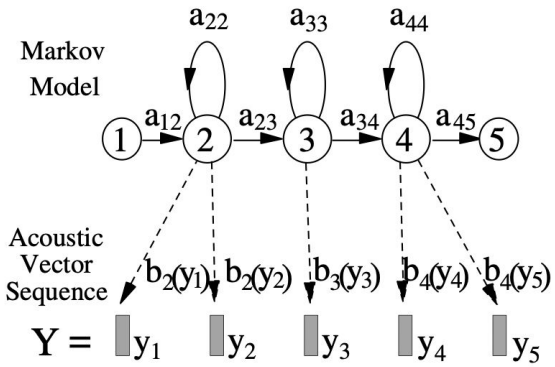
Automatic speech recognition

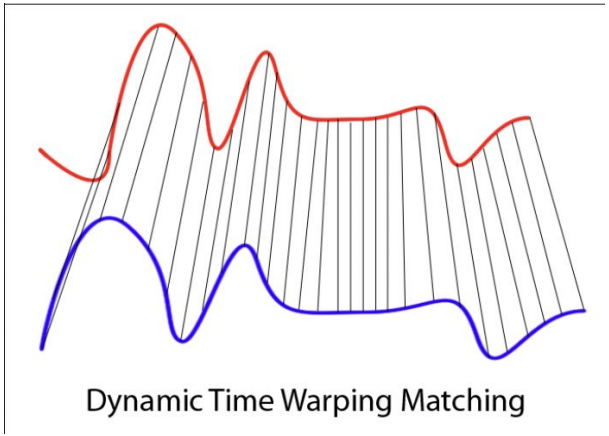
Теплова Анна, 202

До нейронных сетей:

- Hidden Markov Models



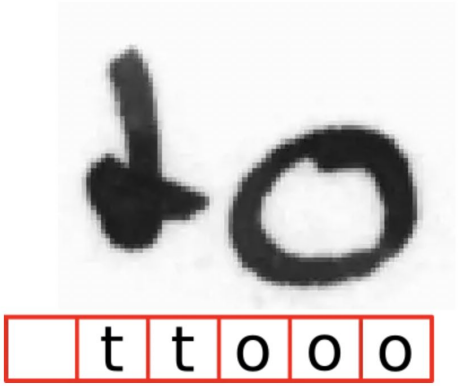
- Dynamic Time Warping



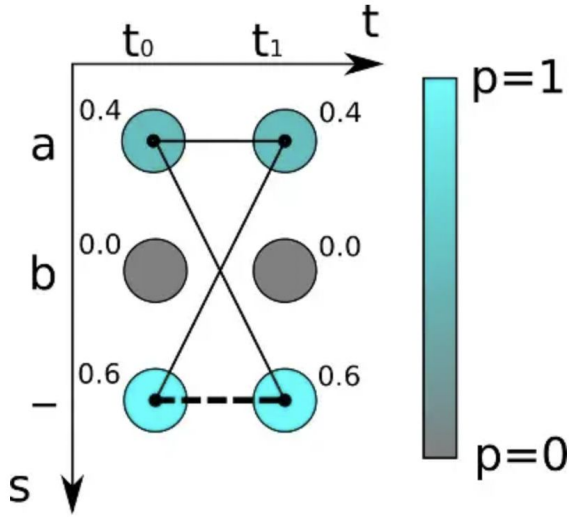
неустойчиво к шумам, другим языкам, требует огромных сил на разметку данных, необходимы изначальные специфичные знания и предположения

Loss

Нарезка слова:



Выход сети:



- “aa”: 0.4·0.4=0.16
- “a-”: 0.4·0.6=0.24
- “-a”: 0.6·0.4=0.24

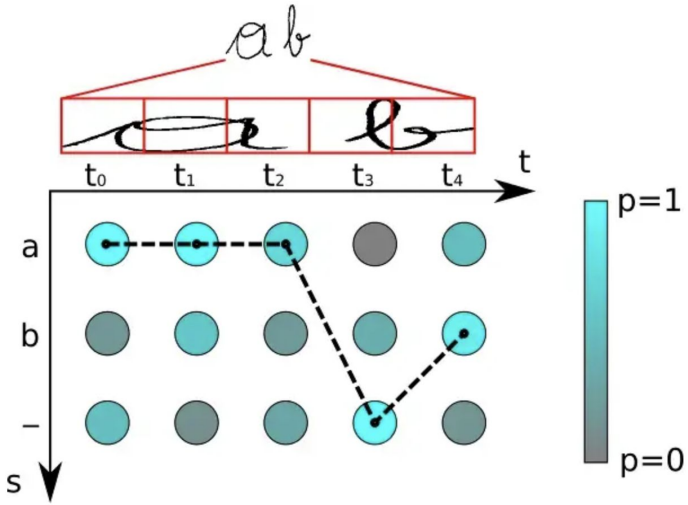
Ответ для GT = “a”: 0.64

Убираем дубликаты:

- “to” → “---ttttttooo”/“-t-o-”/“to”
- “too” → “---ttttto-o”/“-t-o-o-”/“to-o”, **не “too”**

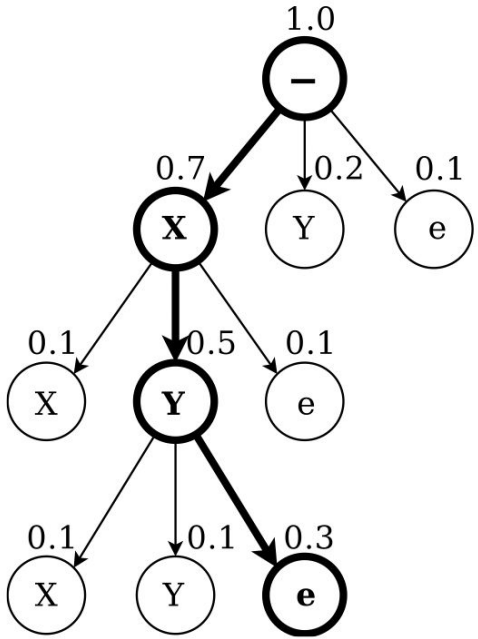
Преобразование над такой матрицей - свертка. Градиент берется по кросс-энтропии

А как на тесте? (наивно)



Декодирование происходит в единицу времени, поэтому сети не нужно понятие “морфема”. Но есть проблема - на предыдущем примере ответом было бы “”

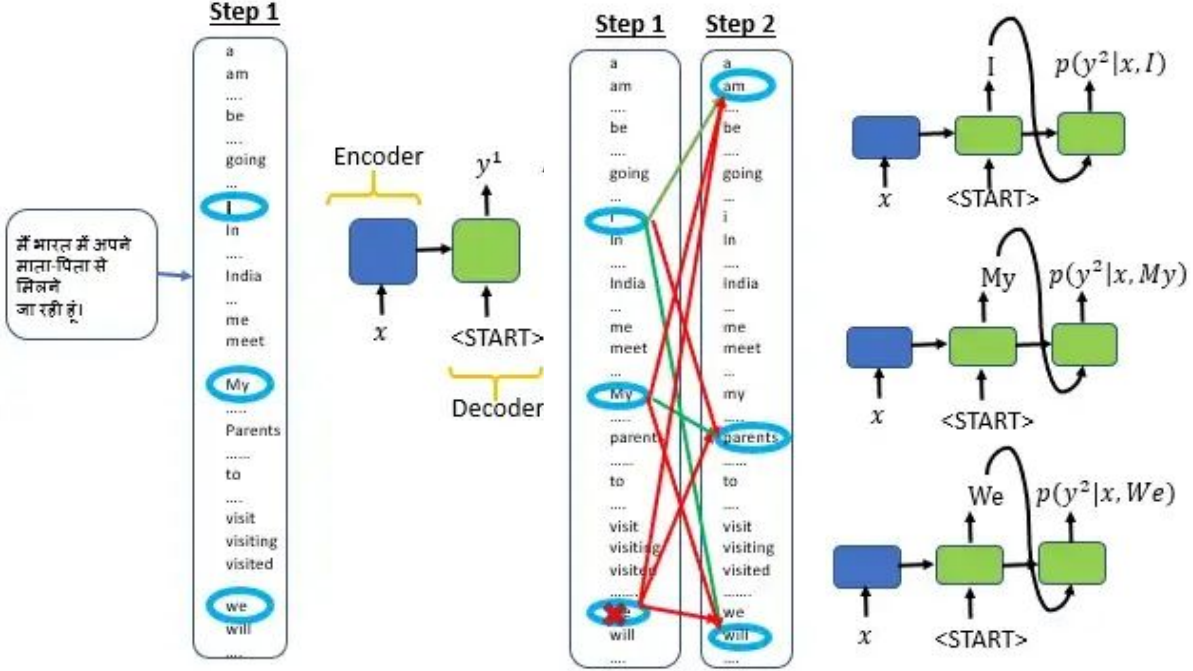
Префиксный алгоритм



Beam search


IF MY AGENT CALLS, TELL HIM HE'S NOT GETTING TEN PERCENT OF MY NEW JOB!

Beam width=3



$P(y^1|x)$ = [I, My, We]

$P(y^1, y^2|x) = p(y^1|x) * p(y^2|x, y^1)$

$P(y^1|x)$ = [I, My, We]

$P(y^1, y^2|x)$= [I am, I will, My parents]

RNN output	Decoded Transcription
what is the weather like in bostin right now	what is the weather like in boston right now
prime miniter nerenr modi	prime minister narendra modi
arther n tickets for the game	are there any tickets for the game

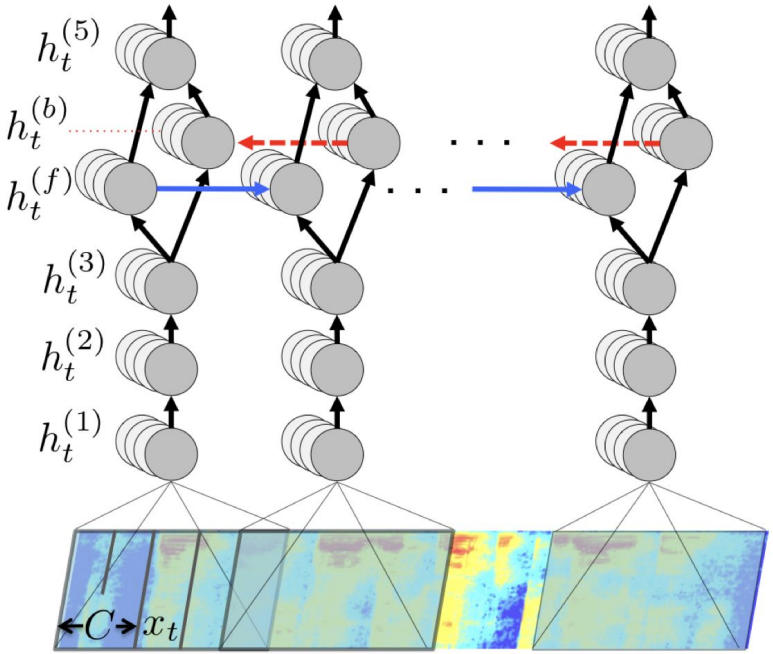
$$Q(c) = \log(\mathbb{P}(c|x)) + \alpha \log(\mathbb{P}_{\text{lm}}(c)) + \beta \, \text{word\_count}(c)$$

максимизируем

LM используется и для подсчета вероятностей последовательности токенов, и для исправления ошибок в выдаваемых токенах

Собственно, Deep Speech

Структура сети



Вход: $x_t^{(i)}, t = 1, \ldots, T^{(i)}$

Первые три слоя:

$$h_t^{(l)} = g(W^{(l)} h_t^{(l-1)} + b^{(l)})$$

Четвертый слой:

$$h_t^{(f)} = g(W^{(4)} h_t^{(3)} + W_r^{(f)} h_{t-1}^{(f)} + b^{(4)})$$

$$h_t^{(b)} = g(W^{(4)} h_t^{(3)} + W_r^{(b)} h_{t+1}^{(b)} + b^{(4)})$$

$$h_t^{(4)} = h_t^{(f)} + h_t^{(b)}$$

Пятый слой:

$$h_t^{(5)} = g\left(W^{(5)} h_t^{(4)} + b^{(5)}\right)$$

Структура сети:

Шестой слой: $$h_{t,k}^{(6)} = \hat{y}_{t,k} \equiv \mathbb{P}(c_t = k|x) = \frac{\exp(W_k^{(6)} h_t^{(5)} + b_k^{(6)})}{\sum_j \exp(W_j^{(6)} h_t^{(5)} + b_j^{(6)})}.$$

Активация: $$g(z) = \min\{\max\{0, z\}, 20\}$$

Optimizer: Nesterov gradient method

Регуляризация: Dropout 5-10%

System	Clean (94)	Noisy (82)	Combined (176)
Apple Dictation	14.24	43.76	26.73
Bing Speech	11.73	36.12	22.05
Google API	6.64	30.47	16.72
wit.ai	7.94	35.06	19.41
Deep Speech	**6.56**	**19.06**	**11.85**

Model	SWB	CH	Full
Vesely et al. (GMM-HMM BMMI) [44]	18.6	33.0	25.8
Vesely et al. (DNN-HMM sMBR) [44]	12.6	24.1	18.4
Maas et al. (DNN-HMM SWB) [28]	14.6	26.3	20.5
Maas et al. (DNN-HMM FSH) [28]	16.0	23.7	19.9
Seide et al. (CD-DNN) [39]	16.1	n/a	n/a
Kingsbury et al. (DNN-HMM sMBR HF) [22]	13.3	n/a	n/a
Sainath et al. (CNN-HMM) [36]	11.5	n/a	n/a
Soltau et al. (MLP/CNN+I-Vector) [40]	**10.4**	n/a	n/a
Deep Speech SWB	20.0	31.8	25.9
Deep Speech SWB + FSH	12.6	**19.3**	**16.0**

Текущий прогресс

- Google, Microsoft, Bing, Apple - закрытый ресерч, но есть API
- Vosk, CMUSphinx - mobile
- KoNLPy
- TensorFlow попытался
- До сих пор используются марковские цепи

Общий слайд

Мы поговорили:

- Что было до RNN
- CTC Loss
- Beam search
- Deep Speech
- Текущий прогресс