N-грамм модели

Модель языка: p: предложение → [0,1]

каждому предложению сопоставляете "вероятность"

$\sum_{s \in \text{предл}} p(s) = 1$.

- выбор самого вероятного в-та из нескольких (→ перевод, → исправл. ошибок)
- генерация текста, выбор вероятных предложений

p(я иду) > p(я идёт) > 0

$p(s) = p(w_1 w_2 w_3 \dots w_N) =$ — n слов в предл.

$= p(w_N | w_1 w_2 \dots w_{N-1}) \cdot p(w_1 \dots w_{N-1})$

$= p(w_N | w_1 w_2 \dots w_{N-1}) \cdot$
$p(w_{N-1} | w_1 w_2 \dots w_{N-2}) \cdot$
$\dots$
$p(w_2 | w_1) \cdot$
$p(w_1)$

$p(AB) = p(B|A) \cdot P(A)$

$P(B|A) = \frac{P(BA)}{P(A)}$

p(я вижу кота) =
= p(кота | я вижу) ·
· p(я вижу)

*** я вижу кота

= в начале предложения есть спец слово *

≈ оценим в-ть $p(w_i | w_1 w_2 \dots w_{i-1})$ как $p(w_i | \underbrace{w_{i-n+1} \dots w_{i-1}}_{n \text{ слов}})$

$$= \left. \begin{aligned} & p(w_N | w_{N-n+1} \dots w_{N-1}) \cdot \\ & p(w_{N-1} | \dots \dots w_{N-2}) \cdot \\ & \dots \\ & p(w_1 | \underbrace{* * * \dots *}_{n-1}) \end{aligned} \right\} (1)$$

"*"

правильная запись

$p(\mathcal{W}_N = w_n | \mathcal{W}_{N-1} = w_{n-1})$

сл. вел — конкр. значение

Если $n=2$.

$$p(w_1 w_2 \dots w_N) = p(w_1 | *) \cdot p(w_2|w_1) \cdot p(w_3|w_2) \cdots p(w_N|w_{N-1})$$

если $n=1$

$$p(w_1 w_2 \dots w_N) = p(w_1) \cdot p(w_2) \cdots p_N(w_N)$$

Как оценить $p(w_i | w_j)$?

Нужен корпус текстов для обучения модели.

Считаем встречаемость n-грамм, т.е. последовательностей из n слов. $c(\dots) =$ кол-во раз (count)

Я иду гулять, Я вижу кота, Я вижу сон.

$c(\text{Я иду}) = 1$ $\quad c(\text{Я вижу}) = 2$

$$p(w_n | w_1 w_2 \dots w_{n-1}) = \text{вероятность увидеть } w_n \text{ при условии, что перед этим прочитали } w_1 \dots w_{n-1}$$

$$= \frac{c(w_1 \dots w_{n-1} w_n)}{\sum_{w \in V} c(w_1 \dots w_{n-1} w)}$$

V – словарь

$$= \frac{c(w_1 \dots w_{n-1} w_n)}{c(w_1 \dots w_{n-1})} \quad (2)$$

$$p(\text{вижу} | \text{Я}) = \frac{c(\text{Я вижу})}{c(\text{Я})} = \frac{2}{3}$$

Всё, модель построена!!!

$p(w_1 \dots w_N) = (1)$ = оценивается через (2).

При генерации текста:

даём $n-1$ слово $w_1, w_2 \dots w_{n-1}$

далее w_n выбираете случайно по распреде-

нужно $p(\cdot | w_1 \dots w_{n-1})$. потом w_{n-1} по распределению $p(\cdot | w_2 w_3 \dots w_n)$.

Сглаживание

Проблема 1. p(я буду тирамисушка) = p(Тирамисушка|...) – не было слова.

p(Тир. | .....) = 0. ⇒ результат тоже 0.

Обычно все незнакомые слова заменяются на <UNK> (unknown)

Считаем, что словарь – это все встреченные слова + слово UNK.

$p(w_n | w_1 \dots w_{n-1}) = 0$, если слово w_n не встречалось после $w_1 \dots w_{n-1}$.

"я буду спать"

в корпусе есть все эти слова, но нет этой фразы!

при $n = 3$

p(спать | я буду) = $\frac{0}{\dots} = 0$

⇒ p всего предложения тоже будет 0.

Сглаживание делает все вероятности $\neq 0$

Варианты сглаживания.

1. Лаплас (Lid...)

$$p(w_n | w_1 \dots w_{n-1}) = \frac{c(w_1 w_2 \dots w_N) + \alpha}{c(w_1 \dots w_{N-1}) + \alpha |V|}$$

(выберем)

V – словарь

$|V|$ – размер словаря

$p(\cdot | w_1 \dots w_{n-1})$ вер-ть

Лаплас: $\alpha = 1$ (можно $\alpha = 0{,}1$ и т.п.)

2. Интерполяция, для откета

$$p(w_n \mid w_1 \dots w_{n-1}) = \frac{c(w_1 w_2 \dots w_n)}{c(w_1 w_2 \dots w_{n-1})} \leftarrow n\text{-грамм}$$

$$\frac{c(w_2 \dots w_n)}{c(w_2 \dots w_{n-1})} \leftarrow n-1 \text{ грамм}$$

$$\frac{c(w_3 \dots w_n)}{c(w_3 \dots w_{n-1})} \leftarrow n-2 \text{ грамм}$$

т.е. $\tilde{p}(\text{снег} \mid \text{я иду}) = p(\text{снег} \mid \text{я иду})$ $p(\text{снег} \mid \text{иду})$

$$\frac{c(\text{я иду снег})}{c(\text{я иду})} \qquad \frac{c(\text{иду снег})}{c(\text{иду})}$$

$$p(\text{снег}) = \frac{c(\text{снег})}{c(\dots)}$$

Как использовать оценки меньшей размерности? Интерполяция. Сложить с коэффициентами

$n=3$ для примера

$$\tilde{p}(w_3 \mid w_1, w_2) = \alpha_3 p(w_3 \mid w_1, w_2) + \alpha_2 p(w_3 \mid w_2) + \alpha_1 p(w_3)$$

$$\alpha_3 + \alpha_2 + \alpha_1 = 1 \qquad \alpha_i \geq 0$$

α_i – надо подобрать.

Простой откат. Если $p(w_3 \mid w_1, w_2) = 0$, взять $p(w_3 \mid w_2)$ Если тоже $= 0$ $p(w_3)$

Чтоб $p(\bullet \mid w_1, w_2)$ – была вероятность, делаем так

$$\tilde{p}(w_3 \mid w_1, w_2) = \begin{cases} 0.6\, p(w_3 \mid w_1, w_2), & \text{если} \neq 0 \\ 0.4\, p(w_3 \mid w_2) \cdot 0{,}6 & \text{иначе, если} \neq 0 \\ 0.4\, p(w_3) \cdot 0{,}6 & \text{иначе, если} \neq 0 \\ 0.4 \cdot 0.4 & \end{cases}$$

Последний метод

Kneser-Ney

для $n=2$

$$\tilde{p}(w_2|w_1) = \frac{c(w_1,w_2)-\delta}{c(w_1)} + \lambda \frac{|\{w : (w,w_2)\}|}{c(w_2\ w_2)}$$

обычно

если $c(w_1,w_2) \neq 0$

сколько разных слов будет w_2

- разных пар слов

оценка разнообразия возможных слев перед w_2.

(New York)

↑ для York

Больше нуля.

Как оценить качество модели?

1. Метрика качества

$p\{\text{текста}\}$.

если текст реален, => его вер-ть должна быть ↑↑ max.

$p\{\text{текста}\} = p\{w_1\ w_2 \dots\dots w_n\} = (1)$.

невозможно посчитать из-за underflow (очень мал)

$$-\log_2 p\{\text{текста}\} = -\log_2 p(w_1|*\ *\ \dots\ *) - \log_2 p(w_2|*\ \dots\ w_1) - \dots$$

↑ будет больше.

Это называется perplexity. Должна быть как можно меньше.

2. корпус

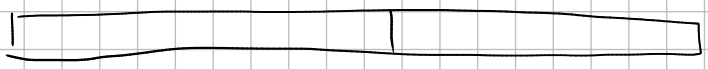
учебный	тестовый
80%	20%

обучения — perplexity.

3. Но мы хотим выбрать лучшую модель.

- Лаплас ($\alpha=1$, $\alpha=0.1$)
- Кнезер-Нез δ
- Simple Backoff

} гиперпараметры



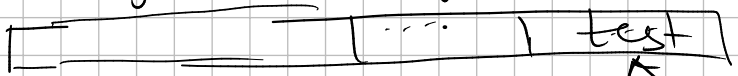
модель 1
модель 2

перплексити модели 1 // 2

← можно выбрать лучшую модель.

Надо

обучение. настройка test



60% 20%

оценить перплексити лучшей модели.

↑
для выбора лучшей модели

Надо

- взять корпус
- настроить разные модели
- выбрать лучшую и оценить её перплексити
- погенерировать тексты.