Языковые модели

Гусев Илья

Московский физико-технический институт

Москва, 2018

Содержание

Языковые модели

Статистичская языковая модель (statistical language model) - вероятностное распределение над последовательностями слов $P(w_1, ..., w_m)$.
Применения:

1. Распознвание речи (ASR)
2. Машинный перевод (MT)
3. PoS-tagging
4. OCR
5. Распознавание рукописных текстов
6. Классификация текстов

В целом, нужны везде, где речь идёт о последовательностях слов. Мы рассматриваем языковые модели именно на уровне слов, но бывают ещё и char-level и subword модели.

Train, validation(dev), test

- Для обучения практически любой языковой модели нужен большой корпус. Достаточно иметь тексты, разбитые на токены.
- Train выборка - собираем статистику или учим модель.
- Validation(dev) выборка - выбираем гиперпараметры модели, таким образом, чтобы языковая модель лучше работала на этой выборке.
- Test выборка - оцениваем качество языковой модели.

Перплексия

$$PP(W) = P(w_1 w_2 ... w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1 w_2 ... w_N)}}$$

Средневзвешенное количество слов, которые могут следовать за данным словом.

Пример: язык из 9 символов $0, 2, ..., 9$, для каждого из них $P = \frac{1}{10}$.

$$PP(W) = (\frac{1}{10}^{N})^{-\frac{1}{N}} = 10$$

Наивные N-граммы

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)...P(w_n|w_1^{n-1}) = \prod_{k=1}^{n} P(w_k|w_1^{k-1})$$

Биграммная модель:

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-1})$$

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)} = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

N-граммная модель:

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

Наивные N-граммы

Пример

<s> I am Sam </s>
<s> Sam I am </s>
<s> I do not like green eggs and ham </s>

$$P(I|\langle s\rangle) = 2/3 = .67$$

$$P(Sam|\langle s\rangle) = 1/3 = .33$$

$$P(am|I) = 2/3 = .67$$

$$P(\langle/s\rangle|Sam) = 1/2 = 0.5$$

$$P(Sam|am) = 1/2 = .5$$

$$P(do|I) = 1/3 = .33$$

Наивные N-граммы

Проблемы

- OOV - out of vocabulary: слова, которых не было в словаре обучающей выборки
  - Выбор размера словаря влияет на перплексию
- Нули в test выборке: n-граммы встретились в test, но не встретились в train
- Недооценка не встретившихся n-грамм
- Переоценка низкочастотных n-грамм

Сглаживание, backoff и интерполяция

Add-k сглаживание

$$P^*_{Add-k}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n) + k}{C(w_{n-1}) + kV}$$

Частный случай: сглаживание Лапласа

$$P^*_{Laplace}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n) + 1}{C(w_{n-1}) + V}$$

$$c^*(w_n|w_{n-1}) = \frac{(C(w_{n-1}w_n) + 1) \times C(w_{n-1})}{C(w_{n-1}) + V}$$

Сглаживание, backoff и интерполяция

Интерполяция

$$\hat{P}(w_n|w_{n-2}w_{n-1}) = \lambda_1 P(w_n|w_{n-2}w_{n-1}) + \lambda_2 P(w_n|w_{n-1}) + \lambda_3 P(w_n)$$

$$\sum_i \lambda_i = 1$$

Усложнённый вариант:

$$\hat{P}(w_n|w_{n-2}w_{n-1}) = \lambda_1(w_{n-2}^{n-1})P(w_n|w_{n-2}w_{n-1}) +$$

$$\lambda_2(w_{n-2}^{n-1})P(w_n|w_{n-1}) + \lambda_3(w_{n-2}^{n-1})P(w_n)$$

Сглаживание, backoff и интерполяция

Katz backoff

$$P_{BO}(w_n|w_{n-N+1}^{n-1}) = \begin{cases} P^*(w_n|w_{n-N+1}^{n-1}), & \text{if } C(w_{n-N+1}^{n}) > 0 \\ \lambda(w_{n-N+1}^{n-1})P_{BO}(w_n|w_{n-N+2}^{n-1}), & \textit{otherwise} \end{cases}$$

Сглаживание, backoff и интерполяция

Не затронуты, но важны

- Good-Turing сглаживание и backoff
- Kneser-Ney сглаживание
- Modified Kneser-Ney сглаживание
- Stupid backoff

Сравнение

Model	Test Perplexity	Number of Params [billions]
Sigmoid-RNN-2048 (Ji et al., 2015a)	68.3	4.1
Interpolated KN 5-gram, 1.1B n-grams (Chelba et al., 2013)	67.6	1.76
Sparse Non-Negative Matrix LM (Shazeer et al., 2015)	52.9	33
RNN-1024 + MaxEnt 9-gram features (Chelba et al., 2013)	51.3	20
LSTM-512-512	54.1	0.82
LSTM-1024-512	48.2	0.82
LSTM-2048-512	43.7	0.83
LSTM-8192-2048 (No Dropout)	37.9	3.3
LSTM-8192-2048 (50% Dropout)	32.2	3.3
2-Layer LSTM-8192-1024 (BIG LSTM)	30.6	1.8
BIG LSTM+CNN Inputs	**30.0**	**1.04**
BIG LSTM+CNN Inputs + CNN Softmax	39.8	**0.29**
BIG LSTM+CNN Inputs + CNN Softmax + 128-dim correction	35.8	**0.39**
BIG LSTM+CNN Inputs + Char LSTM predictions	47.9	**0.23**

Полезные ссылки I

- Speech and Language Processing. Daniel Jurafsky, James H. Martin. Chapter 3
  https://web.stanford.edu/~jurafsky/slp3/3.pdf
- Exploring the Limits of Language Modeling
  https://arxiv.org/pdf/1602.02410.pdf