Applications of NLP

* Machine Translation

Google Translate, Amazon Translate

* Speech Recognition

Mobile telephony, home automation, hands-free computation, virtual assistance, video gams…

* Question Answering/Chatbots

Siri, OK Google, Virtual Assistants

* Automatic Summarization

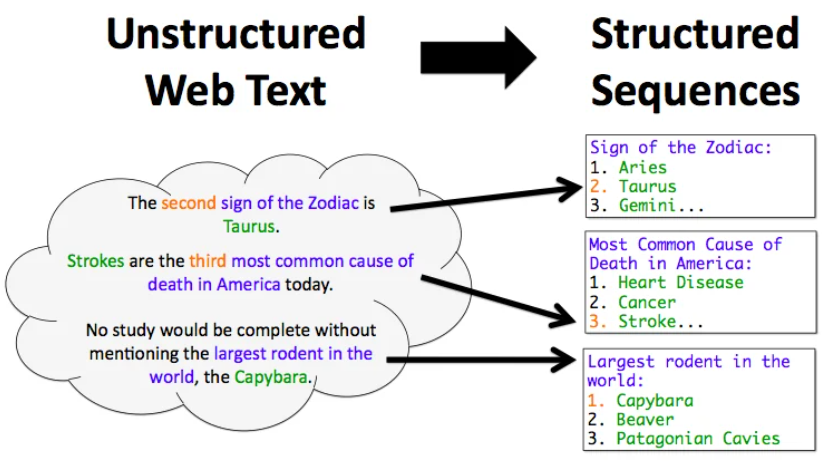
Extraction-based summarization: copying and rearranging passages from the original text

Abstraction-based summarization: generate new phrases, rephrasing or using words that were not in the original text

* Optical Character Recognition (OCR)
* Sentiment Analysis

Social Media Monitoring, customer reaction to product

* Span Filtering (N-gram Modeling, Word Stemming, Bayesian Classification)
* Information Extraction



* Search Auto-Correct and Auto-Complete

Business area:

Customer Service

Chatbots provide swift help in personalized conversations and partially replacing humans in simple scenarios

Hiring and Recruitment (Application Tracking Systems)

Targeting Advertising

Healthcare

Grammar Checkers (Grammarly )

自然语言处理

概率: 从随机试验中的事件到实数域的映射函数，用以表示事件发生的可能性.

估计概率值：相同的情况下重复试验N次，当N足够大时，用相对频率近似概率值

条件概率：如果A,B是样本空间的两个事件，在已知事件B发生的情况下，事件A的概率。

条件概率计算：（贝叶期法则）

随机变量：一个随机试验可能有多种不同的结果，到底会出现哪一种，存在一定的概率。简单地说，就是试验结果的函数 (也称为随机变量的概率分布)

在自然语言处理中，一般以句子为处理单位，为了简化问题的复杂性，通常假设一个句子的出现独立于它前面的其他语句，句子的概率近似地被认为符合二项式分布。

联合概率分布和条件概率分布

贝叶斯决策理论

信息论

信息熵：信息量的量化，即描述一个随机变量的不确定性的数量。物理意义：表示信源X每发一个符号（不论发什么符号）所提供的平均信息量。熵越大，不确定性越大（即信息量越大），那么正确估计其值的可能性越小。

在已知部分知识的前提下，关于未知分布最合理的推断应该是符合已知知识最不确定或最大随机的推断.

最大熵的应用：选择模型

自然语言处理中，通常的做法是，根据已知的样本设计特征函数，假设存在k个特征函数

，它们都在建模过程中对输出有影响，那么，所建立的模型应满足所有这些特征的约束，即所建立的模型p应该属于这k个特征函数约束下所产生的所有模型的集合C.使熵H(p)值最大的模型用来推断某种语言现象存在的可能性，或者作为进行某种处理操作的可靠性依据。

联合熵：描述一对随机变量平均所需要的信息量。

长度为n信息，每一个字符的熵:

互信息：反映的是在知道Y的值以后，X的不确定性的减少量。可以理解为Y的值透露了多少关于X的信息量。互信息体现了两变量之间的依赖程度：如果I(X,Y)>>0,表示X与Y是高度相关的话，如果I(X,Y)=0，表明X和Y是相互独立的.

相对熵:也称KL距离，是衡量相同事件空间里两个概率分布相对差距的测度.

交叉熵：用来衡量估计模型与真实概率分布之间差异情况的。等于真实概率分布的熵+估计模型概率分布和真实概率分布之间的相对熵。在设计模型时，目的是使交叉熵最小，从而使模型最接近真实的概率分布. 设计语言模型时，一般用困惑度来代替交叉熵衡量语言模型的好坏。困惑度是交叉熵的指数函数.

噪声信道模型

信息熵可以定量地估计信息源每发送一个符号所提供的平均信息量。

一般情况下，在信号传输的过程中都要进行双重性处理：一方面要对编码进行压缩，尽量消除所有的冗余； 另一方面又要通过增加一定的可控冗余以保障输入信号经地噪声信道传输以后可以很好地恢复原状。

在自然语言处理中，不需要进行编码，一种自然语言的句子可以视为已编码的符号序列，但需要进行解码，使观察到的输出序列更接近于输入。

模拟信道模型，在自然语言处理中，很多问题都可以归结为在给定输出O（可能含有误传信息）的情况下，如何从所有可能的输入I中求解最有可能的那个，即求出使p(I|O)最大的I所有为输入I。用贝叶斯公式求解：

举例：如何想把一个法语句子f翻译成英语e,那么相应的翻译信道模型就是假定法语句子f作为信道模型的输出，它原本就是一个英语句子e，但通过噪声通道传输时被改变成了法语句子f.那么现在需要做的就是如何根据概率p(e)和p(f|e)的计算求出最接近原始英文句子e的解

NLP

Phonemes 音素

Morphemes 形态素，词素

Taxonomy （生物）分类学

Ontology 本体论

Synonyms 同义词

Syntactic 句法的

traditional, task-specific feature engineering

word vector representations, window-based neural networks, recurrent neural networks, long-short-term-memory models, recursive neural networks, convolutional neural networks

如何表示一个词语？

层级？如百合<花<植物<物体

同义词？好，不错，还行，棒棒哒，…

该representation问题：主观性，难题量化词语相似度，…

1957年，word-document矩阵 （词频统计模型） = window based co-occurrence matrix

Window size作为计算作用域，计算co-occurrence matrix,每个词都有一个vector表示，可以用Euclidean distance倒数，或cosine, 或相关系数表示任意两个词的相似度

1990年, word-document矩阵过于sparse,能否降到低维？(特征值分解即PCA；奇异值分解即SVD)

entropy normalization（见Latent Semantic Analysis）

2009年，correlation normalization （见COALS）

词频-逆文档频度（Term Frequency - Inverse Document Frequency，TF-IDF）

评估一个词对于一个文档集或语料库中某个文档的重要程度。字词的重要性随着它在文件中出现的次数成正比增加 ，但同时会 随着它在语料库中出现的频率成反比下降 。如果某个词比较少见，但是它在这篇文章中多次出现，那么它很可能就反映了这篇文章的特性，正是我们所需要的关键词。

词频 (term frequency, TF)指的是某一个给定的词语在该文件中出现的次数

逆文档频率(nverse Document Frequency，IDF) 是一个词语普遍重要性的度量，它的大小与一个词的常见程度成反比，计算方法是语料库的文档总数除以语料库中包含该词语的文档数量，再将得到的商取对数。

自动提取关键词的算法就很清楚了，就是计算出文档的每个词的TF-IDF值，然后按降序排列，取排在最前面的几个词。

**词向量**

Distributed representation 用来表示词，通常被称为“**Word Representation**”或“Word Embedding”，中文俗称“词向量

如果用传统的稀疏表示法表示词，在解决某些任务的时候（比如构建语言模型）会造成维数灾难[Bengio 2003]。使用低维的词向量就没这样的问题,相似词的词向量距离相近，这就让基于词向量设计的一些模型自带平滑功能，让模型看起来非常的漂亮。

从大量未标注的普通文本数据中无监督地学习出词向量（语言模型本来就是基于这个想法而来的），可以猜测，如果用上了有标注的语料，训练词向量的方法肯定会更多

要从一段无标注的自然文本中学习出一些东西，无非就是统计出词频、词的共现、词的搭配之类的信息。

要从自然文本中统计并建立一个语言模型,必然也需要对语言进行更精细的统计和分析，同时也会需要更好的模型，更大的数据来支撑。目前最好的词向量都来自于此

**语言模型**

语言模型其实就是看一句话是不是正常人说出来的。这玩意很有用，比如机器翻译、语音识别得到若干候选之后，可以利用语言模型挑一个尽量靠谱的结果。

语言模型形式化的描述就是给定一个字符串，看它是自然语言的概率 P(w1,w2,…,wt)。

P(w1,w2,…,wt)=P(w1)×P(w2|w1)×P(w3|w1,w2)×…×P(wt|w1,w2,…,wt−1)

方法：n-gram模型，决策树，最大熵模型，马尔可夫模型，条件随机场，神经网络等

如n-gram模型就是用 P(wt|wt−n+1,…,wt−1) 近似P(w1,w2,…,wt)。

CS224d: Deep Learning for NLP

Deep Learning 算法已经在图像和音频领域取得了惊人的成果，但是在 NLP 领域中尚未见到如此激动人心的结果?

语言（词、句子、篇章等）属于人类认知过程中产生的高层认知抽象实体，而语音和图像属于较为底层的原始输入信号。

traditional method: feature-engineering

Most machine learning methods work well because of human-designed representa7ons and input features

Describing your data with features requires Domain specific

Deep Learning?

Deep learning algorithms a\empt to learn (mul7ple levels of) representa7on and an output

why deep learning?

• Manually designed features are o]en over-specified, incomplete and take a long 7me to design and validate

• Learned Features are easy to adapt, fast to learn

• Deep learning provides a very flexible, (almost?) universal, learnable framework for represen>ng world, visual and linguis7c informa7on.

• Deep learning can learn unsupervised (from raw text) and supervised (with specific labels like positive/negative)

How to represent the meaning of a word?

Taxonomy （生物）分类学 e.g. WordNet

One-hot representation: regard word as atomic symbols

Window-based co-occurrence matrix: Represent word by means of its neighbors

How to reduce dimensionality (around 25 – 1000)

Singular Value Decomposition: 求AAT 对应word的特征向量

缺点：计算量大，而且不易添加新的句子和文档

Word2vec