Noisy Channel Model

$\propto$: proportional

$$p(\text{text}|\text{source}) \propto p(\text{source}|\text{text})\, p(\text{text})$$

Bayes

$$p(\text{text}|\text{source}) = \frac{p(\text{source}|\text{text}) \cdot p(\text{text})}{p(\text{source})}$$

应用场景:

语音识别，机器翻译，拼写纠错，OCR，密码破解 → 文本

$$p(\text{text}|\text{source}) \propto p(\text{source}|\text{text})\, p(\text{text})$$

机器翻译 i.e. 英 → 中

$$\underbrace{P(\text{中文}|\text{英文})}_{\text{argmax}} \propto \underbrace{P(\text{英文}|\text{中文})}_{\text{Translation}} \cdot \underbrace{P(\text{中文})} \rightarrow \text{语言模型}$$

拼写纠错

$$P(\text{正确的写法}|\text{错误的写法}) \propto \underbrace{P(\text{错误的写法}|\text{正确的写法})} \cdot \underbrace{P(\text{正确的写法})}_{\text{语言模型}}$$

$$p(\text{text}|\text{source}) \propto p(\text{source}|\text{text})\, p(\text{text})$$

语音识别 输入：~~~~~

$$P(\text{文本}|\text{语音信号}) \propto \underbrace{P(\text{语音信号}|\text{文本})}_{\substack{\text{Translation}\\ \text{Recognition model}}} \cdot \underbrace{P(\text{文本})}_{\to \text{语言模型}}$$

密码破解 输入：encrypted string (abfde…)

$$P(\text{明文}|\text{暗文}) \propto P(\text{暗文}|\text{明文}) \cdot \underbrace{P(\text{明文})}_{} \to \text{语言模型}$$

Language Model

语言模型用来判断：是否一句话从语法上通顺

Model

↓

pre-trained

P_{LM}(今天是周日) > P_{LM}(今天周日是)

P_{LM}(全民AI是趋势) > P_{LM}(趋势全民AI是)

Recap: Chain Rule

Random variable

$$P(A,B) = P(A|B)\cdot P(B)$$
$$= P(B|A)\cdot P(A)$$

- $p(A, B, C, D) = P(A)\cdot P(B|A)\cdot P(C|A,B)\cdot P(D|A,B,C)$ ← chain rule

$$= P(A,B)\cdot P(C|A,B)\cdot P(D|A,B,C)$$

$$= P(A,B,C)\cdot P(D|A,B,C) = P(A,B,C,D)$$

- $p(w_1, w_2, w_3, w_4, w_5 \dots w_n)$

$$= P(w_1)\cdot P(w_2|w_1)\cdot P(w_3|w_1 w_2)\cdot P(w_4|w_1 w_2 w_3)\cdot P(w_5|w_1 w_2 w_3 w_4)$$

$$\cdots\cdots P(w_n|w_1 w_2 \cdots w_{n-1})$$ ← chain rule

Chain Rule for Language Model

- $p(w_1, w_2, w_3, w_4, w_5 \dots w_n)$

- $p($今天，是，春节，我们，都，休息$)$

$= p($今天$) \cdot p($是$|$今天$) \cdot p($春节$|$今天, 是$) \cdot p($我们$|$今天, 是, 春节$)$
$\cdot p($都$|$今天, 是, 春节, 我们$) \cdot p($休息$|$今天, 是, 春节, 我们, 都$)$

Chain Rule for Language Model

- p(休息 | 今天，是，春节，我们，都) $\doteq 0$

Sparsity

P(运动 | 今天是春节 …) $= 0$

今天是春节我们都休息

今天是春节我们都放假

今天，(今天是)

6个

□ □ □ □ □ □

10^5

$(10^5)^6$

Markov Assumption

- p(休息 | 今天，是，春节，我们，都)

$\approx P$(休息 | 都)

→ 1st order markov assumption

- p(休息 | 今天，是，春节，我们，都)

$\approx P$(休息 | 我们, 都)

2nd order markov assumption

- p(休息 | 今天，是，春节，我们，都)

$\approx P$(休息 | 春节, 我们, 都)

3rd order ..

Markov Assumption

- $p(w_1, w_2, w_3, w_4, w_5 \dots w_n)$

$= p(w_1) \cdot p(w_2|w_1) \cdot p(w_3|w_2) \cdots p(w_n|w_{n-1}) = p(w_1) \prod_{i=2}^{n} p(w_i|w_{i-1})$ 1st order

- $p(w_1, w_2, w_3, w_4, w_5 \dots w_n)$

2nd order

$= p(w_1) \cdot p(w_2|w_1) \cdot p(w_3|w_1, w_2) \cdot p(w_4|w_2, w_3) \cdots p(w_n|w_{n-2} w_{n-1})$

$= p(w_1) \cdot p(w_2|w_1) \cdot \prod_{i=3}^{n} p(w_i|w_{i-2}, w_{i-1})$

- $p(w_1, w_2, w_3, w_4, w_5 \dots w_n)$

3rd order

Language Model (Use ~~2nd~~ 1st Order)

LM

$p(是|今天) = 0.01$

$p(今天) = 0.002$

$p(周日|是) = 0.001$

$p(周日|今天) = 0.0001$

$p(周日) = 0.02,$

$p(是|周日) = 0.0002$

比较: 今天是周日 VS 今天周日是

$P_{LM}(今天是周日)$

$= P(今天) \cdot P(是|今天) \cdot P(周日|是)$

$= 0.002 \cdot 0.01 \cdot 0.001$

$= 2 \times 10^{-8}$

$>$

$P_{LM}(今天周日是)$

$= P_{LM}(今天) \cdot P(周日|今天) \cdot P(是|周日)$

$= 0.002 \cdot 0.0001 \cdot 0.0002$

$= 4 \times 10^{-10}$

Language Model : Unigram

- $p(w_1, w_2, w_3, w_4, w_5 \ldots w_n)$

$= p(w_1) \cdot p(w_2) p(w_3) \cdots p(w_n)$

- p(今天，是，春节，我们，都，休息) ✓

$= p$(今天)$\cdot p$(是)$\cdot p$(春节)$\cdot p$(我们)p(都)p(休息)

- p(今天，春节，是，都，我们，休息)

$= p$(今天)$\cdot p$(春节)p(是)p(都)p(我们)p(休息)

Language Model : Bigram

← 1st order markov assumption

- $p(w_1, w_2, w_3, w_4, w_5 \dots w_n)$

$= p(w_1) \cdot p(w_2|w_1) \cdot p(w_3|w_2) \cdots p(w_n|w_{n-1}) = p(w_1) \cdot \prod_{i=2}^{n} p(w_i|w_{i-1})$

- p(今天，是，春节，我们，都，休息)

$= p$(今天) $\cdot$ p(是|今天) $\cdot$ p(春节|是) $\cdot$ p(我们|春节) $\cdot$ p(都|我们) $\cdot$ p(休息|都)

- p(今天，春节，是，都，我们，休息)

$= p$(今天) $\cdot$ p(春节|今天) $\cdot$ p(是|春节) $\cdot$ p(都|是) $\cdot$ p(我们|都) $\cdot$ p(休息|我们)

Language Model : N-gram

$N>2$

Higher Order

$N=3$

- $p(w_1, w_2, w_3, w_4, w_5 \ldots w_n)$

$= P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1 w_2)\, P(w_4|w_2 w_3) \cdots P(w_n|w_{n-2} w_{n-1}) = P(w_1)\, P(w_2|w_1) \prod_{i=3}^{n} P(w_i|w_{i-2} w_{i-1})$

- p(今天，是，春节，我们，都，休息)

$=$ P(今天) P(是|今天) P(春节|今天,是) P(我们|是,春节) ⋯⋯

- p(今天，春节，是，都，我们，休息)

⋯⋯

Unigram：Estimating Probability

- $p(w_1, w_2, w_3, w_4, w_5 \ldots w_n)$

$= p(w_1)\, p(w_2) \cdots \quad p(w_n)$

$p(w_1) = ?$

$p(w_2) = ?$

$p(w_3) = ?$

$\vdots$

$p(w_n) = ?$

我们 - - .

.. 我们 - - .

- - - 我们 - -

$c(\text{我们}) = 100$ 次

$V = 10^5$

$p(\text{我们}) = \frac{100}{10^5} = 1‰$

Unigram：Estimating Probability

语料库

今天 的 天气 很好 啊
我 很 想 出去 运动
但 今天 上午 有 课程
训练营 明天 才 开始

$V=19$

P_{LM}(今天 开始 训练营 课程)

$= P_{LM}(\text{今天}) \cdot P_{LM}(\text{开始}) \cdot P_{LM}(\text{训练营}) P_{LM}(\text{课程})$

$= \frac{2}{19} \cdot \frac{1}{19} \cdot \frac{1}{19} \cdot \frac{1}{19} = \frac{2}{19^4}$

P_{LM}(今天 没有 训练营 课程)

$= P_{LM}(\text{今天}) \cdot P_{LM}(\text{没有}) P_{LM}(\text{训练营}) \cdot P_{LM}(\text{课程})$

$= \frac{2}{19} \cdot \frac{0}{19} \cdot \frac{1}{19} \cdot \frac{1}{19} = 0$

01:23:41 -01:21:34

Bigram：Estimating Probability

1st order markov assumption

- $p(w_1, w_2, w_3, w_4, w_5 \dots w_n)$

$$= p(w_1) \cdot p(w_2|w_1)\, p(w_3|w_2) \cdots p(w_n|w_{n-1})$$

? ? ? ?

P(是|明天)

明天是
明天我们
明天上课
明天是
明天会

$P(是|明天) = \frac{2}{5}$

$P(我们|明天) = \frac{1}{5}$

$P(上课|明天) = P(会|明天) = \frac{1}{5}$

Bigram：Estimating Probability

语料库

今天 的 天气 很好 啊
我 很 想 出去 运动
但 今天 上午 想 上课
训练营 明天 才 开始

$V=19$

P_{LM}(今天 上午 想 出去 运动)

$= P_{LM}$(今天) $\cdot P_{LM}$(上午|今天) $\cdot P_{LM}$(想|上午) P_{LM}(出去|想) P_{LM}(运动|出去)

$= \frac{2}{19} \cdot \frac{1}{2} \cdot 1 \cdot \frac{1}{2} \cdot 1 = \frac{1}{38}$

P_{LM}(今天 上午 的 天气 很好 呢)

$= P_{LM}$(今天) $\cdot P_{LM}$(上午|今天) $\cdot \underbrace{P(\text{的}|\text{上午})}_{0} \cdot \cdots$

$= 0$

N-gram：Estimating Probability

$N=3$

语料库

今天 上午 的 天气 很好
我 很 想 出去 运动
但 今天 上午 有 课程
训练营 明天 才 开始

今天 上午 有 课程

$= P_{Lm}(\text{今天}) \cdot P_{Lm}(\text{上午} \mid \text{今天}) \cdot \underbrace{P_{Lm}(\text{有} \mid \text{今天}, \text{上午})}_{\frac{1}{2}} \cdot \underbrace{P_{Lm}(\text{课程} \mid \text{上午}, \text{有})}_{1}$

今天 没有 训练营 课程

$= P_{Lm}(\text{今天}) \cdot P_{Lm}(\text{没有} \mid \text{今天}) \cdot \underbrace{P_{Lm}(\text{训练营} \mid \text{没有})}_{0} \cdots$

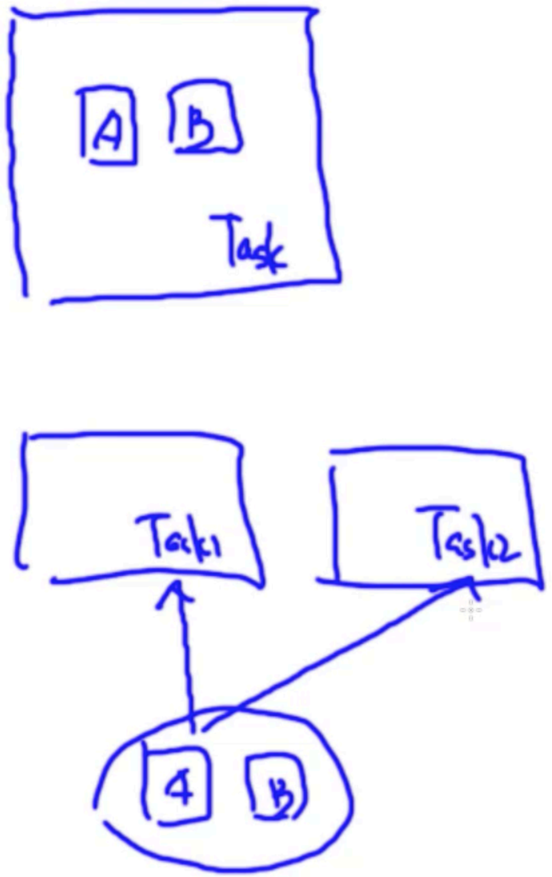
Evaluation of Language Model

Q: 训练出来的语言模型效果好还是坏?

- 理想情况下
  1. 假设有两个语言模型 A,B
  2. 选定一个特定的任务比如拼写纠错
  3. 把两个模型A,B都应用在此任务中
  4. 最后比较准确率，从而判断A,B的表现

A: 91% B: 89%

A > B



Evaluation of Language Model

核心思路

今天____

今天天气____，

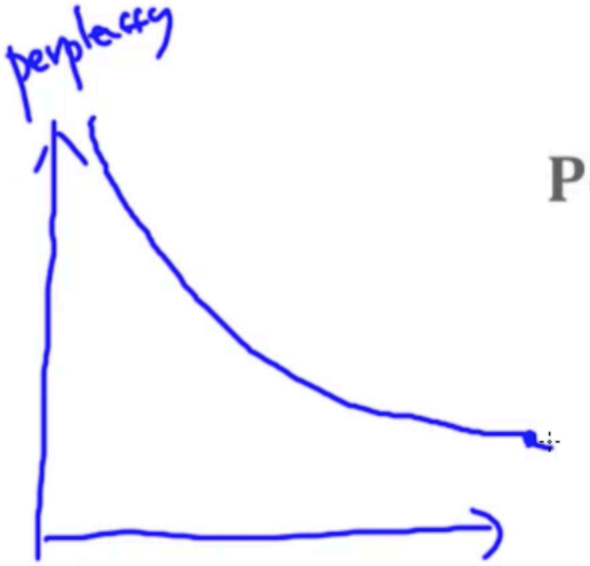
今天天气很好，____

今天天气很好，适合____

今天天气很好，适合出去____

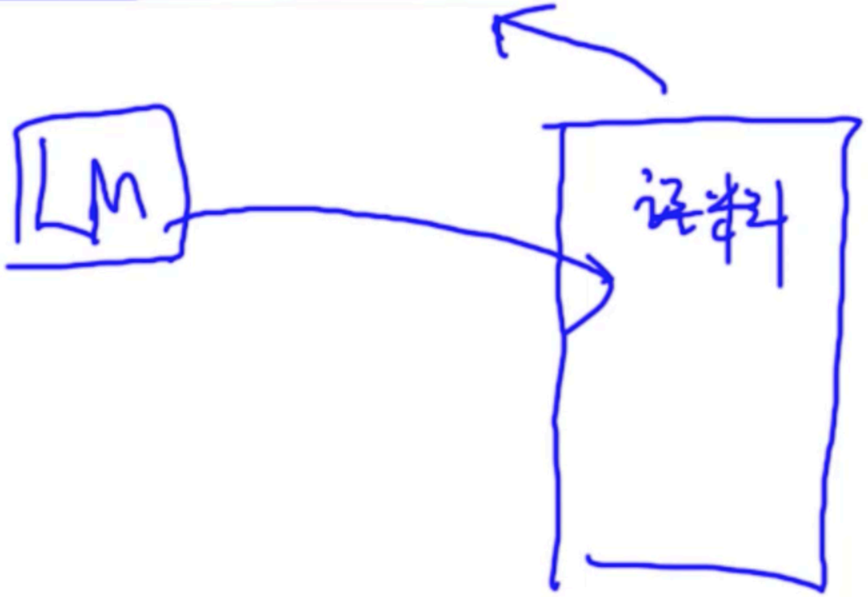
Perplexity

无 (unsupervised)



$$\text{Perplexity} = 2^{-(x)}$$

x: average log likelihood



2^{-x}

perplexity ↓

越小越好

Perplexity

$x = \frac{a_1 - 2 - 1 - 2 + a_2 - 1}{6}$

$\text{Perplexity} = 2^{-(x)}$ x: average log likelihood

训练好的Bigram

p(天气|今天) = 0.01
p(今天) = 0.002
p(很好|天气) = 0.1
p(适合|很好) = 0.01
p(出去|适合) = 0.02,
p(运动|出去) = 0.1

今天 p(今天) = 0.002 ⇒ log p(今天) = a_1

今天天气 p(天气|今天) = 0.01 ⇒ log p(天气|今天) = −2

今天天气很好, p(很好|天气) = 0.1 ⇒ log p(很好|天气) = −1

今天天气很好，适合 p(适合|很好) = 0.01 ⇒ −2

今天天气很好，适合出去 p(出去|适合) = 0.02 = a_2

今天天气很好，适合出去运动 p(运动|出去) = 0.1 = −1

Perplexity

$$\textbf{Perplexity} = 2^{-(x)} \qquad x\text{: average log likelihood}$$

Training 38 million words, test 1.5 million words, WSJ

N-gram Order	Unigram	Bigram	Trigram
Perplexity	962	170	109

Slide Credit: Dan Jurafsky

Recap: Estimating Probability

语料库

Bigram

今天 上午 的 天气 很好

我 很 想 出去 运动

但 今天 上午 有 课程

训练营 明天 才 开始

今天 训练营 没有

$P_{LM}(今天)\cdot P_{LM}(训练营|今天)\cdot P_{LM}(没有|训练营)$

$= ✓ \cdot 0 = 0$

今天 没有 训练营 课程

$= P_{LM}(今天)\cdot P_{LM}(没有|今天)\cdots$

$= \cdots 0 = 0$

Smoothing

- Add-one Smoothing
- Add-K Smoothing
- Interpolation
- Good-Turning Smoothing

Add-one Smoothing (Laplace Smoothing)

Naive Bayes

$$P_{MLE}(w_i|w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_i)} \Rightarrow 0 \times$$

$$P_{Add-1}(w_i|w_{i-1}) = \frac{c(w_{i-1}, w_i) + \mathbf{1}}{c(w_i) + \mathbf{V}} \Rightarrow ?$$

而不是 2V, 100, 1000, 3V ?

V = 20

c(我们) = 3

c(我们, 是) = 0

P_{MLE}(是 | 我们) = 0

P_{Add-1}(是 | 我们) $= \frac{0+1}{3+20} = \frac{1}{23}$

Add-one Smoothing (Laplace Smoothing)

$V = 17$

语料库

今天 上午 的 天气 很好

我 很 想 出去 运动

但 今天 上午 有 课程

训练营 明天 才 开始

$$P_{Add-1}(w_i|w_{i-1}) = \frac{c(w_{i-1}, w_i) + \mathbf{1}}{c(w_i) + \mathbf{V}}$$

$$P_{Add-1}(\text{上午}|\text{今天}) = \frac{2+1}{2+17} = \frac{3}{19}$$

$$P_{Add-1}(\text{的}|\text{今天}) = \frac{0+1}{2+17} = \frac{1}{19}$$

$$P_{Add-1}(\text{天气}|\text{今天}) = \frac{1}{19}$$

$\vdots$

$$P_{Add-1}(\text{有}|\text{今天}) = \frac{1}{19}$$

$$\frac{3}{19} + \frac{16}{19} = \frac{19}{19} = 1$$

Add-K Smoothing (Laplace Smoothing)

语料库

今天 上午 的 天气 很好

我 很 想 出去 运动

但 今天 上午 有 课程

训练营 明天 才 开始

$$P_{Add-k}(w_i|w_{i-1}) = \frac{c(w_{i-1}, w_i) + \mathbf{k}}{c(w_i) + \boldsymbol{kV}}$$

k=1 时 ⇒ Add-one smoothing

$\boxed{k=3}$

$$P_{Add-k}(上午|今天) = \frac{2+3}{2+3\cdot 17} = \frac{5}{53}$$

$$P_{Add-k}(的|今天) = \frac{0+3}{2+3\cdot 17} = \frac{3}{53}$$

Add-K Smoothing (Laplace Smoothing)

$$P_{Add-k}(w_i|w_{i-1}) = \frac{c(w_{i-1}, w_i) + k}{c(w_i) + kV}$$ → 怎么选择?

① $k = 1, 2, 3 \ldots 100$

② 优化 $f(k)$ ..

训练集语料库
今天 上午 的 天气 很好
我 很 想 出去 运动
但 今天 上午 有 课程
训练营 明天 才 开始

$\underline{LM_{训练}}$ → $C(w_{i-1}, w_i)$, $C(w_i)$, :

验证集语料库
今天 上午 想 出去 运动
明天 才 开始 训练营

$\boxed{\text{perplexity} = f(k)}$

Minimize perplexity

$\Updownarrow$

$\boxed{\text{minimize } f(k)}$

$\Downarrow$

$\hat{k} = \arg\min_{k} f(k)$

Interpolation

Tri-gram

Unigram

Bigram

count. 训练数据

Trigram

C(in the kitchen) = 0

C(the kitchen) = 3

C(kitchen) = 4

C(arboretum) = 0

unigram

p(kitchen | in the) = 0

p(arboretum | in the)= 0

in the

kitchen

arboretum

Interpolation

C(in the kitchen) = 0

C(the kitchen) = 3

C(kitchen) = 4

C(arboretum) = 0

p(kitchen | in the) =

p(arboretum | in the)=

核心思路

在计算Trigram概率时同时考虑Unigram, Bigram, Trigram出现的频次

Interpolation

Tri-gram

$$\underline{p(w_n|w_{n-1}, w_{n-2})} = \overset{0.5}{\lambda_1} p(w_n|w_{n-1}, w_{n-2}) \rightarrow \text{trigram}$$

$$\overset{0.3}{+ \lambda_2} p(w_n|w_{n-1}) \rightarrow \text{bigram}$$

$$\overset{0.2}{+ \lambda_3} p(w_n) \rightarrow \text{unigram}$$

$$\boxed{\lambda_1 + \lambda_2 + \lambda_3 = 1}$$

方法

0.6 + 0.1 + 0.3 = 1

方法1 方法2 方法3