Языковые модели

Language models

«Высшая школа экономики»

15.10

Outline

Language models

Language model (LM) - model that predict a next word given a previous.

Language models

- **Machine translation**: translating a sentence saying about height it would probably state that $P(tallman) > P(largeman)$ as the 'large' might also refer to weight or general appearance thus, not as probable as 'tall'
- **Spelling Correction**: Spell correcting sentence: "Put you name into form", so that $P(nameintoform) > P(nameintofrom)$
- **Speech Recognition**: Call my nurse: $P(Callmynurse.) \gg P(coalminers)$, I have no idea. $P(noidea.) \gg P(Noeyedeer.)$
- Summarization, question answering, sentiment analysis etc.

Language models

- Statistical Language Models
- Neural Language Models

Probabilistic language modeling

Sentence:

$$P(S) = P(w_1, w_2, w_3, w_4) \equiv P(w_4|w_1, w_2, w3)$$

Probability Chain Rule:

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

$$P(S) = P(w_1, \ldots, w_n) = P(w_1)P(w_2|w_1)P(w_3)P(w_1, w_2) \ldots P(w_n|w_1, \ldots w_{n-1}) = \prod_i P(w_i|w_1, \ldots w_{i-1})$$

Probabilistic language modeling

Initial way

$$P(w_5|w_1, w_2, w_3, w_4) = \frac{count(w_1, w_2, w_3, w_4, w_5)}{count(w_1, w_2, w_3, w_4)}$$

Markov assumption:

$$P(w_1, \ldots, w_n) \approx \prod_i P(w_i|w_{i-k}, \ldots P(w_{i-1}))$$

Probabilistic language modeling

Bi-gram model

$$P(w_1, \ldots, w_n) \approx \prod_i P(w_i|w_{i-1})$$

Maximum Likelihood Estimate (MLE)

$$P(w_i|w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

Information measure

To remind:

- Surprisal
- Entropy
- Cross-Entropy
- Cross-Entropy Loss
- Perplexity

Information measure

Surprisal

$$s = \log(1/y_i)$$

Surprisal

Entropy

$$e = \sum_{0}^{n} y_i \log(1/y_i)$$

Entropy

Information measure

Cross-Entropy

$$c = \sum_{0}^{n} p_i \log(1/q_i)$$

Cross-Entropy

Cross-Entropy Loss

$$c = \sum_{0}^{1} p_i \log(1/q_i) = p_0 \log(1/q_0) + p_1 \log(1/q_1) = p_0 \log(1/q_0) + (1 - p_0) \log(1/(1 - q_0))$$

Binary Cross-Entropy Loss

Information measure

Perplexity

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