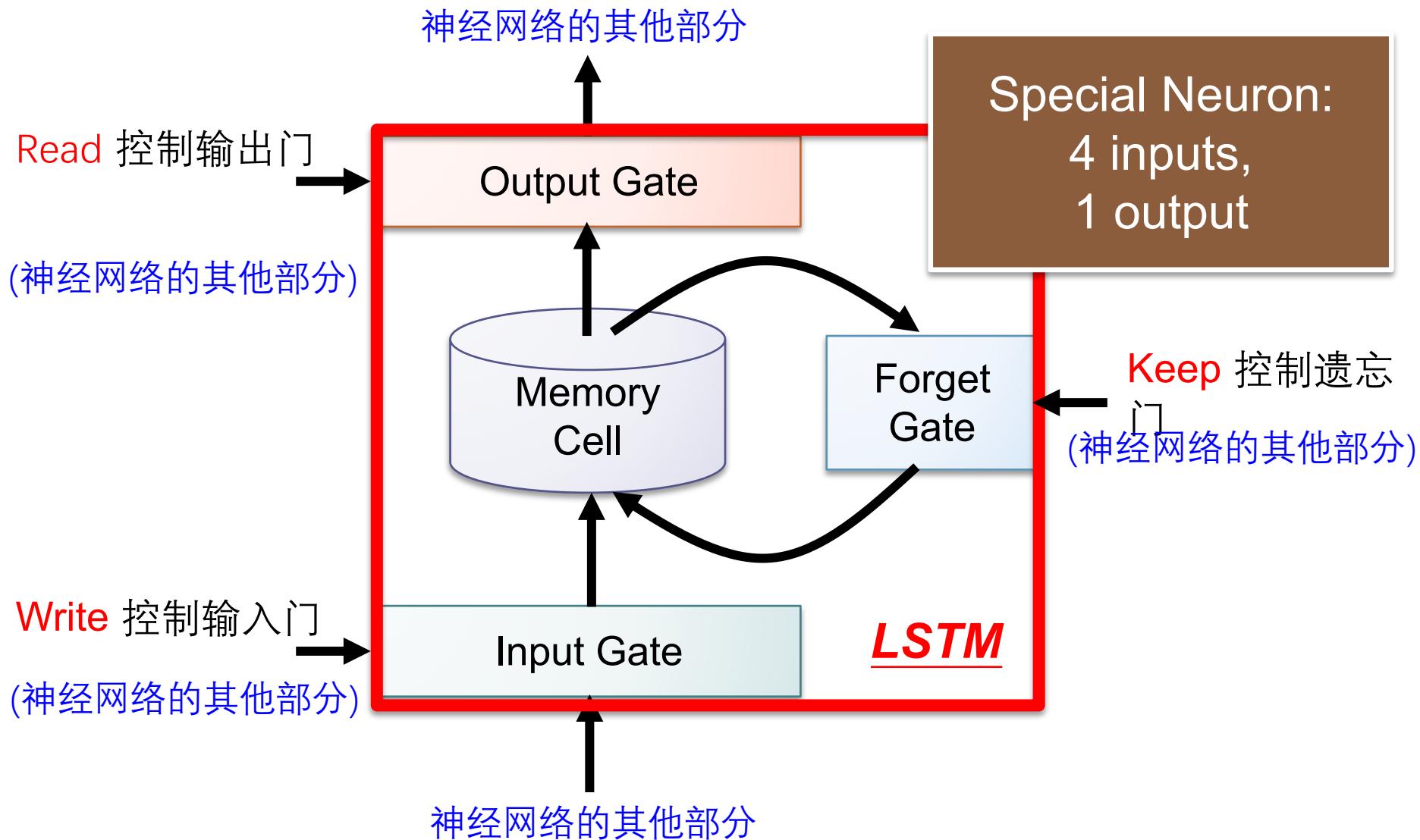


Jürgen Schmidhuber

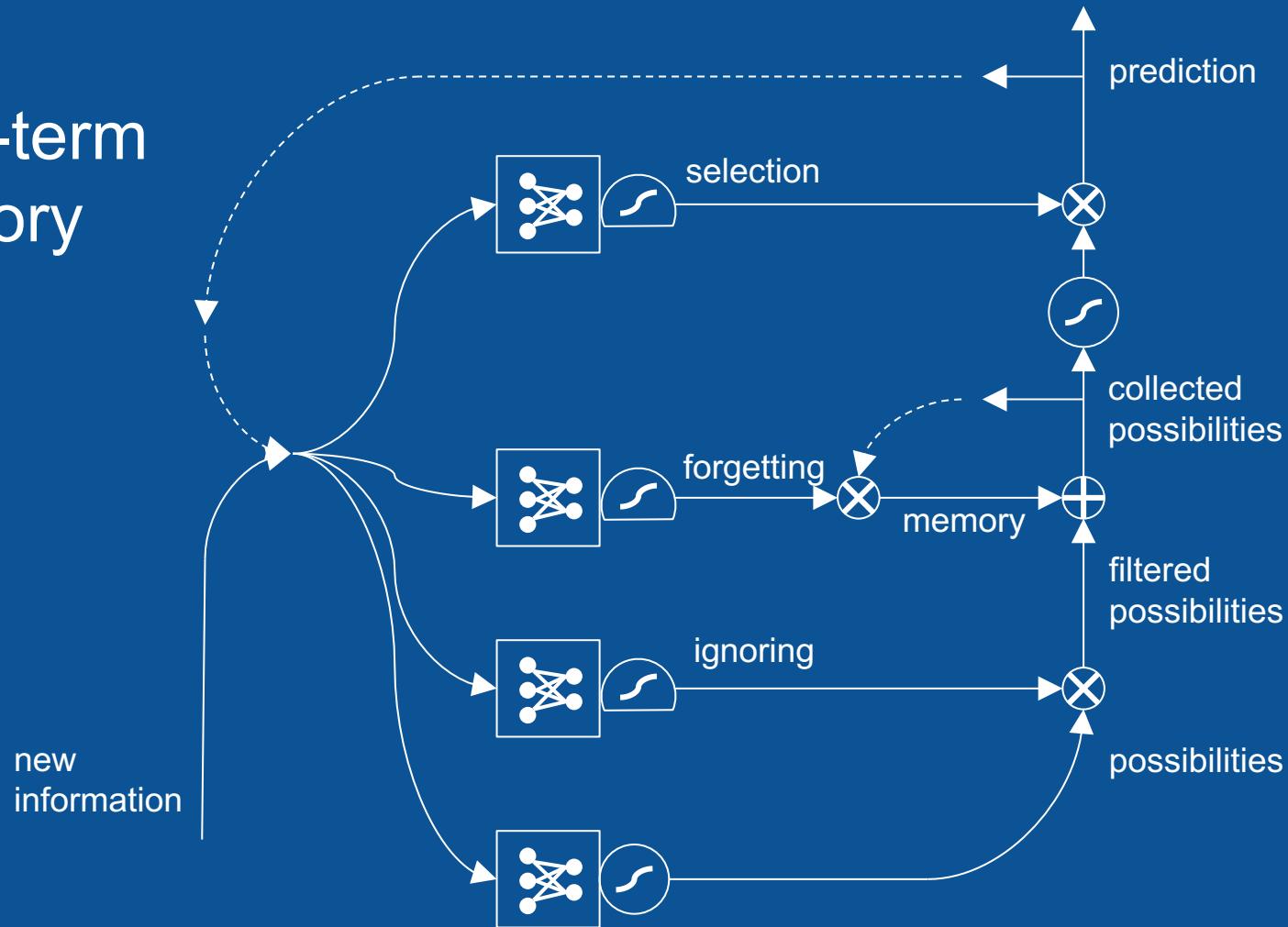
LSTM – Long Short Term Memory

- Hochreiter & Schmidhuber 使用逻辑和线性单元结合 Multiplicative Interaction设计了一个记忆单元。
- 当 “**write**” 门打开的时候，信息进入单元。
- 当 “**keep**” 状态门打开时，信息可以保留在单元很久。
- 当 “**read**” 状态门打开时，信息可以从单元中读取出去。

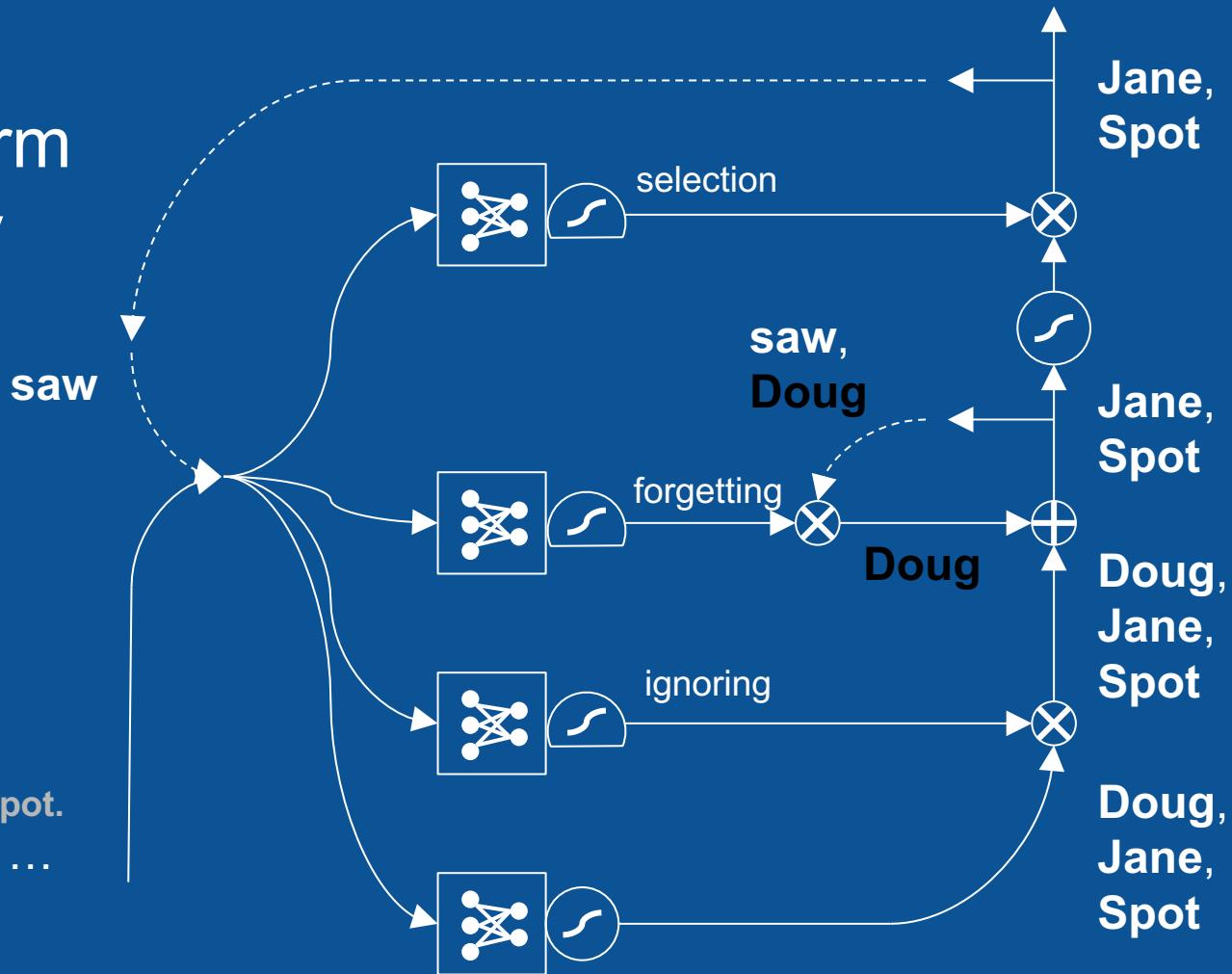
LSTM 结构



long short-term memory

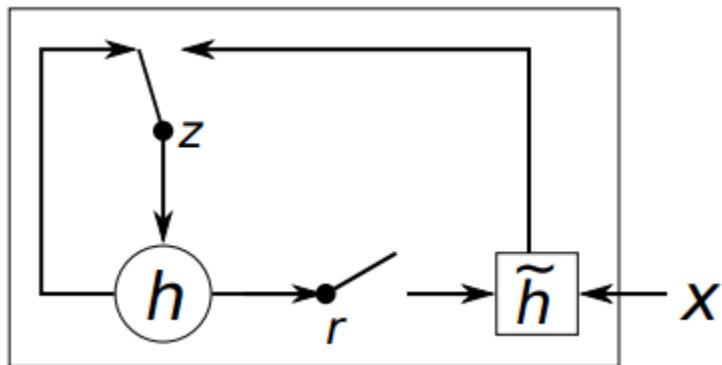


long
short-term
memory



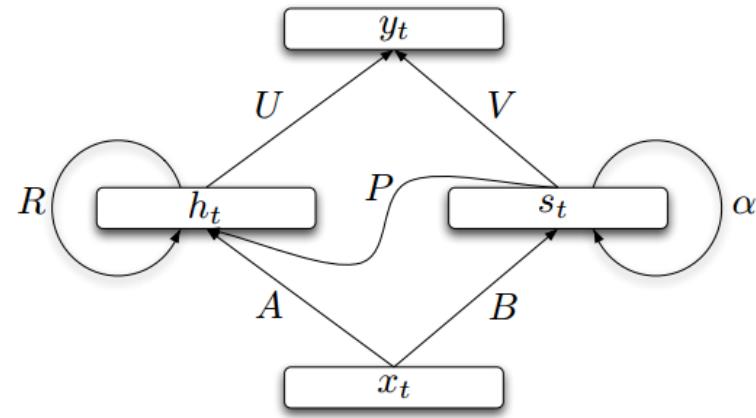
一些简化替代模型

Gated Recurrent Unit (GRU)



[Cho, EMNLP'14]

Structurally Constrained Recurrent Network (SCRN) 结构受限循环神经网络



[Tomas Mikolov, ICLR'15]

GRU 一般比较快

双向 RNN

Outputs

...

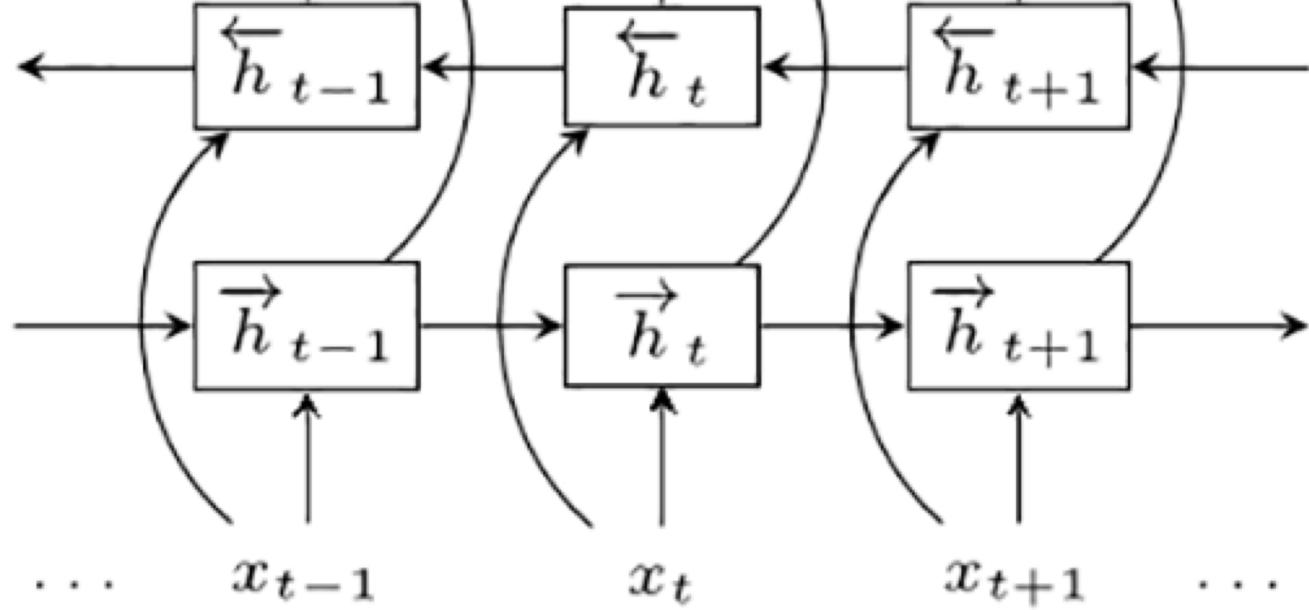
y_{t-1}

y_t

y_{t+1}

...

Backward Layer



Forward Layer

Inputs

...

x_{t-1}

x_t

x_{t+1}

...

Coverage

- Introduction to NLP
- NLP components and basic methods
- Structure of RNN
- Long Short-term Memory (LSTM)
- Basic language models
- BERT
- NLP applications

Acknowledgments

This file is for the educational purpose only. Some materials (including pictures and text) were taken from the Internet at the public domain.