NLP Formula Sheet

04 - Language Modeling

Perplexity Formula:

$$PP(W) = P(w_1, w_2, \dots, w_N)^{-\frac{1}{N}} = \left(\prod_{i=1}^{N} \frac{1}{P(w_i|w_1,w_2,\dots,w_{i-1})}\right)^{\frac{1}{N}}$$

Unigram Model:

$$P(w_i \mid w_{i-1}) = \frac{C(w_{i-1},w_i)+1}{C(w_{i-1})+V}$$

Good-Turing Estimation:

$$P_{GT}(w) = \frac{(C(w)+1)N_{C(w)+1}}{N_{C(w)}N}$$

05 - Spell Correction

Maximum Likelihood Estimation:

$$\text{argmax}_{c\in\mathcal{C}} P(c \mid w) = \text{argmax}_{c\in\mathcal{C}} P(w \mid c)P(c)$$

Minimum Edit Distance:

$$D[i,j] = \min \begin{cases} D(i-1,j)+1 \\ D(i,j-1)+1 \\ D(i-1,j-1)+\delta(a_i,b_j) \end{cases}$$

Cost Function:

$$\delta(a_i,b_j) = \begin{cases} 0 & \text{if } a_i = b_j \\ 2 & \text{if } a_i \neq b_j \end{cases}$$

06 - Word Embedding

Term Frequency (TF):

$$TF(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

Inverse Document Frequency (IDF):

$$IDF(t,D) = \log_{10}\left(\frac{\text{Total number of documents } |D|}{\text{Number of documents containing term } t}\right)$$

TF-IDF:

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

Pointwise Mutual Information (PMI):

$$PMI(w,c) = \log_{10}\left(\frac{P(w,c)}{P(w)\cdot P(c)}\right)$$

Positive Pointwise Mutual Information (PPMI):
$PPMI(w, c) = \max(PMI(w, c), 0)$

Skip-gram Objective:
Maximize $\sum \log_{10} P(w_t \mid w_{t-k}, \ldots, w_{t+k})$

Continuous Bag of Words (CBOW) Objective:
Maximize $\sum \log_{10} P(w_{t-k}, \ldots, w_{t+k} \mid w_t)$

Cosine Similarity:
$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$

GloVe Loss Function:
$J = \sum_{i,j=1}^{V} f(P_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log_{10}(X_{ij}))^2$

07 - Sequence Modeling

LSTM Hidden State Calculation:
$h_t = \sigma(W_h h_{t-1} + W_x x_t + b_h)$

$y_t = \sigma(W_y h_t + b_y)$

$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$

$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$

$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$

$C_t = f_t * C_{t-1} + i_t * \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$

08 - Transformers

Scaled Dot-Product Attention:
$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

Multi-Head Attention:
$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

Positional Encoding:
$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)$

$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)$

10 - POS Tagging

Word Probability Given Tag:
$P(w_i \mid t_i)$

Tag Transition Probability:
$P(t_i \mid t_{i-1})$

CRF Formula:
$P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_j f_j(y_{i-1}, y_i, \mathbf{x}, i)\right)$

Linear CRF Formula:
$P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{i=1}^{n} \sum_{k=1}^{K} \theta_k \cdot g_k(y_{i-1}, y_i, \mathbf{x}, i)\right)$

11 - Machine Translation

Encoder Hidden State Calculation:
$h_t = f(h_{t-1}, x_t)$

Decoder Hidden State Calculation:
$s_t = f(s_{t-1}, y_{t-1}, c_t)$

Attention Weight Calculation:
$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{ik})}$

Decoder Output Calculation:
$P(y_t \mid y_{<t}, x) = \text{softmax}(V s_t)$

Cross-Entropy Loss:
$\text{CE}(y, \hat{y}) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$

BLEU Score:
$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$