第1节 从语言到信息——介绍NLP

From Languages to Information CS124

—— Lecture 1: Introduction http://120.52.51.16/web.stanford.edu/class/cs124/lec/124-2019-introns.pdf

提交作业

http://211.159.187.254:8088

从语言到信息

从以下内容自动提取句意和结构：

- 人类语言文本和演讲（新闻、社交媒体等）
- 社交网络
- 基因组序列

通过语言与人类交互

- 对话系统/聊天机器人
- 问题解答
- 推荐系统

商业世界






社交世界

- 赈灾
- 心理健康聊天机器人
- 改善警民关系
- Body-Cameras

1. 从语言中提取信息

信息检索

每天6,586,013,574次网络搜索（估算）

基于文本的信息检索很可能成为当今软件中最常用的功能

怎样工作？

文本分类：灾难响应

对短信进行分类

Mwen thomassin 32 nan pyron mwen ta renmen jwen yon ti dlo gras a dieu bo lakay mwen anfom se sel dlo nou bezwen

I am in Thomassin number 32, in the area named Pyron. I would like to have some water. Thank God we are fine, but we desperately need water.

2010 海地地震

提取情绪和社会语义

很多的意思都隐藏在**内涵**中

内涵：除单词的字面意思或主要意义之外，由其引发的想法或感觉。

通常称提取内涵为**情感分析**

餐厅评价中的情绪

以Yelp上900000条的餐厅评价为数据集

其中一条一星评论如下：

The bartender... absolutely horrible... we waited 10 min before we even got her attention... and then we had to wait 45 - FORTY FIVE! - minutes for our entrees... stalk the waitress to get the cheque... she didn't make eye contact or even break her stride to wait for a response ...

差评中惯用的语言是什么？

- 消极情绪语言

  horrible, awful, terrible, bad, disgusting…
- 涉及到人物时以过去式叙述

  waited, didn’t, was

  he, she, his, her,

  manager, customer, waitress, waiter
- 经常提及“我”和“我们”

  … we were ignored until we flagged down a waiter to get out

waitress

使用英语的另外一些叙述

- 一种常用类型

  Past tense (过去时), we/us, negative, people narratives (叙述)

- 遭受创伤的人的记录文字
  - James Pennebaker lab at UT Austin
  - Past tense as distancing (过去式表示距离感)
  - Use of "we": seeking solace in community (使用we 以寻求慰藉)

- 经常提及“我”和“我们”

  … we were ignored until we flagged down a waiter to get out waitress

一星评论属于创伤叙述！

计算生物学：比较基因序列

AGGCTATCACCTGACCTCCAGGCCGATGCCC

TAGCTATCACGACCGCGGTCGATTTGCCCGAC

```
-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
 ||  |||||||||  |||||x|  ||  |||  |||||
TAG-CTATCAC--GACCGC--GGTCGATTTGCCCGAC
```

序列比较的关键之处：

- 寻找基因
- 决定功能
- 揭示进化进程

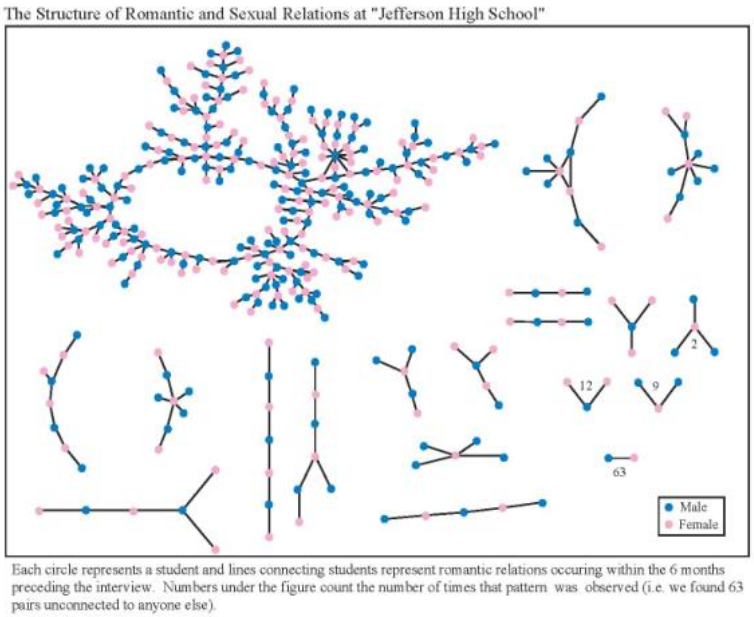
这也是拼写检查工作的方式！

Hint: 编辑距离算法

社交网络

由您的亲朋好友组成的网络离线或在线关系

- 我们可以计算这些网络的属性吗？
- 可以从这些网络中提取信息吗？
- 社交关系的结构是什么？
  - **人作为节点**
  - **链接代表关系的建立**
- 关系图的形状是什么？
  - **一个紧密相连的图？**
  - **一条线？**
  - **一个圆环？**



2. 通过语言与人类交互

私人助理

Siri

Facebook M

A personal assistant inside Messenger

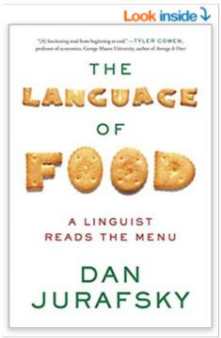
Google Now

Listening...

amazon alexa

推荐引擎

图书推荐



音乐推荐

每日歌曲推荐

根据你的口味生成，
每天6:00更新

网易云音乐

为什么语言解释难以实现?

- 歧义

  例：开刀的是他父亲。（可理解为她父亲是开刀的外科医生，也可理解为她父亲患病，医生给他做了手术。）

- 非标准语言

  Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either

- 新词

  例：排遣式进食、塑料姐妹花、unfriend

如何在这些问题上取得进展 …

需要什么工具？

- 关于语言和世界的知识
- 一种结合各种知识体系的方法

我们通常会怎么做？

- 从语言数据建立概率模型

模型与工具

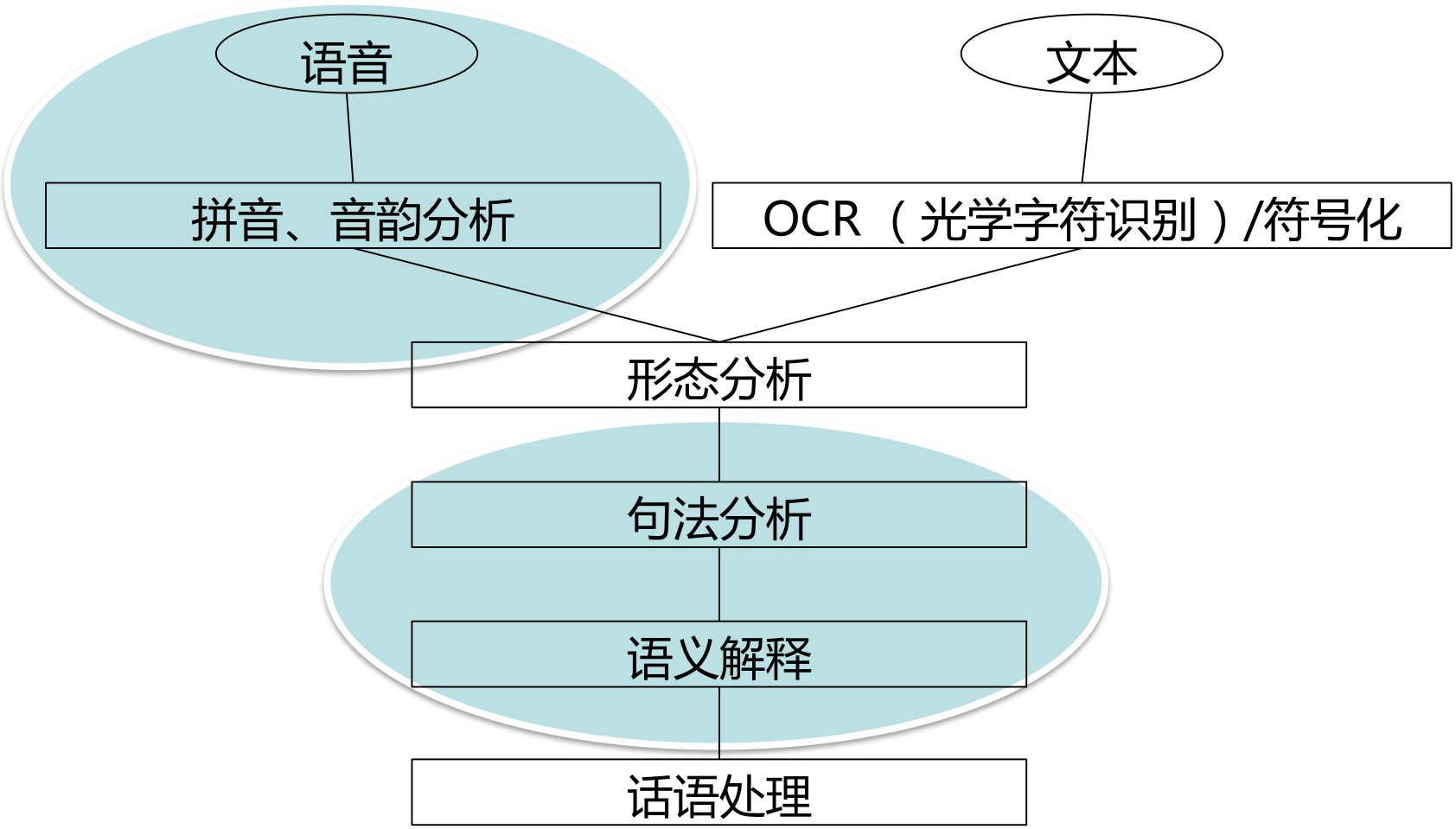
- 正则表达式 (Regular Expressions)
- 编辑距离与对齐方式
- 词嵌入（针对词义的向量/神经网络模型）
- 机器学习分类器（朴素贝叶斯/线性回归/神经网络）
- 推荐算法（协同过滤）
- 网络算法 (PageRank)
- 语言学工具（情绪词典 – Sentiment lexicons）

什么是自然语言处理 (NLP)？

- NLP是计算机科学、人工智能、语言学的交集领域。
- NLP的目标是让计算机处理或者“理解”自然语言，从而执行譬如预约、买东西、问题回答（像Siri、Google assistant、Facebook M、Cortana等一样）等有用的任务。
- 但是充分理解和表达语言的含义（甚至定义语言）是件有难度的事情。达到完美的语言理解被称为AI-complete问题。

注：在人工智能领域，最困难的问题被非正式地称为AI-complete或AI-hard。 ——维基百科

NLP层级



NLP应用 从简单到复杂

- 拼写检查、关键字搜索、同义词查找
- 从网站提取信息，如产品价格、日期、地点、人员或公司名称
- 分类：阅读教材，判断长文本的积极/消极的情绪
- 机器翻译
- 口语对话系统
- 复杂的问答系统

NLP在工业上的应用越发广泛

- 搜索（书面或口头上）
- 在线广告匹配
- 自动/辅助翻译
- 市场或财务/交易的情绪分析
- 语音识别
- 聊天机器人/对话代理：自动化客户支持、控制设备、订购货物

为什么NLP是一门困难的学科？

- 表现、学习和使用语言/情景/世界/视觉知识的复杂性
- 人类语言具有不明确性（与编程和其他正式语言不同）
- 人类语言的解释取决于现实世界、常识和语境知识

深度学习在NLP中的应用

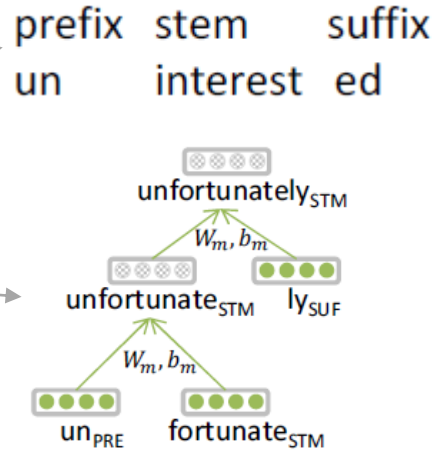
深度NLP = 深度学习 + NLP

- 结合自然语言处理的思想和目标，采用表示学习和深度学习的方法来解决这些问题。
- 近年来NLP有了很大的改进与不同（后面说明）：
  - **层次：语音，文字，语法，语义**
  - **工具：词性，实体，解析**
  - **应用：机器翻译，情感分析，对话代理，问答系统**

NLP层级的表示：形态学

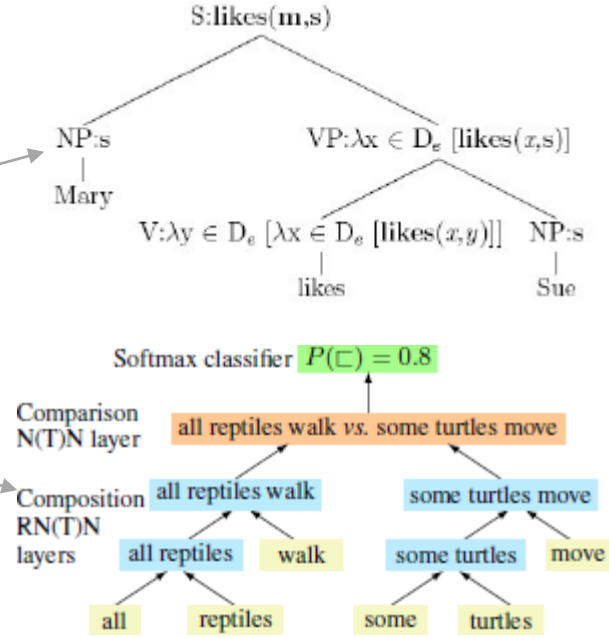
- 传统：词语是由语素组成的。
- 深度学习：每个语素都是一个向量，神经网络将两个向量组合成一个向量

注：最大的语法单位是句子，比句子小的语法单位，依次是短语、词、语素。



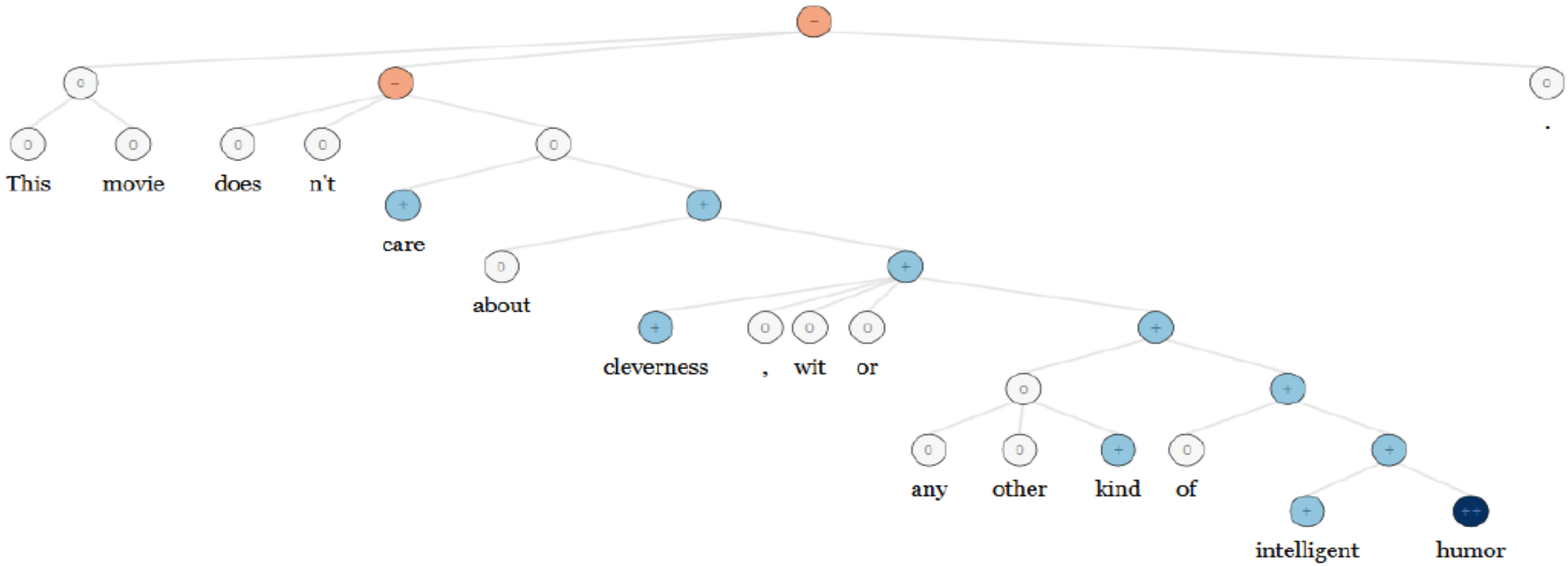
NLP层级的表示：语义

- 传统：Lambda演算：精心设计的功能、作为输入具体的其他功能、没有语言的相似性或模糊性的概念。
- 深度学习：每个单词，每个词组和每个逻辑表达式都是一个向量，神经网络将两个向量组合成一个向量。



NLP应用：情感分析

- 传统：精心策划的情感词典结合袋表示（忽略词序）或人工设计否定特征（不会捕捉所有内容）
- 深度学习：可以使用用于形态学、语法和逻辑语义学的相同的深度学习模型（递归NN ）



NLP应用：问答系统

- 传统：很多用于捕捉世界和其他知识的特征，例如正则表达式。

Is main verb trigger?

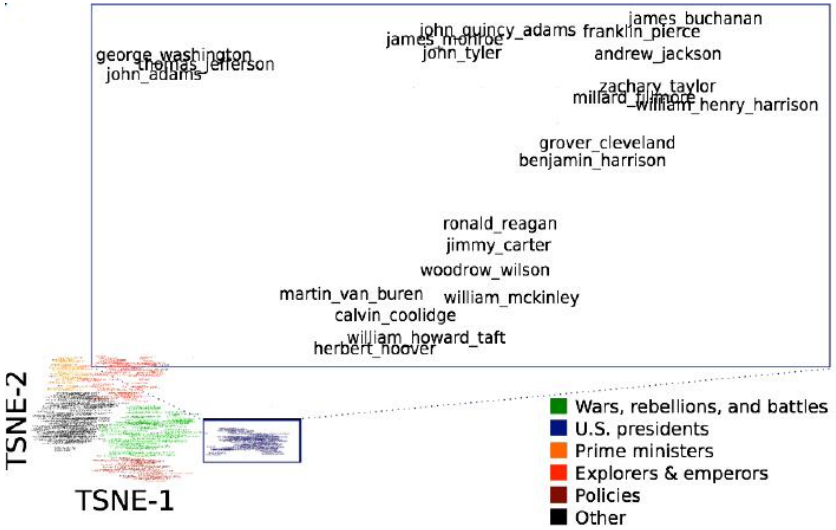
Yes

Condition	Regular Exp.
Wh- word subjective?	AGENT
Wh- word object?	THEME

No

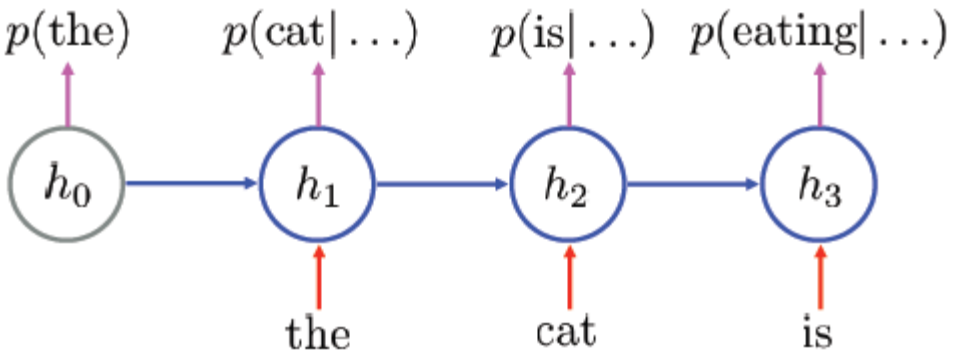
Condition	Regular Exp.
default	(ENABLE|SUPER)+
DIRECT	(ENABLE|SUPER)
PREVENT	(ENABLE|SUPER)* PREVENT(ENABLE|SUPER)*

- 深度学习：可以使用深度学习架构。事实存储在向量中。



NLP应用：对话代理/回复生成

- 一个简单而成功的例子是Google Inbox APP中提供的自动回复功能。
- 这是一个递归神经网络的实例，是强大且通用的神经语言模型的应用。



NLP应用：机器翻译

- 过去曾经尝试过很多层次的翻译：传统的MT系统是非常庞大的复杂系统。
- 神经机器翻译：源语句被映射到向量，然后输出生成的语句。现在被应用在谷歌翻译（等）的一些语言翻译上，大大降低了错

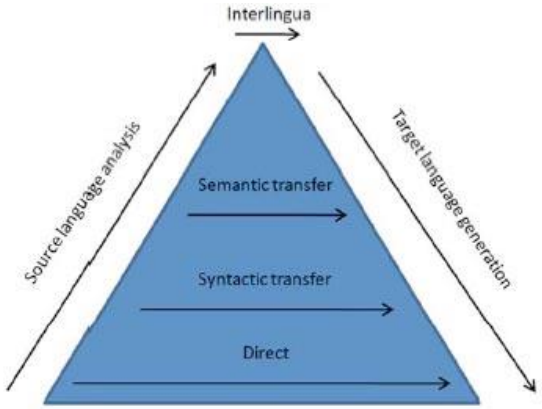
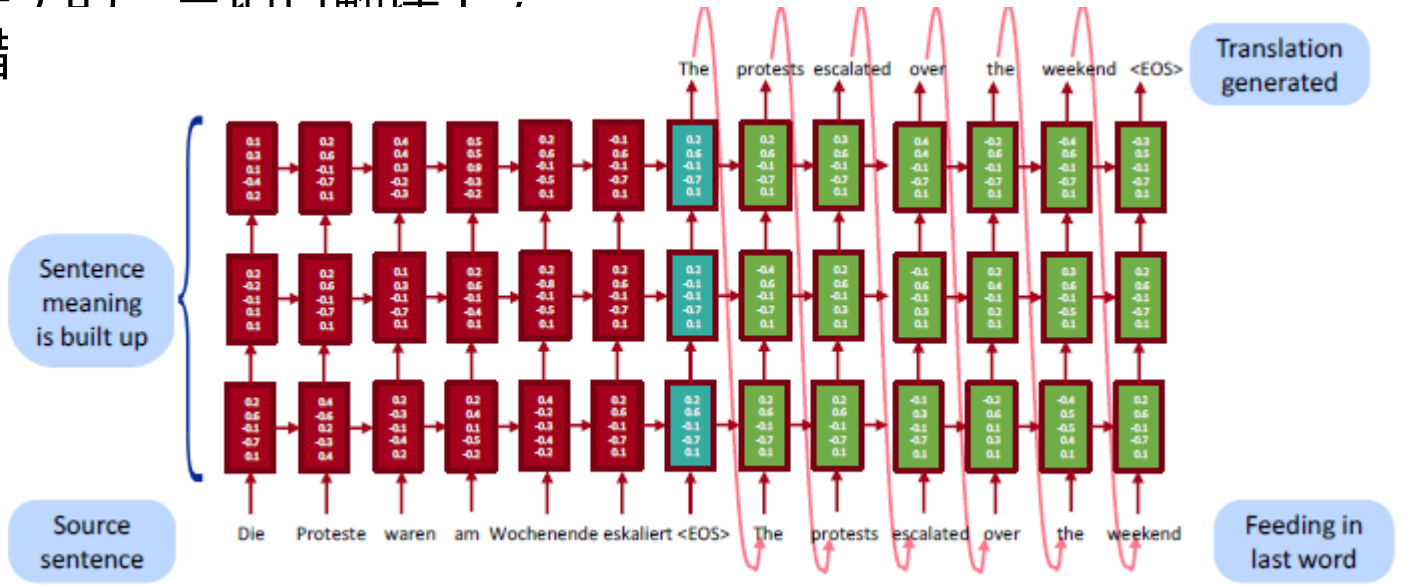
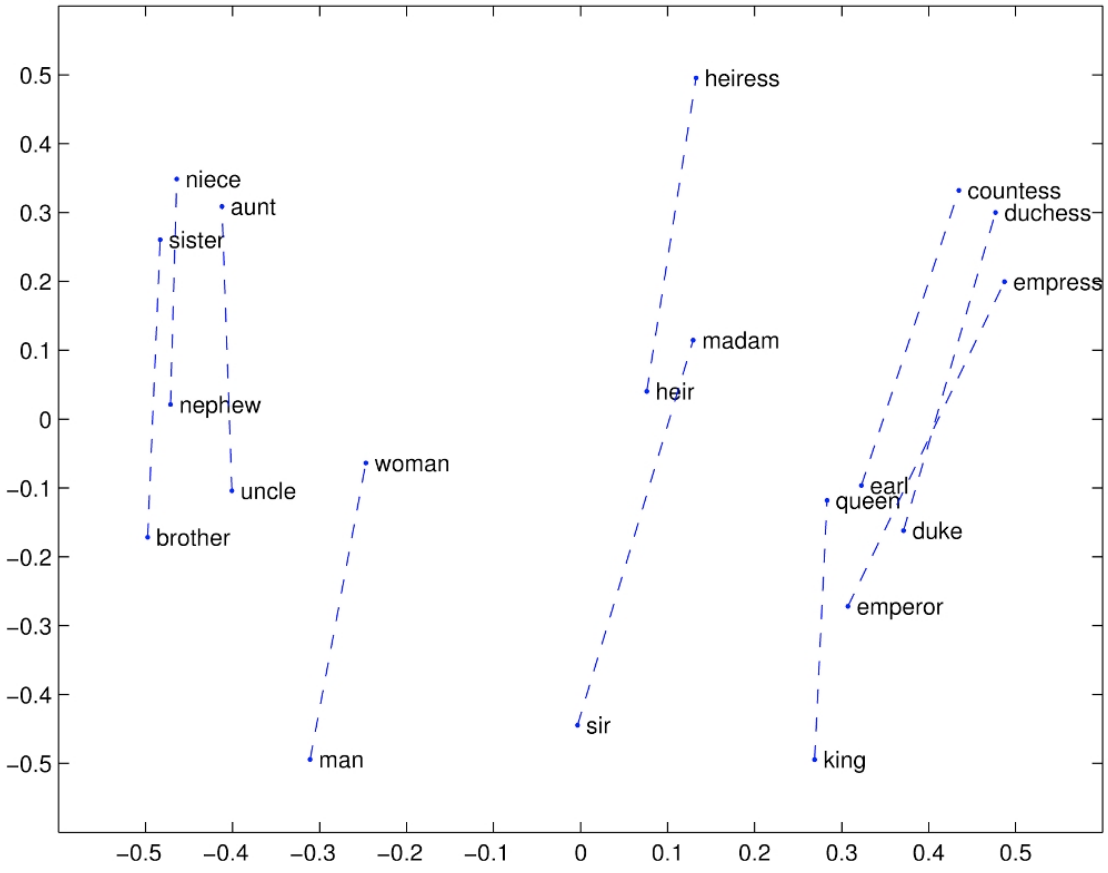


Figure 1: The Vauquois triangle



结论：在各个层级上的表示方式——向量



Thank you !