Sec 6.2

Simultaneous ST

Simultaneous Translation

- Generate translation while speaker speaks
- Tradeoff:
  - *More context* improves speech translation
    - Wait as long as possible
  - *Low latency* is important for user experience
    - Generate translation as early as possible
- Challenge:
  - Different word order in the language
    - SOV vs SVO

German	Ich	melde	mich	zum	E2E	Tutorial	an
Gloss	I	register/ cancel	myself	to	E2E	tutorial	
English	I	????					

Simultaneous Translation

- Approaches:
  - Learn optimal segmentation strategies
    - Create segments that optimizing tradeoff between segment length and translation quality
    - Advantages:
      - No changes to the system
    - Disadvantage:
      - Shorter context during translation
    - Mainly used in cascaded approaches (e.g. Oda et al., 2014)

Example:

Ich melde mich

zur Konferenz an

Simultaneous Translation

- Approaches:
  - Learn optimal segmentation strategies
  - Re-translate / Iterative -update
    - Directly output first hypothesis
    - If more context is available:
      - Update with better hypothesis
    - Cascade
      - (Niehues et al, 2018; Arivazhagan et al, 2020)
    - End-to-end
      - (Weller et al, 2021)

Example:

Ich
I

Ich melde mich
I register

Ich melde mich von
I cancel my registration for

Re-translation

- Challenge:
  - Flickering
- Ideas:
  - Output masking
    - Do not output last tokens
  - Constrained decoding:
    - Fixed part of the previous translation

Example:

Ich
I

Ich melde mich
I register

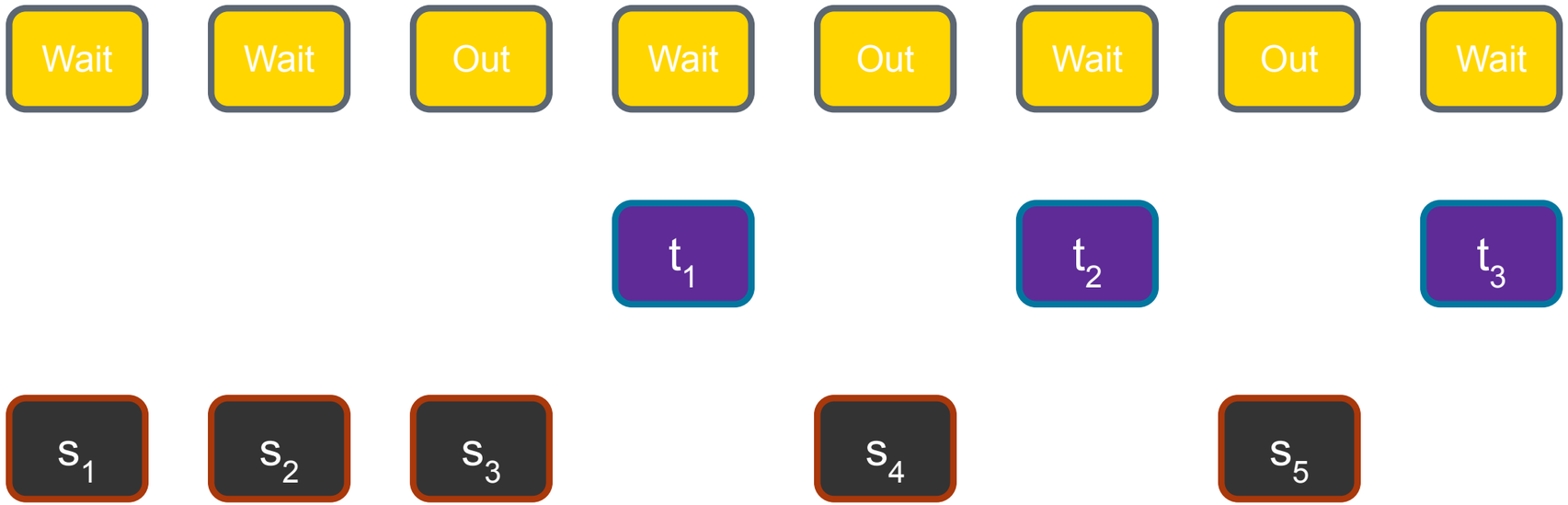
Ich melde mich von
I cancel my
registration for

Simultaneous Translation

- Approaches:
  - Learn optimal segmentation strategies
  - Re-translate
  - Stream decoding
    - Dynamically learn when to generate a translation
    - At each time step:
      - Decided to output word
      - Wait for additional input

Stream decoding

- Methods:
  - Fixed schedule (Ma et al, 2019)
    - Wait-k policy



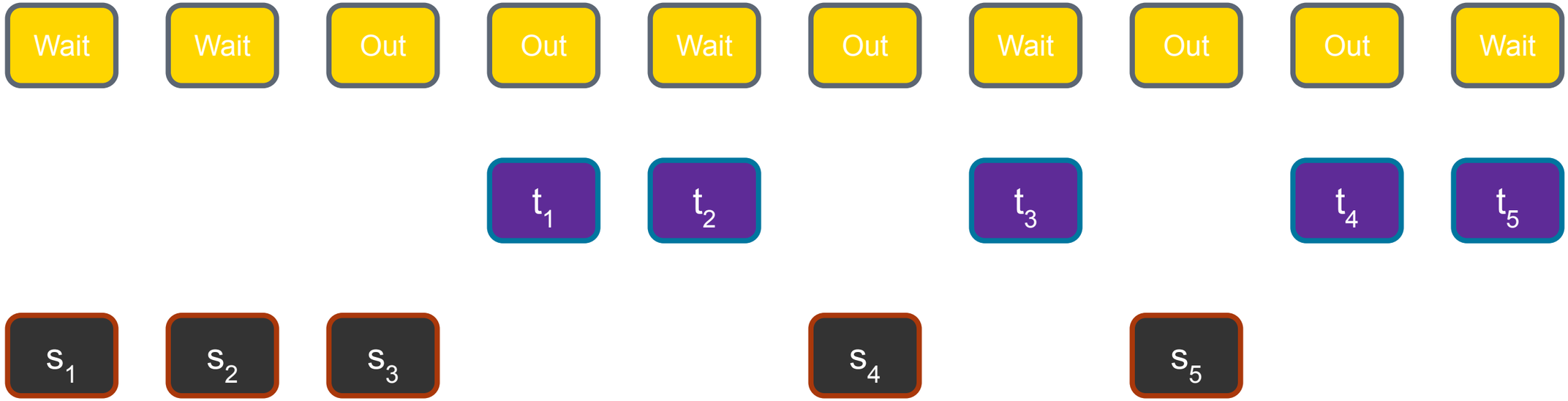
Stream decoding

- Challenges:
  - Assumes constant rate between input and output
    - Speaking speed varies

- Ideas:
  - Estimate word boundaries on the source side (Ma et al. 2020)
    - Predict using CTC Loss (Ren et al, 2020)

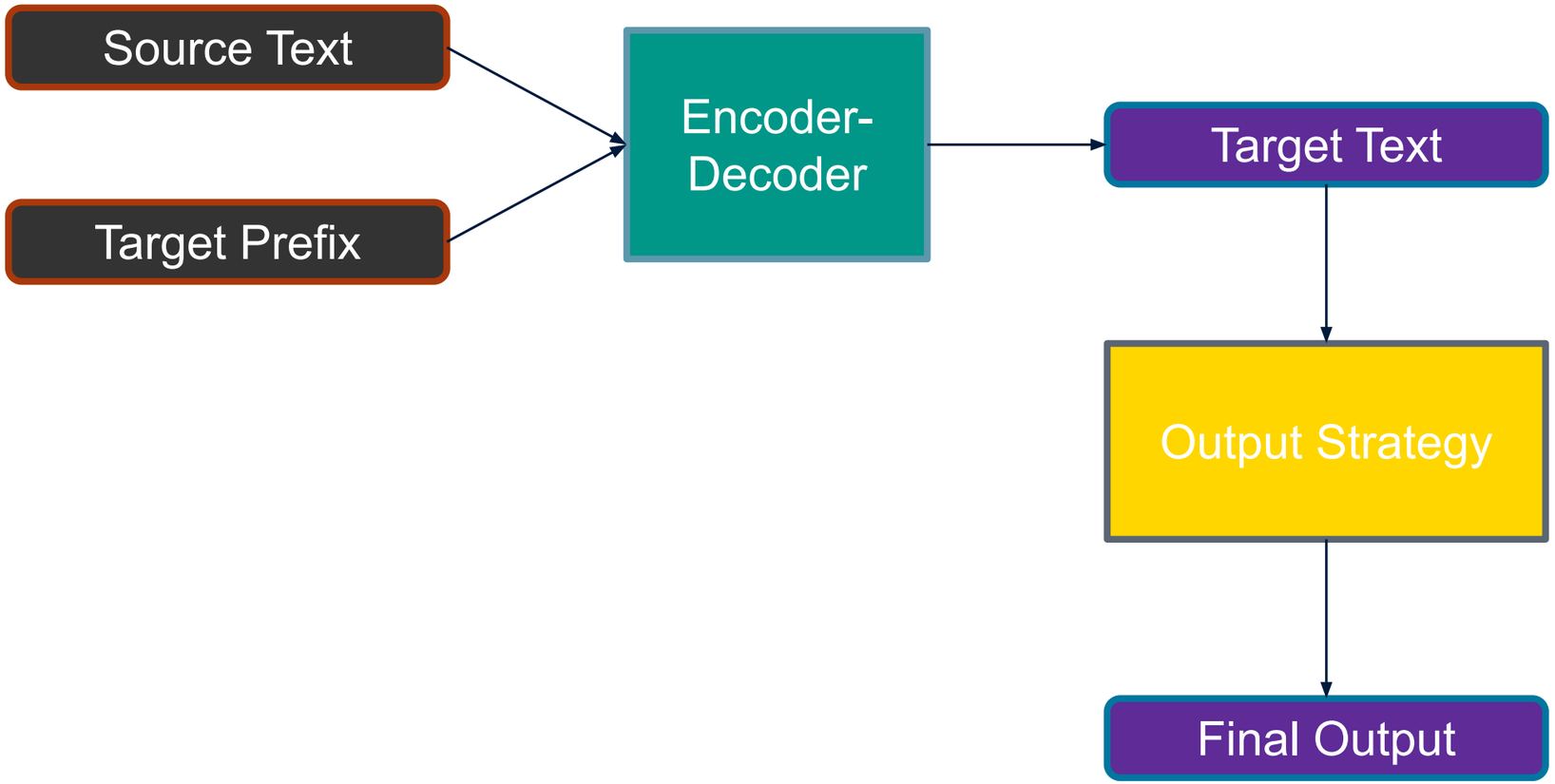
Stream decoding

- Methods:
  - Fixed schedule (Ma et al, 2019)
  - Dynamic decision (Cho et al, 2016; Gu et al, 2017; Dalvi et al, 2018)
    - End-to-end:
      - Estimate output probability based on confidence



Stream decoding using Retranslation

- Decoding with fixed target prefix



Stream decoding strategies

- Local agreement (Liu et al, 2020)
  - Output if previous and current output agree on prefix
  - Variation (Yao et al., 2020):
    - Predict the next source word instead of relying on the previous input

Input	Prefix	Target Text	Final Output
1	Ø	All model trains	Ø
1,2	Ø	All models art	All
1,2,3	All	All models are wrong	All models
1,2,3,4	All models		
…			