数据挖掘导论
Introduction to Data Mining

第五章 文本挖掘

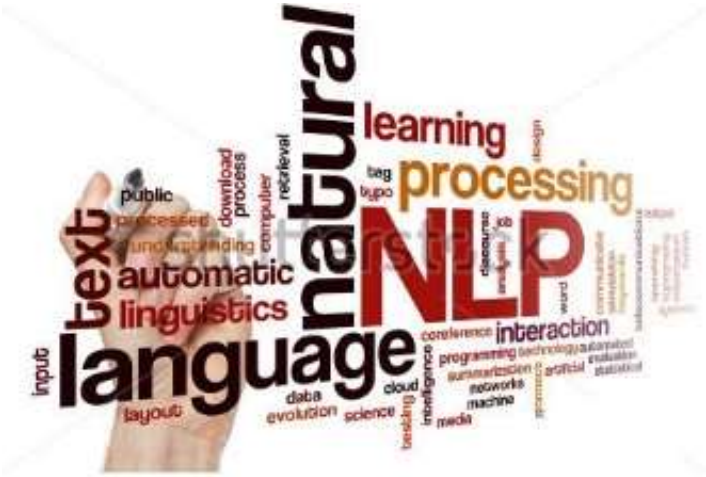
刘 淇
Email: qiliuql@ustc.edu.cn
课程主页：
http://staff.ustc.edu.cn/~qiliuql/DM2017YZ.html

目录

- 传统的文本处理
  - 自然语言处理
  - N-Gram语言模型
- 文本挖掘
  - 文本挖掘简介
  - 特征抽取
  - 特征选择
  - 文本分类
  - 文本聚类
  - 文本挖掘的常用方法
    - 主题模型
    - word2vec

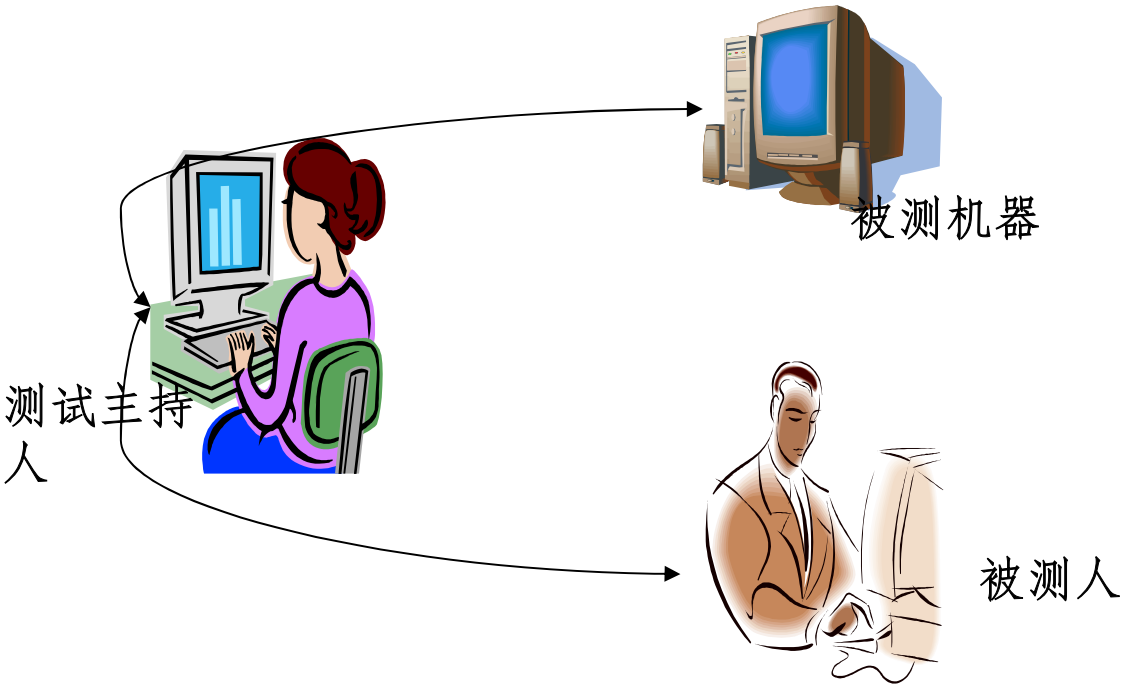
自然语言处理

- 计算机发明以来，人类首先想到的计算机的应用之一，就是利用机器进行交流，如翻译。然而时至今日，计算机处理自然语言的能力在很多情况下仍不能满足人类信息化时代的需求。
- 因此，对于文本数据，打破不同语言之间的固有壁垒，对其有效的理解，是个既有趣又有挑战性的研究。



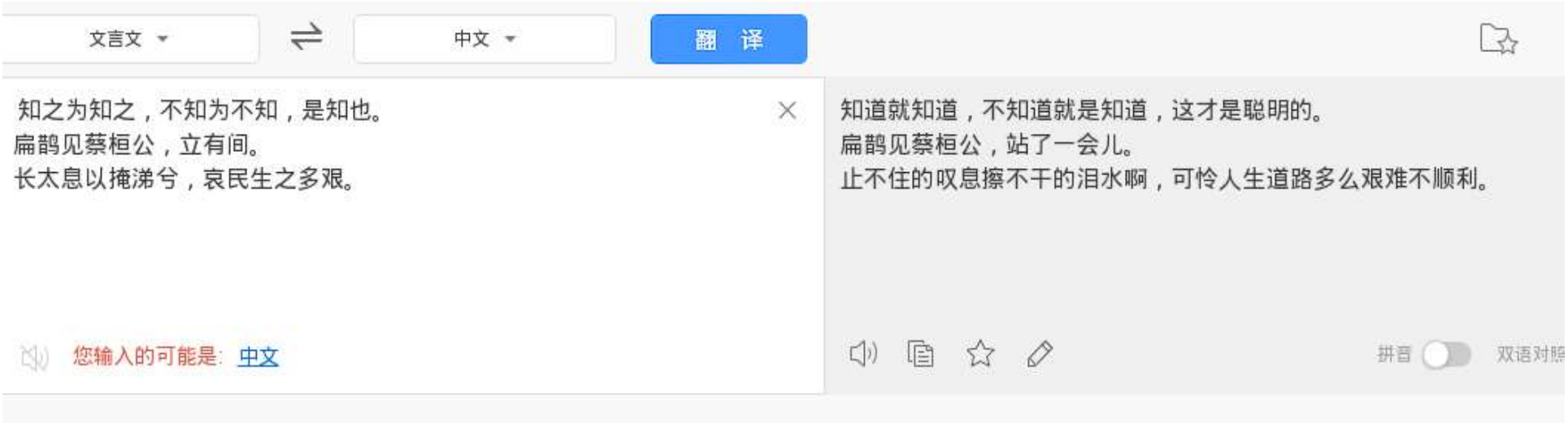
自然语言处理

- Turing Test图灵测试
  - 进行多次测试后，如果有超过30%的测试者不能确定出被测试者是人还是机器，那么这台机器就通过了测试，并被认为具有人类智能。



自然语言处理

- 自然语言处理研究的内容：
- 机器翻译（Machine translation, MT）：实现一种语言到另一种语言的自动翻译。
  - 应用：文献翻译、网页搜索和辅助浏览等。



自然语言处理

- 自然语言处理研究的内容：
- 自动摘要（Automatic summarization）：将原文档的主要内容或某方面的信息自动提取出来，并形成原文档的摘要或缩写。
  - 应用：电子图书管理、情报获取、游记生成等。

A Neural Attention Model for Abstractive Sentence Summarization

Alexander M. Rush
Facebook AI Research /
Harvard SEAS
srush@seas.harvard.edu

Sumit Chopra
Facebook AI Research
spchopra@fb.com

Jason Weston
Facebook AI Research
jase@fb.com

Abstract

Summarization based on text extraction is inherently limited, but generation-style abstractive methods have proven challenging to build. In this work, we propose a fully data-driven approach to abstractive sentence summarization. Our method utilizes a local attention-based model that generates each word of the summary conditioned on the input sentence. While the model is structurally simple, it can easily be trained end-to-end and scales to a large amount of training data. The model shows significant performance gains on the DUC-2004 shared task compared with several strong baselines.

自然语言处理

- 自动摘要（使用Large-scale Chinese Short Text Summarization dataset （LCSTS））：
- SOURCE:
  - 当前 西方 世界 最 畅销 的 书， 是 法国 经济学家 托马斯·匹克迪 的 新作 《21 世纪 资本论》。 英国 《经济学人》 网站 刊文 指出， 这 本书 大受欢迎 之 处 在于， 它 指出 持续增长 的 财富 集中化 是 资本主义 固有 现象， 同时 呼吁 向 全球 富人 收税， 以此 作为 改善 现状 的 方案。
- Result:
  - 托马斯 新作 《21 世纪 资本论》 呼吁 向 全球 富人 收税
- Target:
  - 当代 资本论 缘何 成 西方 最 畅销书

自然语言处理

- 自动摘要（使用Large-scale Chinese Short Text Summarization dataset （LCSTS））：
- 有一些keywords 抓不住
- SOURCE:
  - 1 . 陕西 汉中 ； 2 . 湖北 荆门 ； 3 . 江苏 兴化 ； 4 . 浙江 瑞安 ； 5 . 云南 罗平 ； 6 . 重庆 潼南 ； 7 . 青海 门源 ； 8 . 上海 奉贤 ； 9 . 江西 婺源 ； 10 . 贵州 贵定 。 漫步 于 乡村 的 石板路 ， 穿过 溪河 的 石拱桥 ， 这个 周末 去 油菜花 花海 寻找 春天 吧
- Result:
  - 中国 最美 的 十大 经典 地方
- Target:
- 中国 十大 最美 油菜花 海

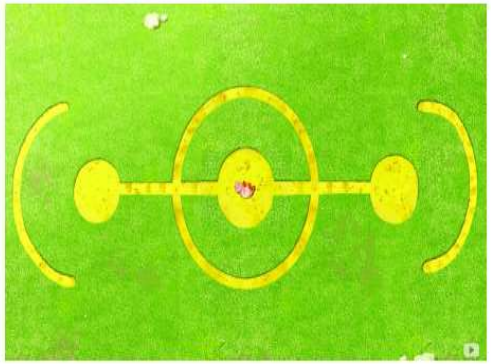
自然语言处理

- 自动摘要（使用Large-scale Chinese Short Text Summarization dataset （LCSTS））：
- 输出的重点和target不相符
- SOURCE:
  - 日前， 有 网友 微博 爆料： 安康 汉滨区 许家 台 古墓 遗址 前 的 石马 、 武士 俑 等 雕像 风化 严重 。 记者 随后 从 汉滨区 文物部门 获悉 ， 当地 正在 准备 对 古墓 前 的 文物 搬迁 保护 。 此外， 根据 墓碑 记载 ， 南宋 名将 王彦 很 可能 是 绥德 人 。
- Result:
  - 网 曝 安康 汉滨区 许家 台 古墓 遗址 雕像 风化 严重
- Target:
  - 安康 汉滨区 发现 珍贵 墓碑 记载 南宋 名将 王彦 或 为 绥德 人

自然语言处理

- 自动评论、回复

p：除草机- -
p：Weeder - -



h：神吃啊=.=!!
h：Eating very skillfully=.=!!

微博：八个英格兰球员都待在禁区内

回复：这真是一场精彩的比赛

回复：比赛还是0：0，没有进球

$$\begin{aligned} \arg\min_{L_x,L_y} \quad & \max(1-\sum_i \mathbf{x}_i^\top L_x L_y^\top \mathbf{y}_i, 0) \\ s.t. \quad & \|L_{n,x}\|_1 \le \mu_1,\ n=1,2,\cdots,N_x \\ & \|L_{m,y}\|_1 \le \mu_1,\ m=1,2,\cdots,N_y \\ & \|L_{n,x}\|_2 = \mu_2,\ n=1,2,\cdots,N_x \\ & \|L_{m,y}\|_2 = \mu_2\ m=1,2,\cdots,N_y. \end{aligned}$$

Hao Wang, Zhengdong Lu, Hang Li and Enhong Chen, A Dataset for Research on Short-Text Conversation, EMNLP'2013.

自然语言处理

- 能够进行多轮语言沟通的智

偶然心儿跳动的时候我看到了你
舞台上不必张惶
眼睛充满生命的火焰
幻化成水滴飘在空中
——少女诗人小冰

长按二

自然语言处理

- 2017.11.1袁晶博士与Satya在tech summit上的keynote给大家介绍小冰在AI Creation 方向的一些尝试

自然语言处理

- 自动写稿

△《四川阿坝州九寨沟县发生7.0级地震》速报，内容包括速报参数、震中地形、热力人口、周边

机器人写稿时代来了！今日头条、腾讯、南周齐发力，媒体人将迎下岗潮？

2017-07-19 10:58 腾讯 / 机器人 / 媒体

自然语言处理

- 自然语言处理研究的内容：
- 信息检索（Information retrieval）：利用计算机系统从大量文档中找到符合用户需求的相关信息。
  - 代表系统：google、百度等。
- 文档分类（Doucument categorization）：利用计算机系统对大量文档按照一定的标准实现分类。
  - 应用：图书管理、内容管理、信息安全监控。

自然语言处理

- 自然语言处理研究的内容：
- 信息检索（Information retrieval）：利用计算机系统从大量文档中找到符合用户需求的相关信息。
  - 代表系统：google、百度等。
  - 序列模式挖掘

SID	Search Session
1	丰田➔雷诺➔宝马➔奔驰
2	宝马➔奔驰➔法拉利
3	本田➔丰田➔通用➔宝马
4	吉利➔奇瑞➔长城➔江淮
5	比亚迪➔吉利➔江淮➔长安
6	长城➔江淮➔华泰➔长安

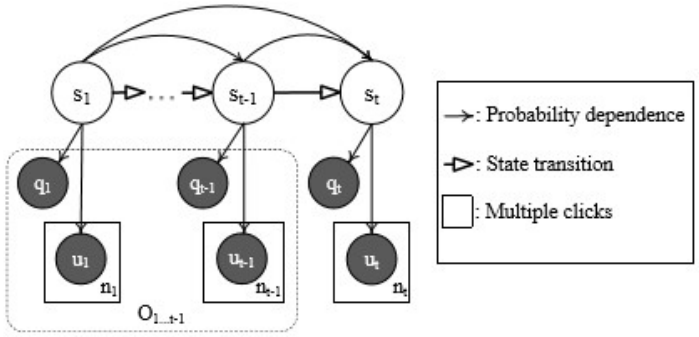
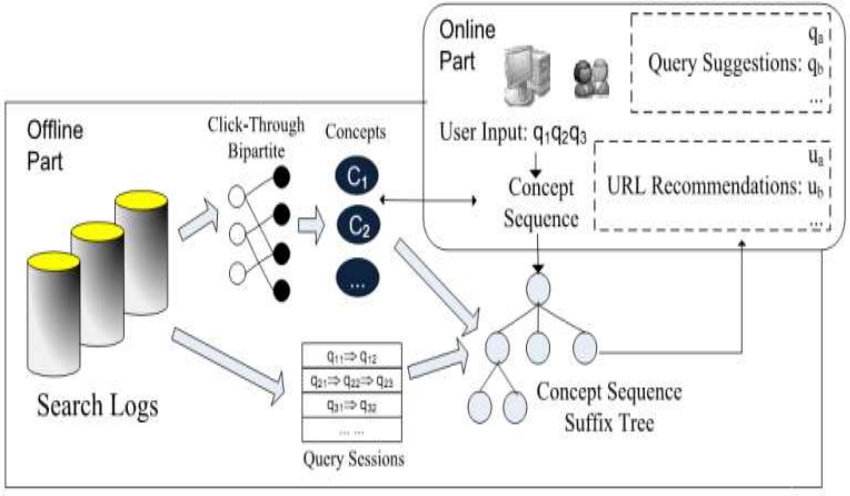


Figure 1: Graphical structure of the vlHMM.

- **Huanhuan Cao, Daxin Jiang, Jian Pei, Enhong Chen, Hang Li, Towards Context-Aware Search by Learning A Very Large Variable Length Hidden Markov Model from Search Logs, *WWW'2009*.**

自然语言处理

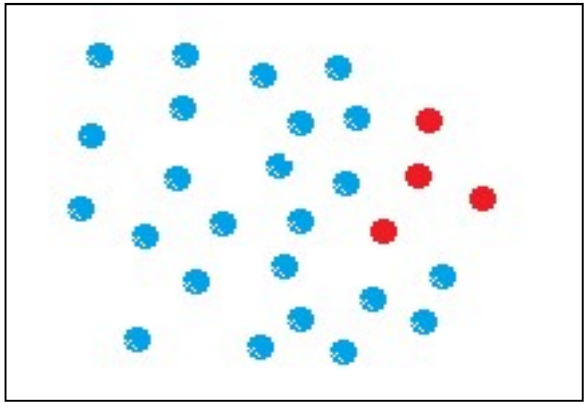
- 自然语言处理研究的内容：
- 信息检索（Information retrieval）：利用计算机系统从大量文档中找到符合用户需求的相关信息。
  - 代表系统：google、百度等。
  - 序列模式挖掘



- **Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen liao, Enhong Chen, Hang Li, Context-Aware Query Suggestion by Mining Click-Through and Session Data, *KDD'2008*.(Best Application Paper Award).**

自然语言处理

- 自然语言处理研究的内容：
- 文档分类（Doucument categorization）：利用计算机系统对大量文档按照一定的标准实现分类。
  - 应用：图书管理、内容管理、信息安全监控。
  - 常以词向量为特征



WIKIPEDIA
The Free Encyclopedia

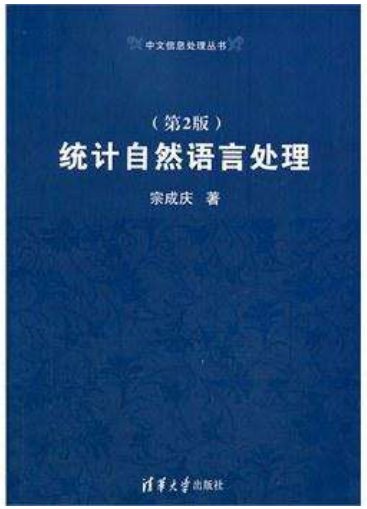
Table 1
Three short text documents.

ID	Document content
Doc 1	Puma, an American Feline Resembling a Lion.
Doc 2	Puma, a Famous Sports Brand from German.
Doc 3	Welcome to Zoo, an Animal World.

[1] Enhong Chen, Yanggang Lin, Hui Xiong, et al., Exploiting Probabilistic Topic Models to Improve Text Categorization under Class Imbalance, IPM, 2011.
[2] Zongda Wu, Hui Zhu, Guiling Li, et al., An efficient Wikipedia semantic matching approach to text document classification, Information Sciences, 2017

自然语言处理

- 自然语言处理的关键技术：
  - 词法分析
  - 句法分析
  - 语义分析
  - 语用分析
  - 语句分析



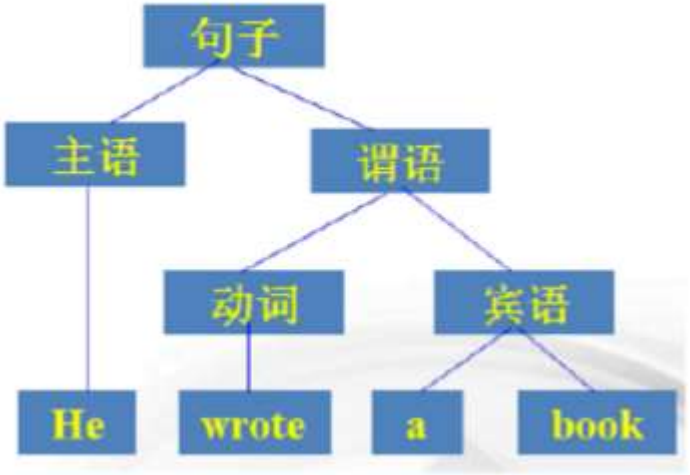
- 《统计自然语言处理》，宗成庆，清华大学出版社
- 《Foundations of Statistical Natural Language Processing》，Christopher D. Manning / Hinrich Schütze， The MIT Press

自然语言处理

- 词法分析
  - 词法分析的主要目的是从句子中切分出单词，找出词汇的各个词素，并确定其词义。
  - 不同的语言对词法分析有不同的要求，例如英语和汉语就有较大的差距。汉语中的每个字就是一个词素，所以要找出各个词素是相当容易的，但要切分出各个词就非常难。
  - “我们研究所有东西”，
  - “我们——研究所——有——东西”
  - “我们——研究——所有——东西” 。
  - 英语容易切分一个单词，但有词性、数、时态、派生、变形等变化，因而要找出各个词素就复杂得多，需要对词尾和词头进行分析。如uncomfortable可以是un-comfort-able或uncomfort-able，因为un、comfort、able都是词素。

自然语言处理

- 句法分析
- 目的是识别句子的句法结构，实现自动句法分析过程。
- 识别句子的句法结构，实现自动句法分析过程。一个句子是由各种不同的句子成分组成的。这些成分可以是单词、词组或从句。句子成分还可以按其作用分为主语、谓语、宾语、宾语补语、定语、状语、表语等。这种关系可用一棵树来表示，如对句子：He wrote a book.



自然语言处理

- 语义分析
  - 语义分析是基于自然语言语义信息的一种分析方法，其不仅仅是词法分析和句法分析这样语法水平上的分析，而是涉及到了单词、词组、句子、段落所包含的意义。
- 语用分析
  - 语用分析相对于语义分析又增加了对上下文、语言背景、环境等的分析
  - He wrote a book. “He”指的是上文中的John。
- 语境分析
  - 将自然语言与客观的物理世界和主观的心理世界联系起来，补充完善了词法、语义、语用分析的不足。

目录

- 传统的文本处理
  - 自然语言处理
  - N-Gram语言模型
- 文本挖掘
  - 文本挖掘简介
  - 特征抽取
  - 特征选择
  - 文本分类
  - 文本聚类
  - 文本挖掘的常用方法
    - 主题模型
    - word2vec

N-Gram语言模型

- N-Gram（N元模型）是自然语言处理中一个非常重要的概念，通常在NLP中，人们基于一定的语料库，可以利用N-Gram来预计或者评估一个句子是否合理。另外一方面，N-Gram的另外一个作用是用来评估两个字符串之间的差异程度。这是模糊匹配中常用的一种手段。
- 香农游戏（Shannon Game）
  - 给定前n-1个词，预测下一个词
  - 例如，给定一个词“姚明”，那么下一个词更容易想到的是“篮球”，而不是“足球”

N-Gram语言模型

- N-Gram是大词汇连续文本识别中常用的一种语言模型，对中文而言，我们称之为汉语语言模型(CLM, Chinese Language Model)。
- 该模型基于这样一种假设，第N个词的出现只与前面N-1个词相关，而与其它任何词都不相关，整句的概率就是各个词出现概率的乘积。这些概率可以通过直接从语料中统计N个词同时出现的次数得到。

$$P(T)=P(S)=P(w_1w_2...w_n)=p(w_1)p(w_2\mid w_1)p(w_3\mid w_1w_2)...p(w_n\mid w_1w_2...w_{n-1})$$

N-Gram语言模型

- 这种假设参数空间过大，且数据稀疏，不实用
  - “马尔科夫假设”
  - 下一个词的出现仅仅依赖于它前面的一个词或者几个词
- 假设下一个词的出现依赖于它前面的一个词

$$P(I)=P(S)=P(w_1w_2...w_n)=p(w_1)p(w_2|w_1)p(w_3|w_1w_2)...p(w_n|w_1w_2...w_{n-1})$$

$$\approx p(w_1)p(w_2|w_1)p(w_3|w_2)...p(w_n|w_{n-1}) \quad \text{:bigram}$$

$$\approx p(w_1)p(w_2|w_1)p(w_3|w_1w_2)...p(w_n|w_{n-2}w_{n-1}) \quad \text{:trigram}$$

N-Gram语言模型

- 最大似然估计

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

- N-Gram：N个词构成的序列
  - Unigram
  - Biggram
  - Trigram
  - …

N-Gram语言模型

- N元文法对下一个单词的条件概率逼近的通用公式为：

$$P(w_n \mid w_1^{n-1}) \approx P(w_n \mid w_{n-N+1}^{n-1})$$

- 训练N-Gram语言模型：在训练语料库中统计获得N-Gram的频度信息

N-Gram语言模型

- 假设语料库总词数为13748词

I	3437
want	1215
to	3256
eat	938
Chinese	213
food	1506
lunch	459

N-Gram语言模型

	I	want	to	eat	Chinese	food	lunch
I	8	1087	0	13	0	0	0
want	3	0	786	0	6	8	6
to	3	0	10	860	3	0	12
eat	0	0	2	0	19	2	52
Chinese	2	0	0	0	0	120	1
food	19	0	17	0	0	0	0
lunch	4	0	0	0	0	1	0

N-Gram语言模型

- P(I want to eat Chinese food)

=P(I)*P(want|I)*P(to|want)*P(eat|to)*P(Chinese|eat)*P(food|Chinese)

=0.25*1087/3437*786/1215*860/3256*19/938*120/213

= 0.000154171

- P(I want to eat Chinese food lunch)=?

N-Gram语言模型

- N的选择
- 更大的N：对下一个词出现的约束性信息更多，更大的辨别力
- 更小的N：在训练语料库中出现的次数更多，更可靠的统计结果，更高的可靠性

- “我 正在 _____”
  - 讲课？图书馆？听课？借书？学习？……
- “我 正在 图书馆 _____”
  - 借书？学习？……

N-Gram语言模型

- 词表中词的个数|V|=20,000

n	所有可能的n-gram的个数
2 (bigrams)	400,000,000
3 (trigrams)	8,000,000,000,000
4 (4-grams)	1.6×10^{17}

N-Gram语言模型

- 数据稀疏问题
  - Good-Turing平滑
  - Back-off平滑
  - 线性插值平滑
  - Witten-Bell平滑
  - 等等

N-Gram语言模型的应用

- 统计所有bigram的频度，作为文本分类的特征
  - 多字词，专业研究领域影响很大（氢氧化钠）
  - 在进行bi-gram切分时，不仅统计gram的出现频度，而且还统计某个gram与其前邻gram的情况，并将其记录在gram关联矩阵中。对于那些连续出现频率大于事先设定阈值的，就将其合并成为多字特征词。
- 新闻报道（记者，采访…）
- 军事文本（AK47，轰炸机…）
- 天气预报（晴转多云，降水量…）