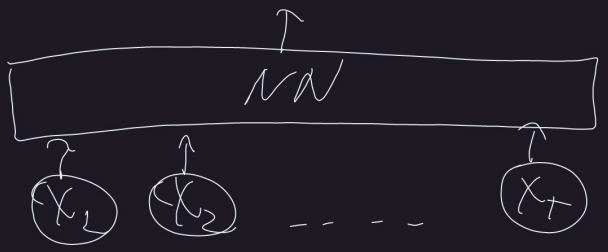
NLP for texts



Stemming

models
modelling
modelled
---

stemming → model

lem("modelling", pos="v") = "model"

lem("modelling", pos="s") = "modelling"

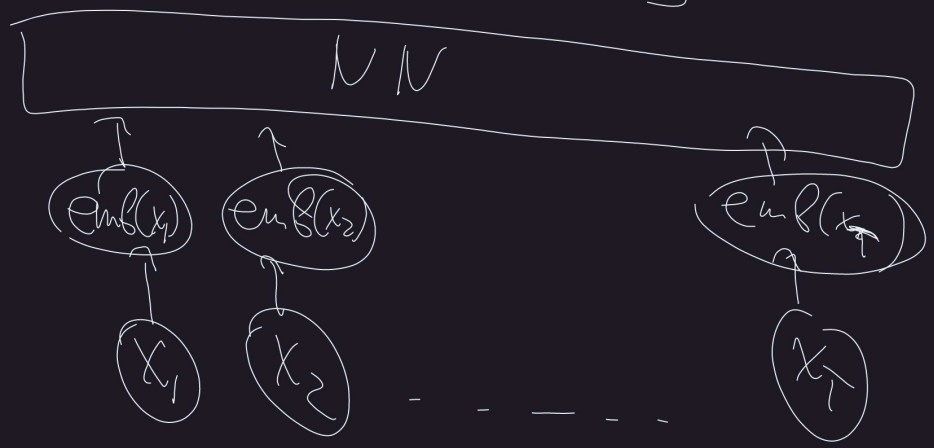
Lemmatization

"went" → "go"
"better" → "good"
"corpora" → "corpus"

stop words: "the", "a", "an", "and", "in"

pipeline: words → stemming / lemmatization → remove least freq. tokens → remove stop words

$x_t \in \{1, 2, \ldots, V\} \rightarrow emb(x_t) \in \mathbb{R}^{emb\_dim}$



<UNK>, <PAD>, ...

(byte pair encoding)

BPE tokenizer

Corpus

"hug" – 10
"pug" – 5
"pun" – 12
"bun" – 4
"hugs" – 5

Vocab

"h", "u", "g",
"p", "n", "b",
"s"

All pairs

"hu" – 15
"ug" – 20
"ug" – 0
...

Vocab.

"h", "u", "g",
"p", "n", "b",
"s", "hu"
...
"gs", "hug"

inference: "hugs" → ("hu", "gs")

"hu", "g", "s" → "hu", "gs"...

WordPiece tokenizer

"word" → "w", "##o", "##r", "##d"

$$score(token1, token2) = \frac{freq(token1, token2)}{freq(token1) \cdot freq(token)}$$

"un", "##able"

"##o", "##r" → "##or"

"bre", "##men" → "bremen"

"w", "##o" → "wo"

Unigram tokenizer

"hug" - 10 → p_2 "h", "u", "g", "p", "n", "b", "s", "hu", "ug",

"pug" - 5 p_2 "pu", "un", "bu", "hug", "ugs"

"pun" - 12 p_3

"bun" - 4 p_4 "hug" → p("hu", "g") = p("hu") · p("g") =

"hugs" - 5 p_1 = $\frac{15}{210} \cdot \frac{20}{216}$ inference:

loss = $-(\log p_1)10 - (\log p_2)5 \ldots$ "hugs" → "hug", "s"

teacher forcing

exposure bias fix

$$p(y_1, \ldots, y_T | x, \theta) = p(y_1 | x, \theta) \cdot p(y_2 | y_1, x, \theta) \cdots =$$

$$= \prod_{t=1}^{T} p(y_t | y_1 \cdots y_{t-1}, x, \theta)$$

beam search

$$\log p(y_1 \dots y_t | x, \theta) + \lambda \log p_{LM}(y_1 \dots y_t | \theta) + \beta t$$

$$\underset{y}{\operatorname{argmax}}\, p(y | x, \theta) = \underset{y}{\operatorname{argmax}} \prod_{t=1}^{T} p(y_t | y_1 \dots y_{t-1}, x, \theta)$$

T, V $\quad O(V^T)$

K - beam size

$O(T \cdot K \cdot V)$

t=1

top-2 (
- "a" − 0.5
- "b" − 0.4
- "c" − 0.1
)

t=2

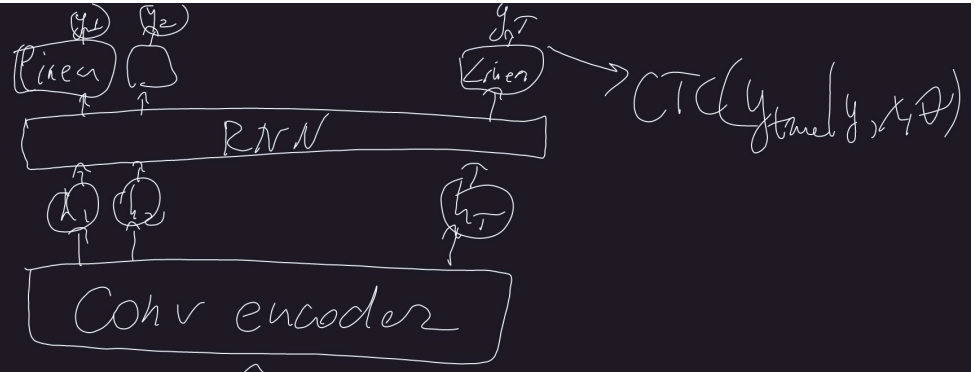
"aa"	0.5 · 0.5
"ab"	0.5 · 0.4
"ac"	0.5 · 0.1
"ba"	0.4 · 0.45
"bb"	0.4 · 0.45
"bc"	0.4 · 0.1

top-2

t=3

- "aaa"
- "aab"
- "aac"
- "aba"
- "abb"
- "abc"

CTC loss



handwritten text rec.

image

→ "h", "e", "l", "l", "o"

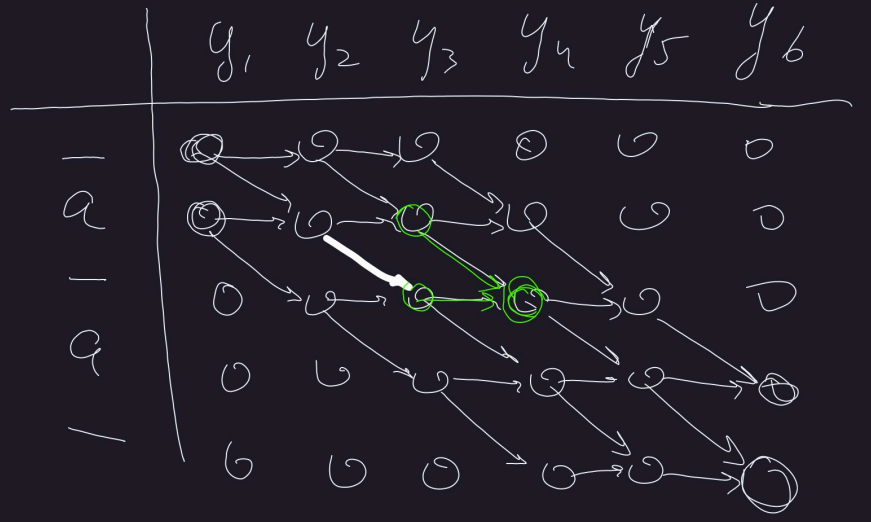
CTC-loss:

$$p(y_{true}|x,\theta) = \sum_{y:\,ctc(y)=y_{true}} \prod_{t=1} p_t(y_t|x,\theta)$$

"hello"

$$CTC(\_\_h\_ee\_l\_l\_o\_) = \_h\_e\_l\_l\_o\_ \rightarrow hello$$

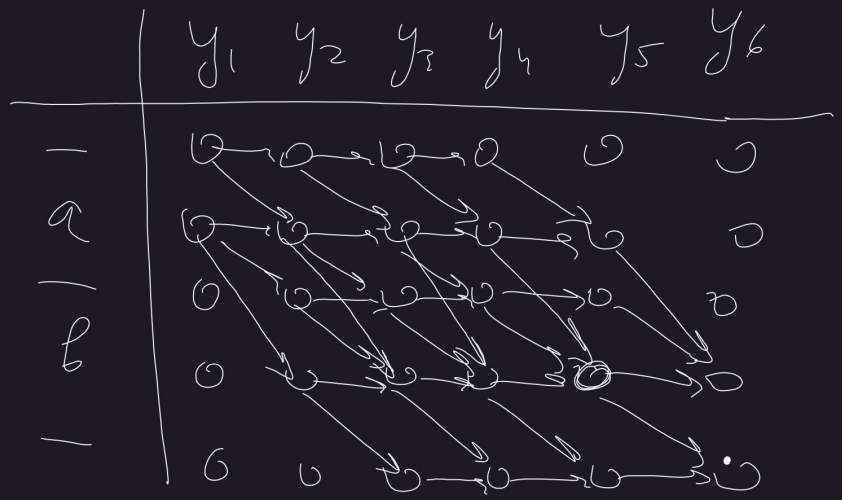
$y_{true} = "aa"$, $T = 6$



$\Gamma(-\,-\,-\,a\,-\,a) = aa$

$\alpha_{s,t} = p_t(y_t \mid x, \theta)\left(\alpha_{s,t-1} + \alpha_{s-1,t-1}\right)$

y_{true} = "a b", $T = 6$



T_{true}

$O(T_{true} \cdot T)$

$$\alpha_{s,t} = \left(\alpha_{s,t-1} + \alpha_{s-1,t-1} + \alpha_{s-2,t-1}\right) p_t(y_t \mid x_t, \theta)$$

if $y_{true,s} \neq y_{true,s-2}$ and $y_{true,s} \neq$ –

$$\alpha_{s,t} = \left(\alpha_{s,t-1} + \alpha_{s-1,t-1}\right) p_t(y_t \mid x_t, \theta)$$

otherwise

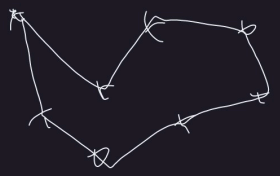
$$CTC = \alpha_{4,6} + \alpha_{5,6}$$

Non-diff. losses

$BLEU(y, y_{true})$ — machine translation

$CIDEr(y, y_{true})$ — image2caption

$\sum_i \rho(x_{y(i)}, x_{y(i+1)})$ — TSP



$$F(\theta) = \mathop{\mathbb{E}}_{p(y|x,\theta)} f(y) \to \min_{\theta}$$

(f — non-diff.)

$$\nabla_\theta F(\theta) = \nabla_\theta \mathop{\mathbb{E}}_{p(y|x,\theta)} f(y) = \nabla_\theta \sum_y p(y|x,\theta) f(y) =$$

$$= \sum_y \nabla_\theta p(y|x,\theta) \cdot f(y) = \sum_y p(y|x,\theta) \cdot \frac{\nabla_\theta p(y|x,\theta)}{p(y|x,\theta)} f(y) =$$

$$= \mathop{\mathbb{E}}_{p(y|x,\theta)} \nabla_\theta \log p(y|x,\theta) f(y) \approx \nabla_\theta \log p(y|x,\theta)$$

$$\approx \frac{1}{M} \sum_{j=1}^{M} \nabla_\theta \log p(y_j|x,\theta) f(y_j), \quad y_j \overset{i.i.d.}{\sim} p(y|x,\theta)$$

Good choices for $b(x)$:

1. $b(x) = \mathbb{E}_{p(y|x,\theta)} f(y)$
2. $b(x) = f(y_{\text{greedy rollout}})$

baseline $b(x)$

$$\nabla_\theta F(\theta) = \mathbb{E}_{p(y|x,\theta)} \nabla_\theta \log p(y|x,\theta)\big(f(y) - b(x)\big)$$

$$\mathbb{E}_{p(y|x,\theta)} \nabla_\theta \log p(y|x,\theta)\, b(x) = \sum_y p(y|x,\theta) \frac{\nabla_\theta p(y|x,\theta)}{p(y|x,\theta)} b(x) =$$

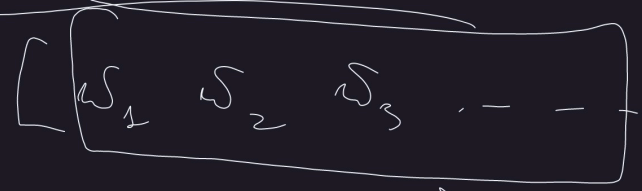
$$= b(x) \sum_y \nabla_\theta p(y|x,\theta) = b(x) \nabla_\theta \underbrace{\sum_y p(y|x,\theta)}_{=1} = 0$$

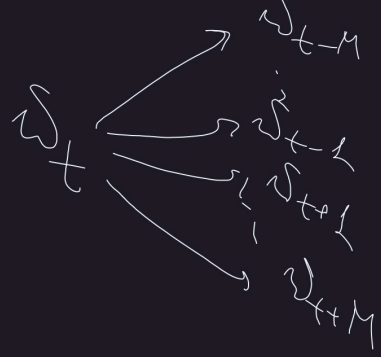
REINFORCE / log-derivative trick

Word Embedding

$x_t \in \{1, 2, \ldots, V\} \rightarrow emb(x_t) \in \mathbb{R}^{emb\_dim}$

Word2Vec





$$\frac{\exp(emb(w_c)^T emb(w_t))}{\sum_{j=1} \exp(emb(w_j)^T emb(w_t))}$$

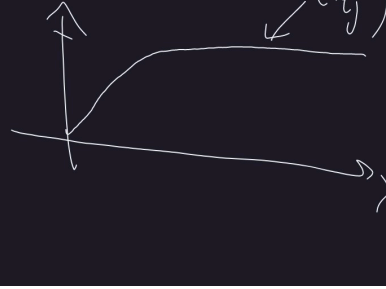
$$\sum_t \sum_{c \in C_t} \log p(w_c \mid w_t) \rightarrow \max_{emb}$$

negative sampling:

$$\sum_t \left[\sum_{c \in C_t} \log \sigma(emb(w_c)^T emb(w_t)) + \sum_{n \in N_{t,c}} \log \sigma(-emb(w_n)^T emb(w_t)) \right] \to \max_{emb}$$

fast-text: $score(w_c, w_t) = emb(w_c)^T \sum_{w \in n\text{-}grams(w_t)} emb(w)$

GloVe $\sum_{i,j} f(X_{ij}) \left(emb(x_i)^T emb(x_j) + b_i + b_j - \log X_{ij} \right)^2 \to \min_{emb}$



CoVe

ELMo

BERT