RELEVANT LINKS

DATE
24/03/2025

CS217: ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Introduction to Speech Recognition

presented by: Darshan Prabhu

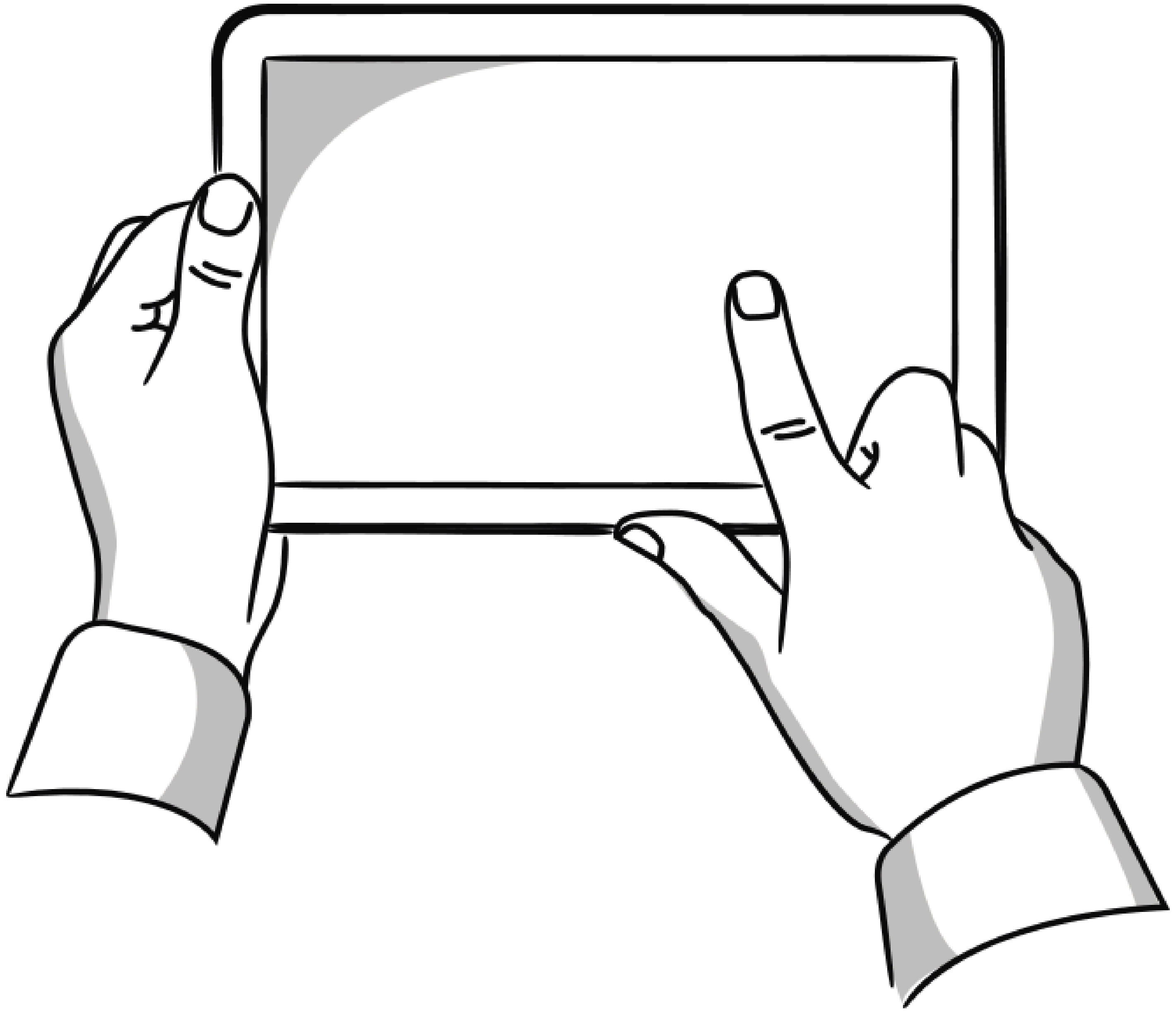
CSALT Research
Indian Institute of Technology Bombay

RELEVANT LINKS

DATE
24/03/2025

Chapter 1:
What is Automatic Speech Recognition?

Definition, Challenges, History and Evaluation Metrics



RELEVANT LINKS

DATE
24/03/2025

What is ASR ?

Reference: "How Alexa and Siri work !", https://bridgethedigitaldivide.us/blog/how-alexa-and-siri-work/.

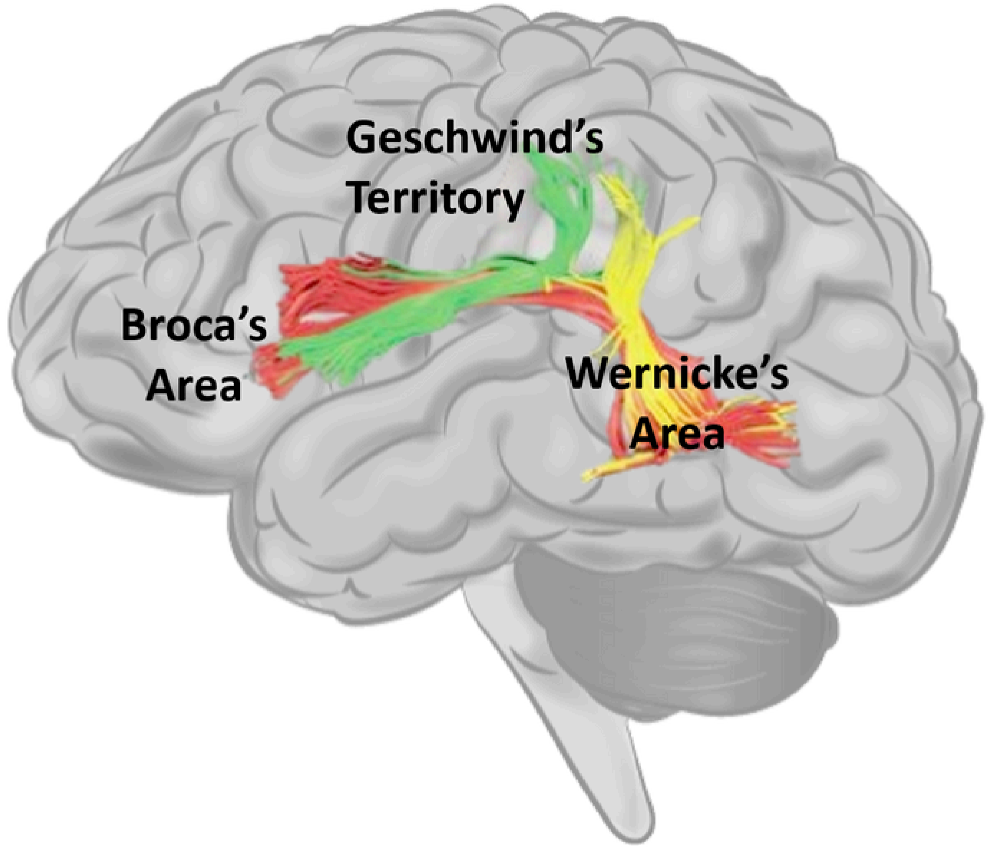
What is ASR ?

Reference: "How Alexa and Siri work !", https://bridgethedigitaldivide.us/blog/how-alexa-and-siri-work/.

RELEVANT LINKS

DATE
24/03/2025

What is ASR ?



Wernicke's area, located in the posterior segment of the superior temporal gyrus

Reference: "How Alexa and Siri work !", https://bridgethedigitaldivide.us/blog/how-alexa-and-siri-work/.

RELEVANT LINKS

DATE
24/03/2025

What is ASR ?

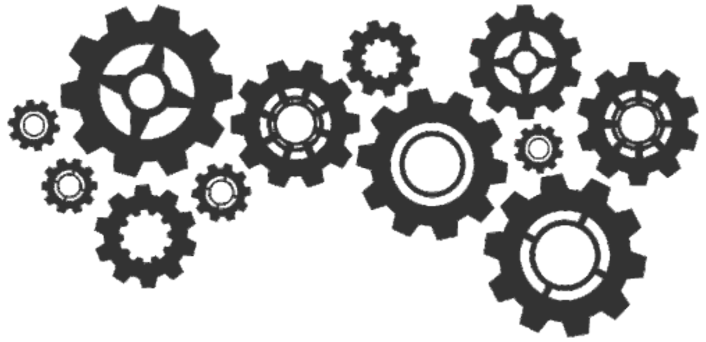
- A task of automatically converting the speech signal into words.

RELEVANT LINKS

DATE
24/03/2025

What is ASR ?

- A task of automatically converting the speech signal into words.



WHY SO SERIOUS?

What is ASR ?

- A task of automatically converting the speech signal into words.
- The recognized words can be
  - the final output, or
  - the input to Natural Language Processing

RELEVANT LINKS

DATE
24/03/2025

What is ASR ?

- A task of automatically converting the speech signal into words.
- The recognized words can be
    - the final output, or
    - the input to Natural Language Processing
- Downstream applications of ASR
    - Spoken language understanding
    - Spoken translation
    - Intelligent video editing
    - ASR from brain signals
    - ASR for speakers with speech pathologies

RELEVANT LINKS

DATE
24/03/2025

Main Challenges

RELEVANT LINKS

DATE
24/03/2025

Main Challenges

1. Speaker's Influence: Accent or Dialect variations, Non-native speakers, Disfluencies etc

RELEVANT LINKS

DATE
24/03/2025

Main Challenges

1. Speaker's Influence: Accent or Dialect variations, Non-native speakers, Disfluencies etc

2. Environmental factors: Background noise, Co-articulation, Reverberation etc

RELEVANT LINKS

DATE
24/03/2025

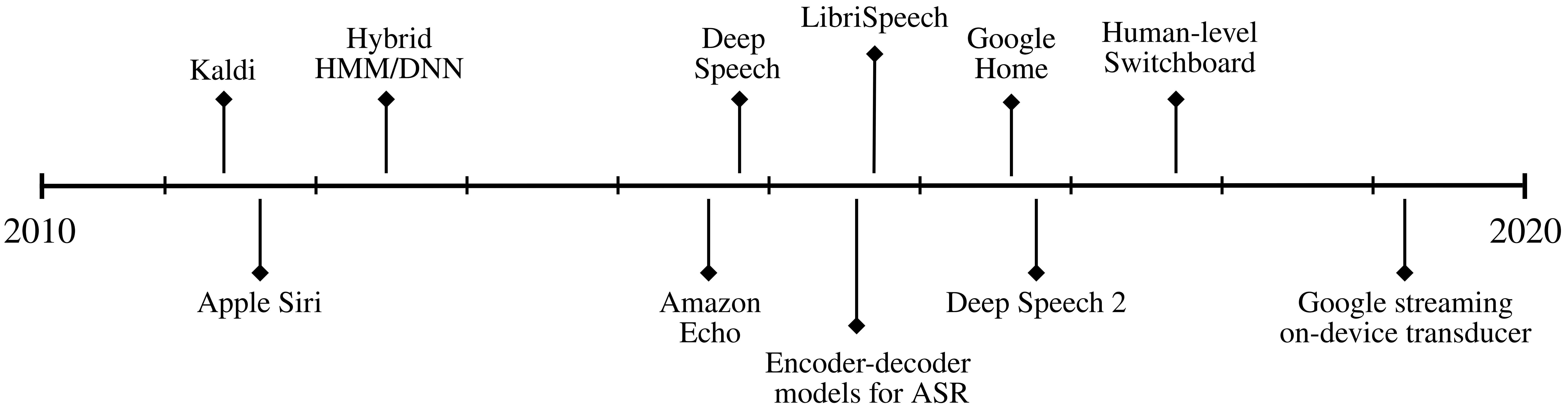
Main Challenges

1. Speaker's Influence: Accent or Dialect variations, Non-native speakers, Disfluencies etc
2. Environmental factors: Background noise, Co-articulation, Reverberation etc
3. Data related problems: Microphone quality, Lack of diversity etc

RELEVANT LINKS

DATE
24/03/2025

A historical perspective of ASR



Reference: “The History of Speech Recognition to the Year 2030” by Awni Hannun,https://awni.github.io/future-speech/.

RELEVANT LINKS

DATE

24/03/2025

A historical perspective of ASR (continued)

📐 The 🤗 Open ASR Leaderboard ranks and evaluates speech recognition models on the Hugging Face Hub.
We report the Average WER (⬇️ lower the better) and RTFx (⬆️ higher the better). Models are ranked based on their Average WER, from lowest to highest. Check the 📈 Metrics tab to understand how the models are evaluated.
If you want results for a model that is not listed here, you can submit a request for it to be included ✉️✨.
The leaderboard currently focuses on English speech recognition, and will be expanded to multilingual evaluation in later versions.

🏅 Leaderboard | 📈 Metrics | ✉️✨ Request a model here!

model	Average WER ⬇️	RTFx ⬆️	AMI	Earnings22	Gigaspeech	LS Clean	LS Other	SPGISpeech	Tedlium
microsoft/Phi-4-multimodal-instruct	6.14	62.12	11.45	10.5	9.77	1.67	3.82	3.11	2.89
nvidia/canary-1b-flash	6.35	1045.75	13.11	12.77	9.85	1.48	2.87	1.95	3.12
nvidia/canary-1b	6.5	235.34	13.9	12.19	10.12	1.48	2.93	2.06	3.56
nyrahealth/CrisperWhisper	6.67	84.05	8.71	12.89	10.24	1.82	4	2.7	3.2
nvidia/parakeet-tdt-1.1b	7.01	2390.61	15.87	14.49	9.52	1.4	2.6	3.16	3.59
nvidia/parakeet-rnnt-1.1b	7.12	2053.15	17.01	13.94	9.89	1.45	2.5	2.93	3.83
nvidia/canary-180m-flash	7.12	1233.58	14.86	12.33	10.51	1.73	4.35	2.26	3.13
efficient-speech/lite-whisper-large-v3-acc	7.23	117.8	16.1	11.04	10.1	2	3.91	2.89	3.71
nvidia/parakeet-ctc-1.1b	7.4	2728.52	15.67	13.75	10.28	1.83	3.51	4.02	3.57
efficient-speech/lite-whisper-large-v3	7.43	115.83	16.9	11.55	10.26	2.1	4.4	2.85	3.73
openai/whisper-large-v3	7.44	145.51	15.95	11.29	10.02	2.01	3.91	2.94	3.86
nvidia/parakeet-tdt_ctc-110m	7.49	5345.14	15.89	12.37	10.52	2.4	5.22	2.54	4.07

Evaluation Metrics

Reference: I want to go to the cse office

Prediction: I want to go see a office

RELEVANT LINKS

DATE
24/03/2025

Evaluation Metrics

Reference: I want to go to the cse office

Prediction: I want to go see a office

Method 1: Sentence Error Rate

An entire sentence is either correct or not.

- 100% error rate in the case above.

RELEVANT LINKS

DATE
24/03/2025

Evaluation Metrics

Reference: I want to go to the cse office

Prediction: I want to go see a office

Method 1: Sentence Error Rate

An entire sentence is either correct or not.

- 100% error rate in the case above.
- Problem: Too strict. Need to consider some measure of local correctness.

RELEVANT LINKS

DATE
24/03/2025

Evaluation Metrics

Reference: I want to go to the cse office

Prediction: I want to go see a *** office

Method 2: Word Error Rate (WER)

insertion errors = 0, # substitution errors = 2, # of deletion errors = 1 → Edit distance = 3

Word Error Rate (%): Edit distance (=3) / # reference words (=8) * 100 = 37.5 %

- Calculated using Levenshtein distance.

RELEVANT LINKS

DATE
24/03/2025

Evaluation Metrics

Reference: I want to go to the cse office

Prediction: I want to go see a *** office

Method 2: Word Error Rate (WER)

insertion errors = 0, # substitution errors = 2, # of deletion errors = 1 → Edit distance = 3

Word Error Rate (%): Edit distance (=3) / # reference words (=8) * 100 = 37.5 %

- Calculated using Levenshtein distance.
- Problem: How to handle languages that do not have word boundaries? (Ex: Japanese)

RELEVANT LINKS

DATE
24/03/2025

Evaluation Metrics

Reference: hello world

Prediction: heldoo world

Method 3: Character Error Rate (CER)

insertion errors = 1, # substitution errors = 1, # of deletion errors = 0 → Edit distance = 2

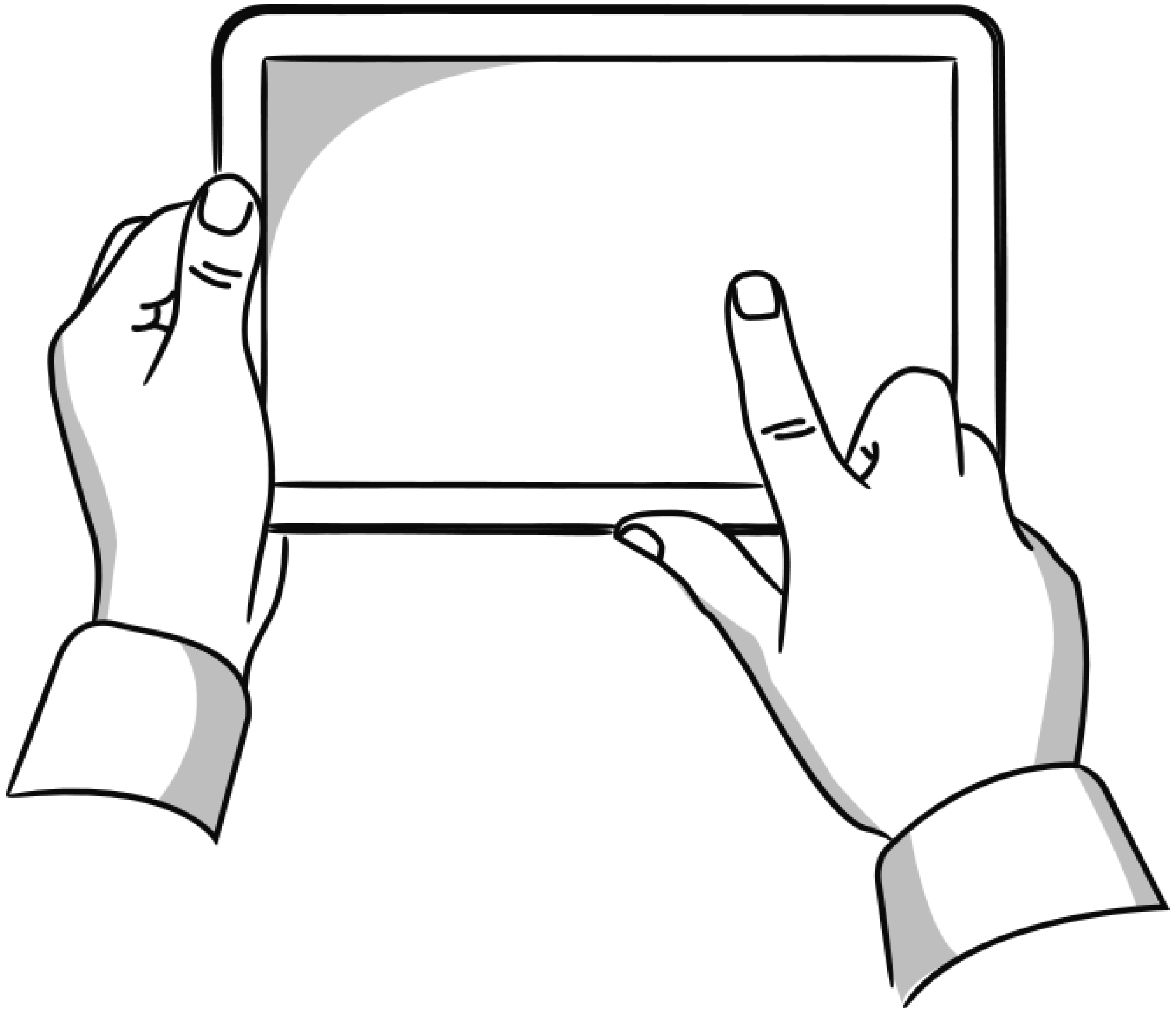
Character Error Rate (%): Edit distance (=2) / # reference chars (=10) * 100 = 20 %

RELEVANT LINKS

DATE
24/03/2025

Chapter 2:
Connectionist Temporal Classification

Background, Problem Setting and Formulations



Reference: "Sequence ModelingWith CTC", https://distill.pub/2017/ctc/.

RELEVANT LINKS

DATE
24/03/2025

Mathematical Formulation of ASR

RELEVANT LINKS

DATE
24/03/2025

Mathematical Formulation of ASR

$$\mathbf{O} = \{O_1, O_2, \ldots, O_T\}$$ → a sequence of **acoustic features** corresponding to a speech signal

$$O_i \in \mathbb{R}^d$$ → **d-dimensional** acoustic feature vector and T is the length of the sequence

RELEVANT LINKS

DATE
24/03/2025

Mathematical Formulation of ASR

$\mathbf{O} = \{O_1, O_2, \ldots, O_T\}$ → a sequence of **acoustic features** corresponding to a speech signal

$O_i \in \mathbb{R}^d$ → **d-dimensional** acoustic feature vector and T is the length of the sequence

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} Pr(\mathbf{W} \mid \mathbf{O})$$

$\mathbf{W} = \{W_1, W_2, \ldots, W_M\}$ → a sequence of **words** and M is the length of this sequence

RELEVANT LINKS

DATE
24/03/2025

Training of ASR

$$\sum_{(\mathbf{O},\mathbf{y})\in\mathcal{D}} -\log Pr(\mathbf{y} \mid \mathbf{O})$$

Training of ASR

$$\sum_{(\mathbf{O},\mathbf{y}) \in \mathcal{D}} - \log Pr(\mathbf{y} \mid \mathbf{O})$$

Dataset

RELEVANT LINKS

DATE
24/03/2025

Training of ASR

$$\sum_{(\mathbf{O},\mathbf{y})\in\mathcal{D}} -\log Pr(\mathbf{y} \mid \mathbf{O})$$

Dataset

Input sequence

Output sequence

RELEVANT LINKS

DATE
24/03/2025

Training of ASR

$$\sum_{(\mathbf{O},\mathbf{y})\in\mathcal{D}} -\log Pr(\mathbf{y} \mid \mathbf{O})$$

Input sequence

Output sequence

Dataset

O and y are different length sequences.

How do we handle it?

Alignment, Blank and Beta Operator

What is an alignment?

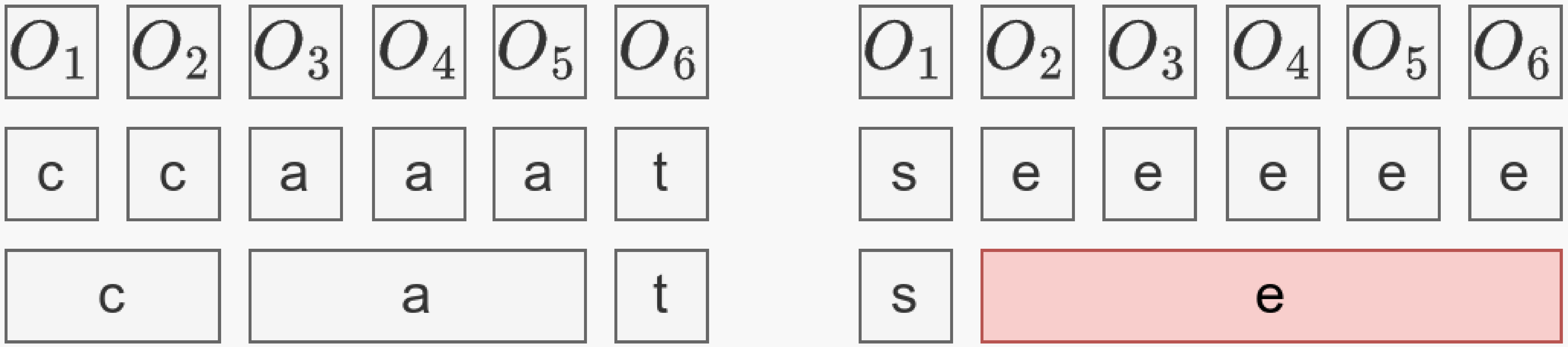
<table>
<tr><td>O_1</td><td>O_2</td><td>O_3</td><td>O_4</td><td>O_5</td><td>O_6</td></tr>
<tr><td>c</td><td>c</td><td>a</td><td>a</td><td>a</td><td>t</td></tr>
<tr><td colspan="2">c</td><td colspan="3">a</td><td>t</td></tr>
</table>

RELEVANT LINKS

DATE
24/03/2025

Alignment, Blank and Beta Operator

What is an alignment?

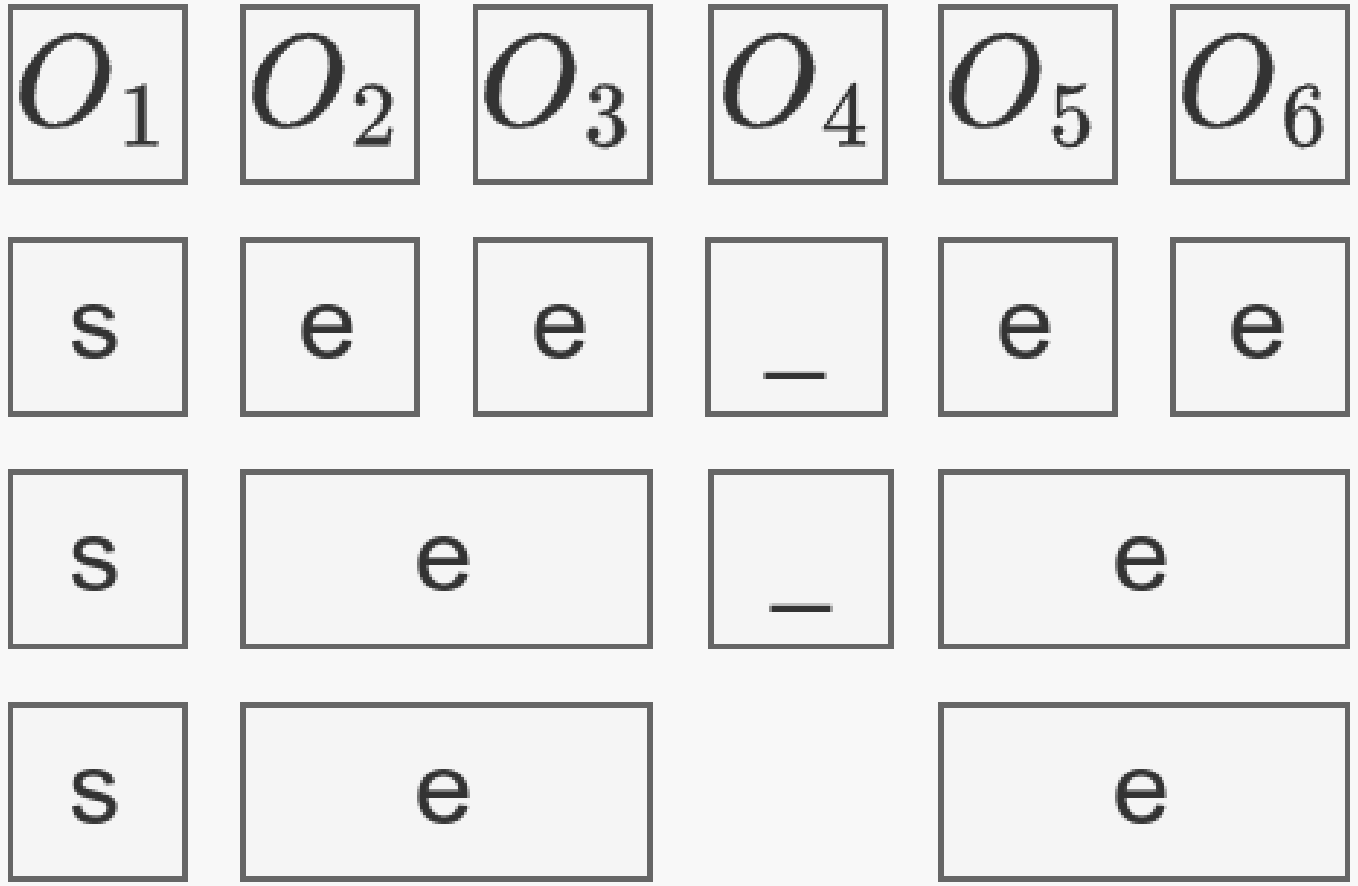


RELEVANT LINKS

DATE
24/03/2025

Alignment, Blank and Beta Operator

How to handle repetitions in alignments?



RELEVANT LINKS

DATE
24/03/2025

Alignment, Blank and Beta Operator

- **Blank symbol (_):** Added to the vocabulary. It represents "empty".
- For a given label sequence, there can be multiple alignments: (x, y, z) could correspond to (x, _, y, _, _, z) or (_, x, x, _, y, z)

RELEVANT LINKS

DATE
24/03/2025

Alignment, Blank and Beta Operator

- **Blank symbol (_):** Added to the vocabulary. It represents "empty".
- For a given label sequence, there can be multiple alignments: (x, y, z) could correspond to (x, _, y, _, _, z) or (_, x, x, _, y, z)
- Define a **2-step operator** $\mathcal{B}$ that reduces a label sequence by: *first* removing repeating labels and *second* removing blanks.
  - $\mathcal{B}$("x, _, y, _, _, z") = $\mathcal{B}$("_, x, x, _, y, z") = "x, y, z"

Alignment, Blank and Beta Operator

- **Blank symbol (_):** Added to the vocabulary. It represents "empty".
- For a given label sequence, there can be multiple alignments: (x, y, z) could correspond to (x, _, y, _, _, z) or (_, x, x, _, y, z)
- Define a **2-step operator** $\mathcal{B}$ that reduces a label sequence by: *first* removing repeating labels and *second* removing blanks.
  - $\mathcal{B}$("x, _, y, _, _, z") = $\mathcal{B}$("_, x, x, _, y, z") = "x, y, z"
- $\mathcal{B}^{-1}$("x, y, z") = {"x, _, y, _, _, z", "_, x, x, _, y, z", ...} is the set of all T-length alignments that collapse to the string "x, y, z" on applying the operator $\mathcal{B}$

RELEVANT LINKS

DATE
24/03/2025

Rethinking the Probability

$$Pr(\mathbf{y} \mid \mathbf{O}) = \sum_{a \in \mathcal{B}^{-1}(\mathbf{y})} Pr(a \mid \mathbf{O})$$

RELEVANT LINKS

DATE
24/03/2025

Doubt

Rethinking the Probability

$$Pr(\mathbf{y} \mid \mathbf{O}) = \sum_{a \in \mathcal{B}^{-1}(\mathbf{y})} Pr(a \mid \mathbf{O})$$

$$= \sum_{a \in \mathcal{B}^{-1}(\mathbf{y})} \prod_{t=1}^{T} Pr(a_t \mid \mathbf{O})$$

conditional independence

RELEVANT LINKS

DATE
24/03/2025

Rethinking the Probability

$$
\begin{aligned}
Pr(\mathbf{y} \mid \mathbf{O}) &= \sum_{a \in \mathcal{B}^{-1}(\mathbf{y})} Pr(a \mid \mathbf{O}) \\
&= \underbrace{\sum_{a \in \mathcal{B}^{-1}(\mathbf{y})}}_{\text{marginalizes over valid alignments}} \underbrace{\prod_{t=1}^{T} Pr(a_t \mid \mathbf{O})}_{\text{computing the probability for a single alignment step-by-step.}}
\end{aligned}
$$

conditional independence

RELEVANT LINKS

DATE
24/03/2025

Dynamic Programming

0	$\varnothing$
1	s
2	$\varnothing$
3	e
4	$\varnothing$
5	e
6	$\varnothing$

CTC
states

Dynamic Programming

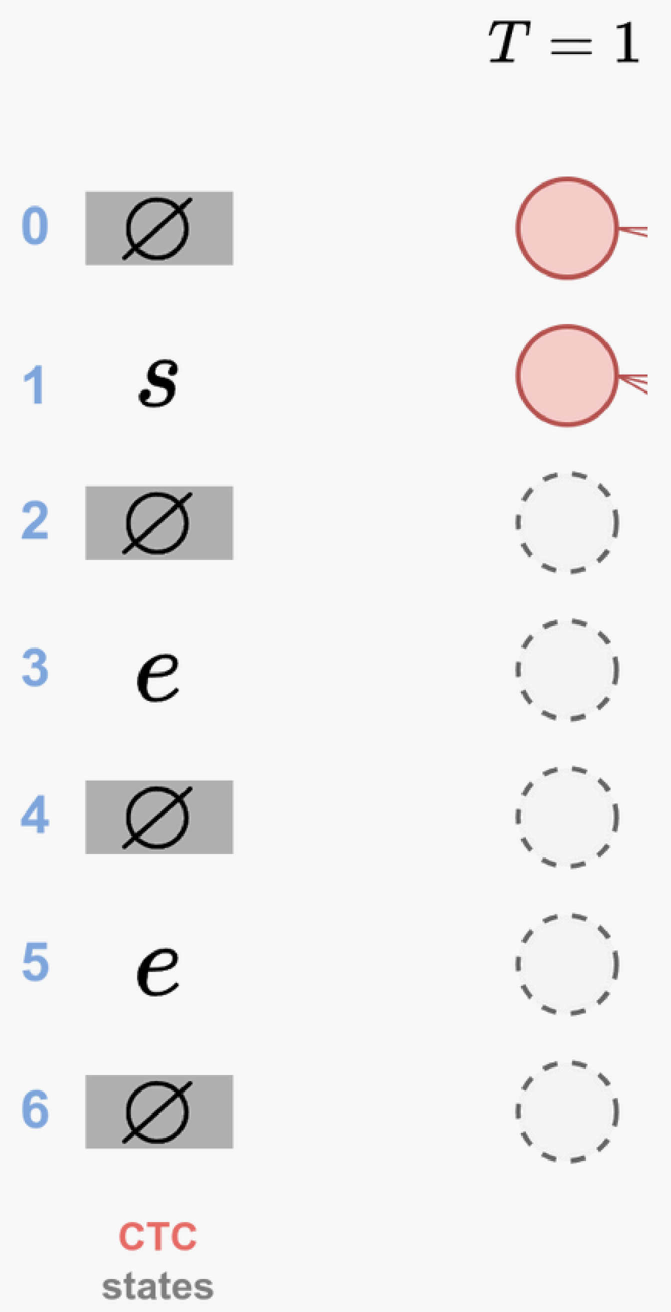
CTC
states

If the orignal label sequence is of length l, then the new sequence would have 2l+1 labels

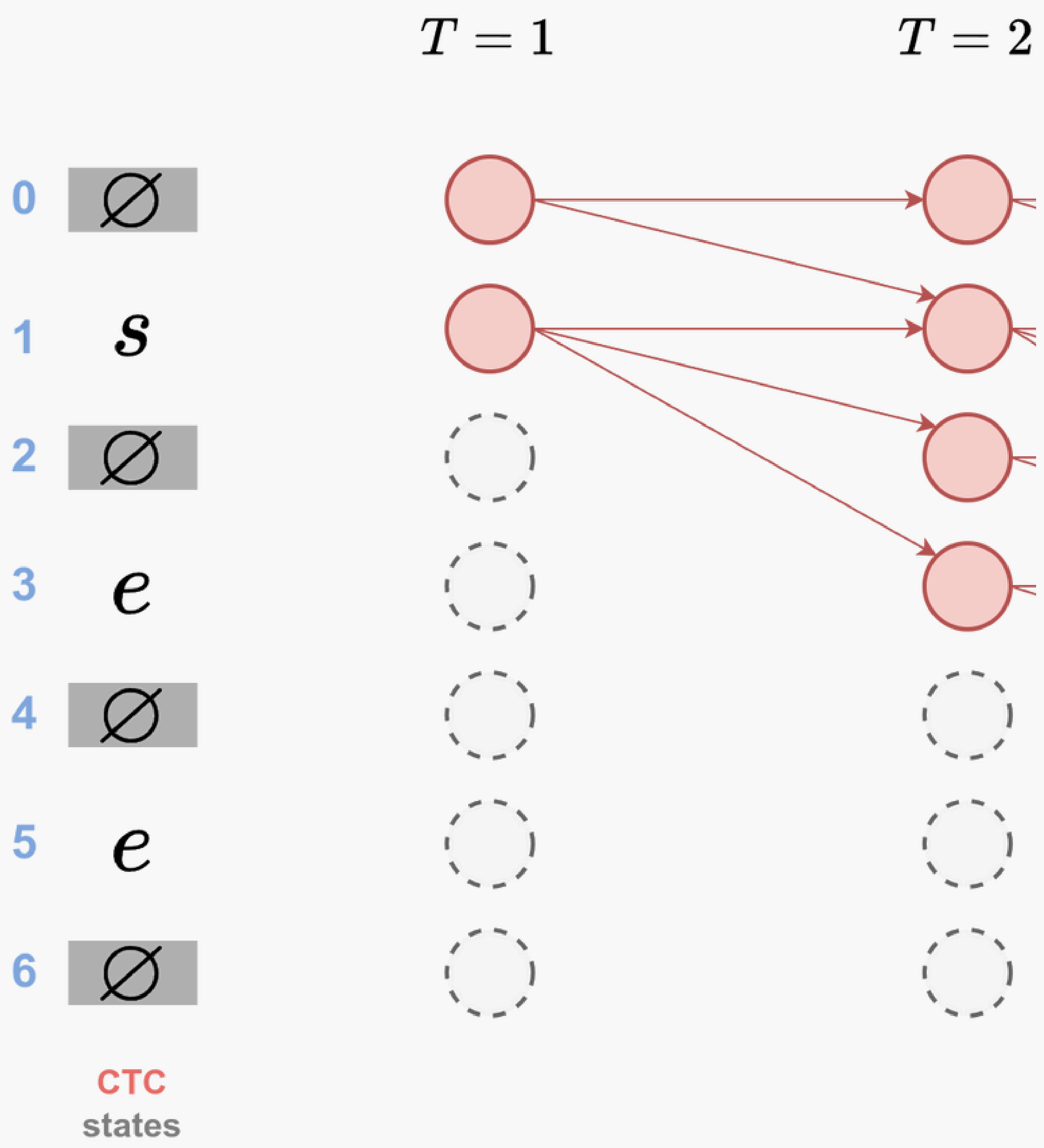
RELEVANT LINKS

DATE
24/03/2025

Dynamic Programming



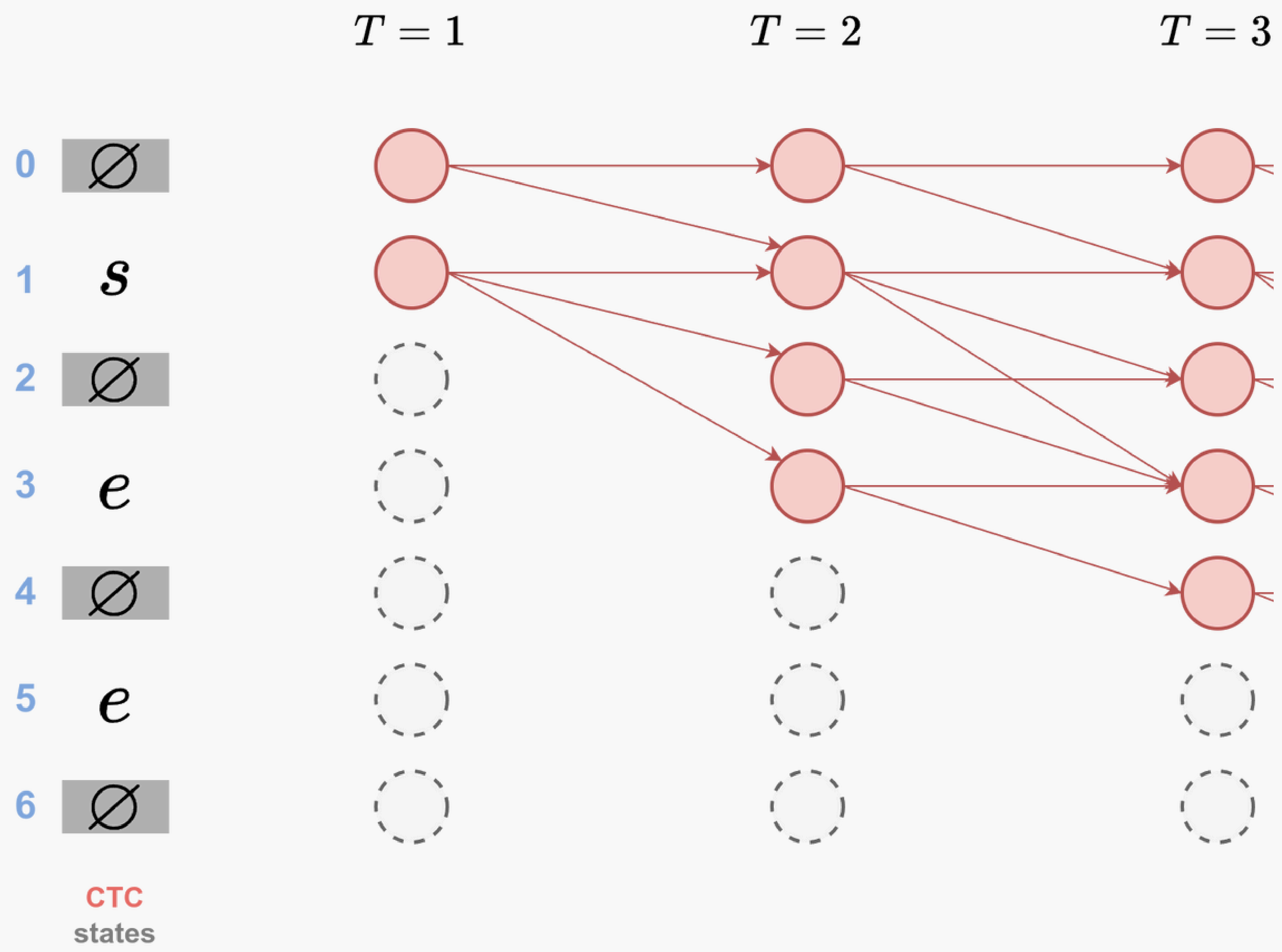
Dynamic Programming



RELEVANT LINKS

DATE
24/03/2025

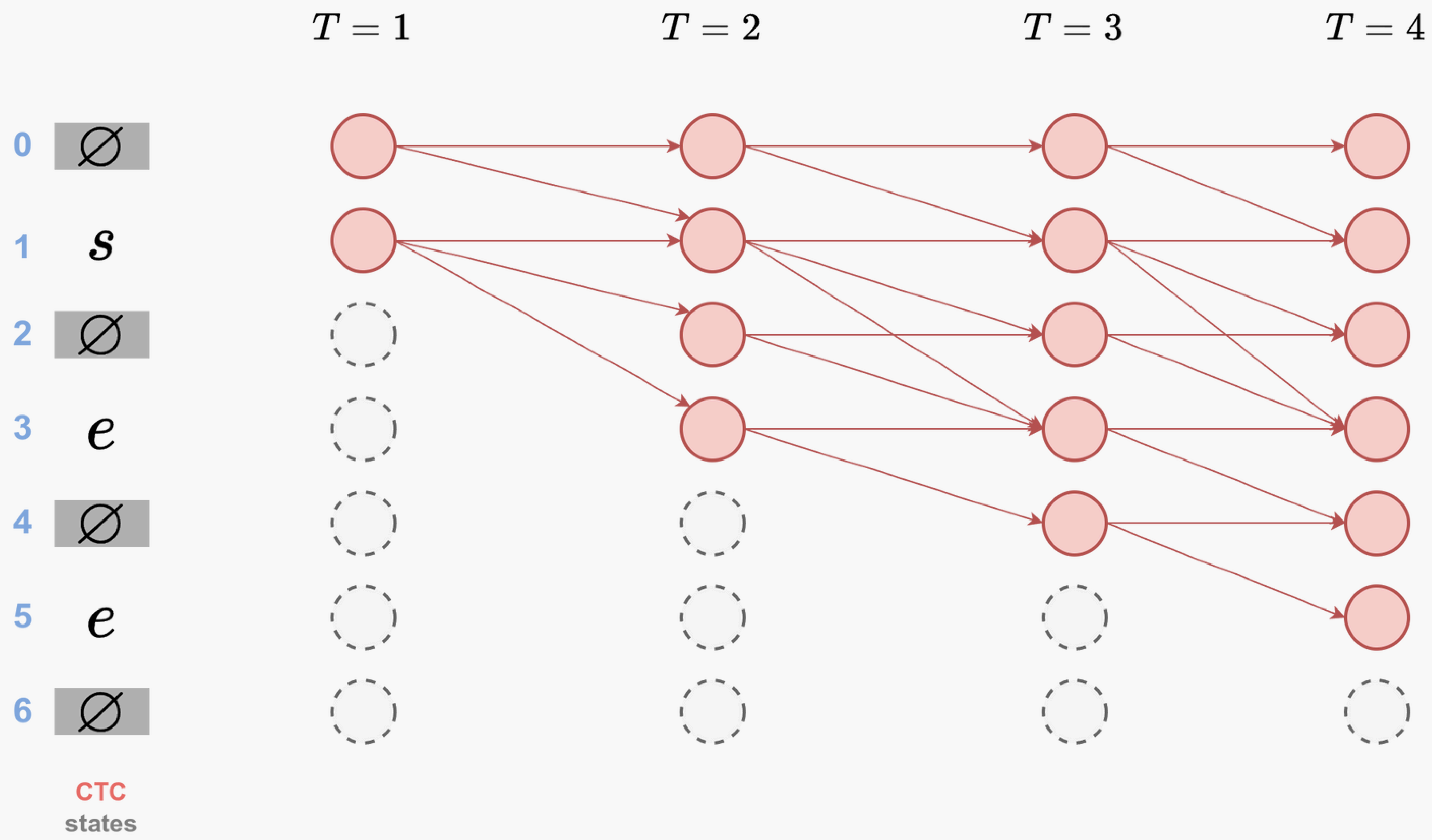
Dynamic Programming



RELEVANT LINKS

DATE
24/03/2025

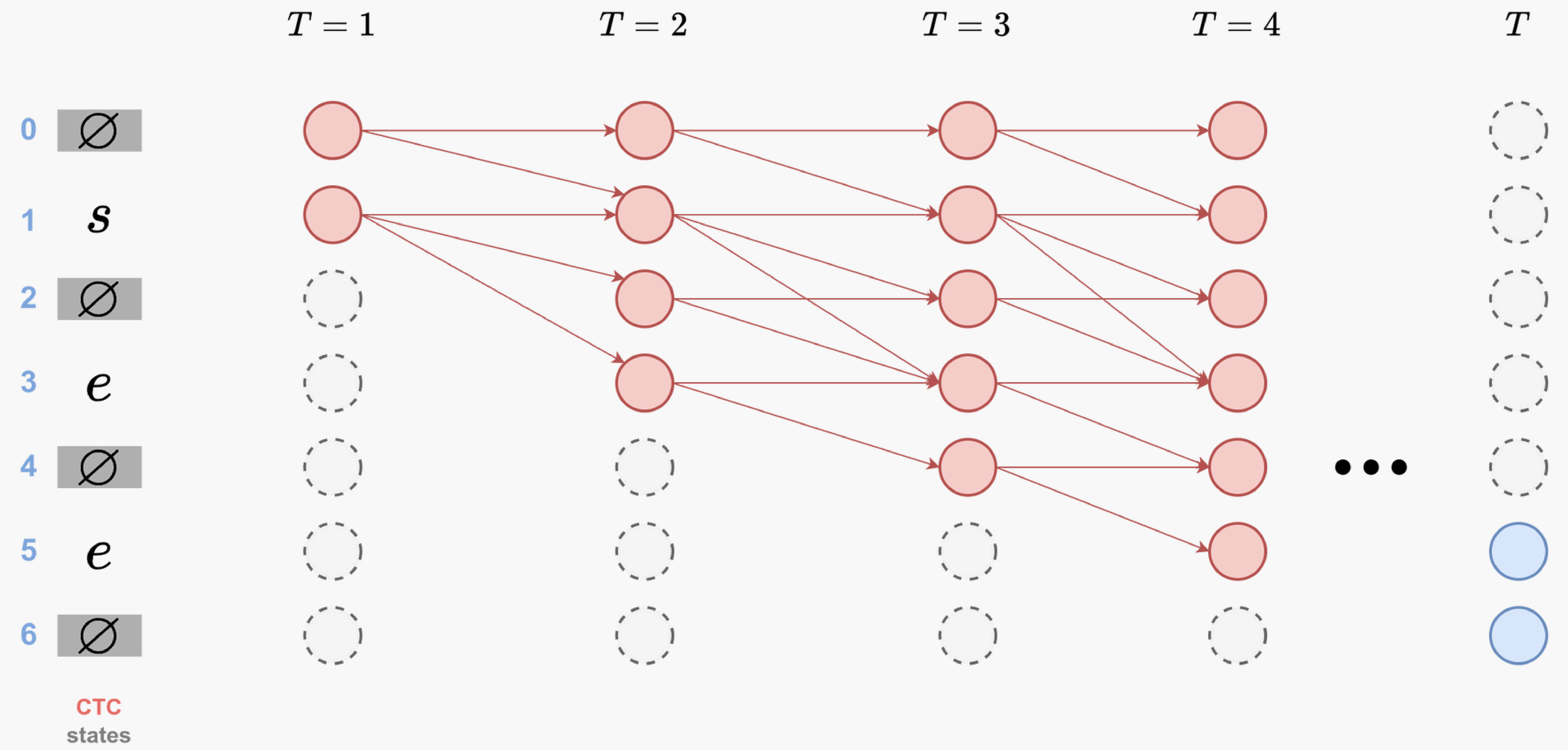
Dynamic Programming



RELEVANT LINKS

DATE
24/03/2025

Dynamic Programming



RELEVANT LINKS

DATE
24/03/2025

CTC formulation

$$\alpha_t(j) = \sum_{i=j-2}^{j} \alpha_{t-1}(i) a_{ij} b_t(y'_j)$$

$$t = 1 \ldots T$$

$$j = 1 \ldots 2l + 1$$

$$|\,\mathbf{O}\,| = T$$

$$|\,y\,| = l$$

RELEVANT LINKS

DATE
24/03/2025

CTC formulation

$$\boxed{\alpha_t(j) = \sum_{i=j-2}^{j} \alpha_{t-1}(i) a_{ij} b_t(y'_j)}$$

$$t = 1 \dots T \qquad | \mathbf{O} | = T$$

$$j = 1 \dots 2l + 1 \qquad | y | = l$$

$b_t(y'_j)$ → the probability given by NN to the symbol y'_j

CTC formulation

$$\boxed{\alpha_t(j) = \sum_{i=j-2}^{j} \alpha_{t-1}(i) a_{ij} b_t(y'_j)}$$

$$t = 1 \ldots T$$

$$j = 1 \ldots 2l + 1$$

$$|\,\mathbf{O}\,| = T$$

$$|\,y\,| = l$$

$b_t(y'_j)$ $\rightarrow$ the probability given by NN to the symbol y'_j

$$y'_j = \begin{cases} y_{j/2} & \text{if } j \text{ is even} \\ \varnothing & \text{otherwise} \end{cases}$$

RELEVANT LINKS

DATE
24/03/2025

CTC formulation

$$\boxed{\alpha_t(j) = \sum_{i=j-2}^{j} \alpha_{t-1}(i) a_{ij} b_t(y'_j)}$$

$$t = 1 \ldots T \qquad | \mathbf{O} | = T$$

$$j = 1 \ldots 2l + 1 \qquad | y | = l$$

$b_t(y'_j)$ → the probability given by NN to the symbol y'_j

$$y'_j = \begin{cases} y_{j/2} & \text{if } j \text{ is even} \\ \varnothing & \text{otherwise} \end{cases}$$

$$a_{ij} = \begin{cases} 1 & \text{if } i = j \text{ or } i = j - 1 \\ 1 & \text{if } i = j - 2 \text{ and } y'_j \neq y'_{j-2} \\ 0 & \text{otherwise} \end{cases}$$

RELEVANT LINKS

DATE
24/03/2025

CTC formulation

$$Pr(\mathbf{y} \mid \mathbf{O}) = \sum_{a \in \mathcal{B}^{-1}(\mathbf{y})} Pr(a \mid \mathbf{O}) = \underbrace{\alpha_T(2l)}_{\text{alignment ends with last token}} + \underbrace{\alpha_T(2l+1)}_{\text{alignment ends with blank}}$$

RELEVANT LINKS

DATE
24/03/2025

Prediction with CTC

Pick the single most probable output at every time step

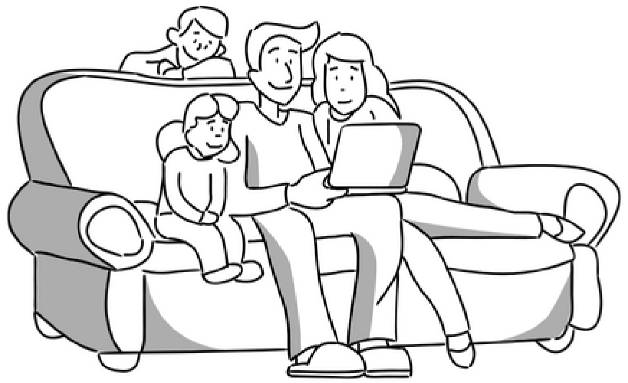
$$\arg\max_{\mathbf{y}} Pr(\mathbf{y} \mid \mathbf{O}) \approx \mathcal{B}(\arg\max_{a} Pr(a \mid \mathbf{O}))$$

RELEVANT LINKS

DATE
24/03/2025

Thank you!

Email: darshanp@cse.iitb.ac.in



Want to know more about ASR? Check out:

CS753: Automatic Speech Recognition