(Simultaneous) Speech Translation: Challenges and Techniques

Jan Niehues

27/10/2020

jan.niehues@maastrichtuniversity.nl

Maastricht University

Overview

- Motivation
- Speech Translation Models
  - Cascaded approach
  - End-to-End Speech Translation
- Challenges
  - Segmentation
  - Simultaneous translation

Use cases

- Conferences / Lectures

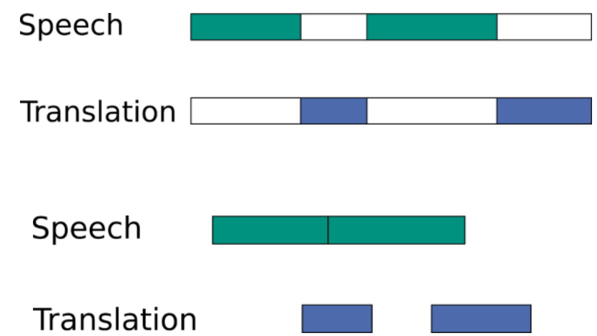
- Internet videos
  - Youtube, Facebook, …

- Television
- Meetings

- Telephone conversations

Different Application scenarios

- Sequence
  - Consecutive translation
  - Simultaneous translation
    - Differences:
      - Segmentation
      - Speech overlap

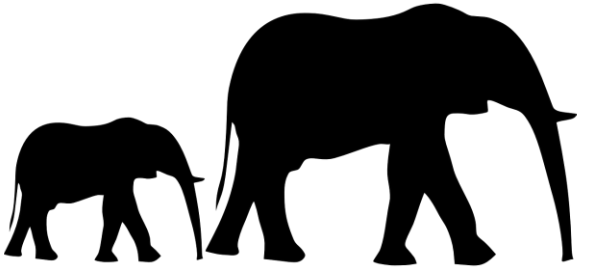


Different Application scenarios

- Sequence
- Number of speakers
  - Single speaker
    - e.g. Presentations
  - Multiple speaker
    - e.g. Meetings
    - Challenges:
      - Overlapping voice
  - Mainly increases difficulty for speech recognition

Different Application scenarios

- Sequence
- Number of speakers
- Online/Offline systems
  - Online: Translate during production of speech
  - Offline: Translate full audio
    - e.g. movies
  - Real-time translations:
    - Translation as fast as speech input
  - Latency
    - Time passes between speech and translation



Different Application scenarios

- Sequence
- Number of speakers
- Online/Offline systems
- Output Modality
  - Text:
    - Most commonly used
    - Reviseble
  - Speech
    - More natural?

Data

- Fischer data [Post et al., 2013]
  - Languages: Spanish to English
  - Domain: Telephone conversation

- MuST-C Corpus [Di Gangi et al., 2019]
  - Languages: English to 8 European Languages
  - Domain: TED
- LIBRI-TRANS [Kocabiyikoglu et al., 2018]
  - Languages: English to French
  - Domain: Audio books
- MASS [Boito et al, 2019], STC [Shimizu et al., 2014], BSTC, ..

The Model

- Speech Translation

- Important technology

Automatic Speech Recognition

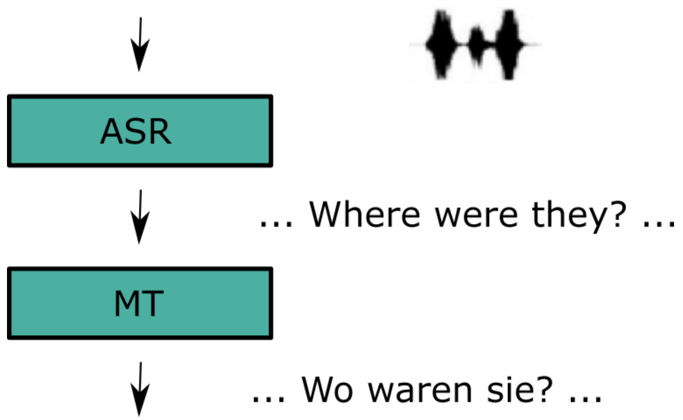
Machine Translation

- Architectures
  - Cascade
  - End-to-End

Cascade Translation

- Serial combination of several models
  - Automatic speech recognition (ASR)
  - Machine translation (MT)

- Advantages:
  - Data availability
  - Modular system
  - Easy incorporation of new ASR/MT developments

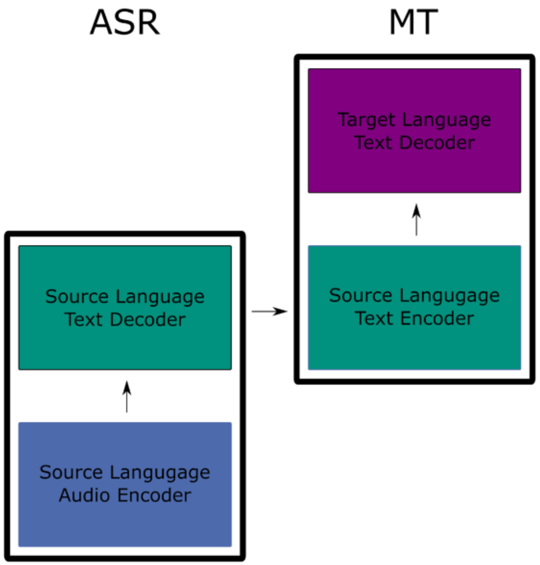


Challenges

- Error propagation
  - Even best components lead to errors
  - Techniques
    - Ignore
    - Represent different hypothesis
      - N-Best lists
      - Lattices [Saleem et al, 2005;Matusov et al, 2005]
    - Robust to errors [Tsvetok et al. 2014;Lewis et al., 2015;Sperber et al, 2017]
- Separate optimization
- Script for source language is needed
- Computational complexity
- Information loss

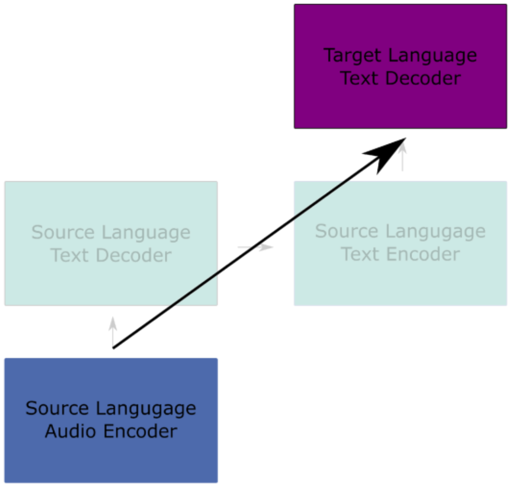
End-to-End SLT

- Opportunity:
  - Sequence to Sequence models successfully applied to both tasks



End-to-End SLT

- Opportunity:
  - Sequence to Sequence models successfully applied to both tasks
  - [Duong et al., 2016;Berard et al., 2016; Weiss et al., 2017]



E2E SLT - Challenges

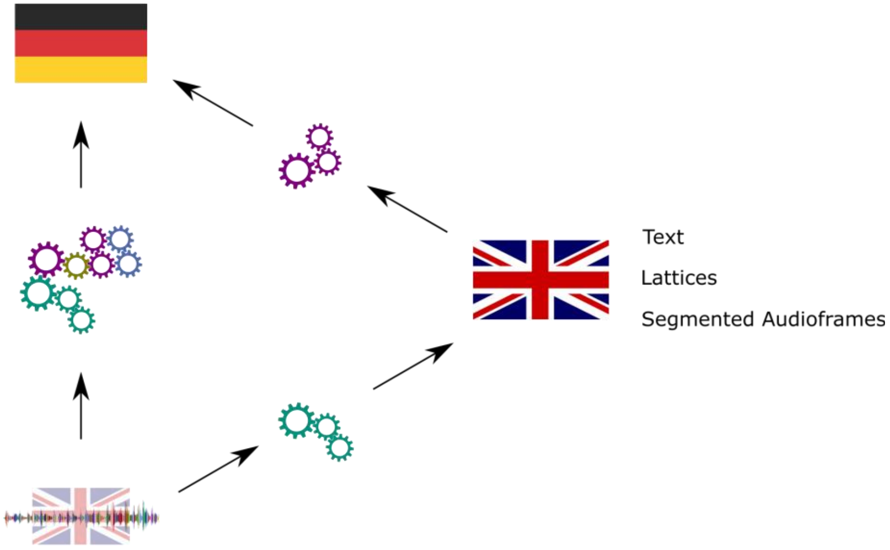
- Task complexity
  - Complicated mapping between source and target sequence
    - Source transcript can be intermedia supervised signal

- Data availability
  - Few end-to-end speech translation corpora available

Intermediate Representation

- Idea:
  - Reintroduce intermediate representation
    - Use additional data based on intermediate representation
    - Simplify task

- Representations:
  - Source Language Transcript
  - Segmented audio frames
  - Lattices

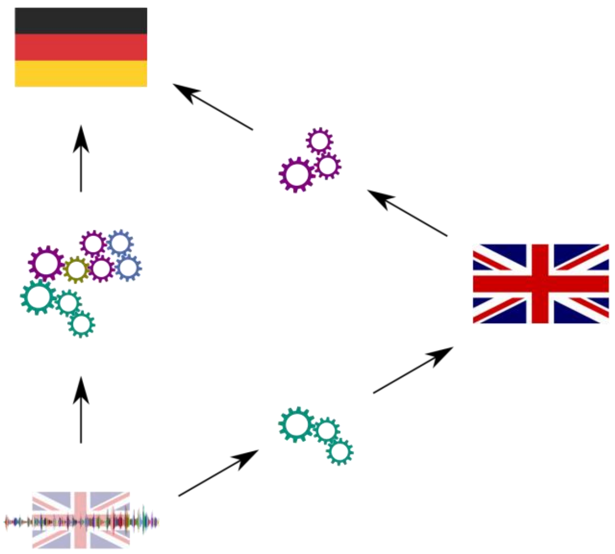


Integration

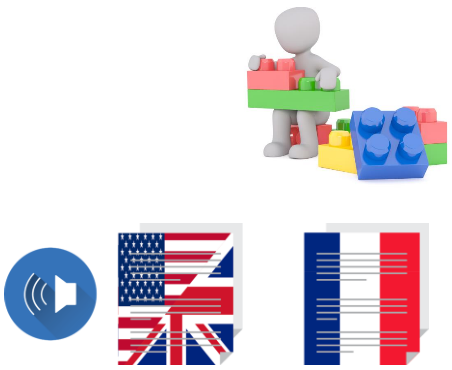
- When to use what component?
  - Training
  - Inference

- How to use the components?
  - Data Generation
  - Add. Loss function

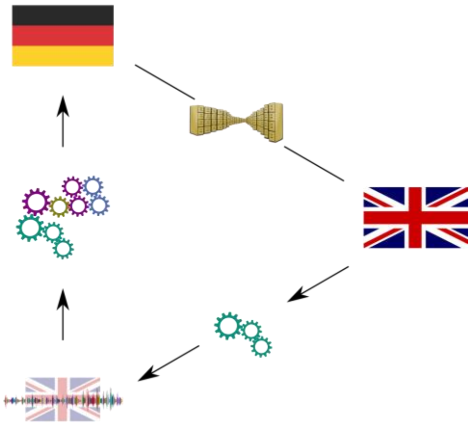
- What parameters to share?
  - Share parameters between different tasks



Synthetic data



- Automatic generation by using TTS
  - [Berard et al, 2016; Kano et al, 2018;]
- Challenge:
  - Generalization from TTS output to real audio signal

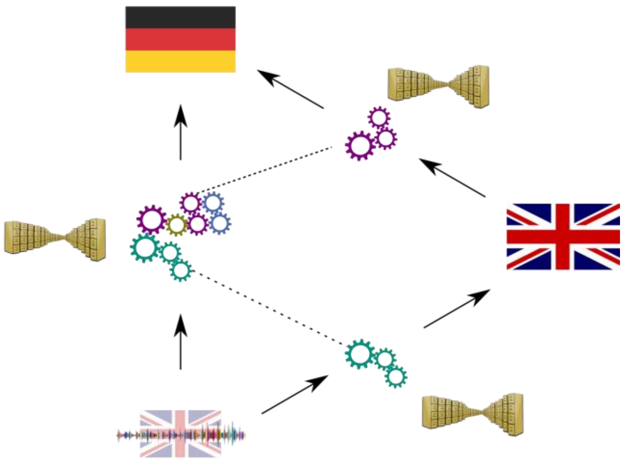


Training

Inference

Multi-task learning

- Available data:
  - Speech data
  - Parallel MT data
- Idea:
  - Share parts of the network
  - Train SLT system using speech or MT data

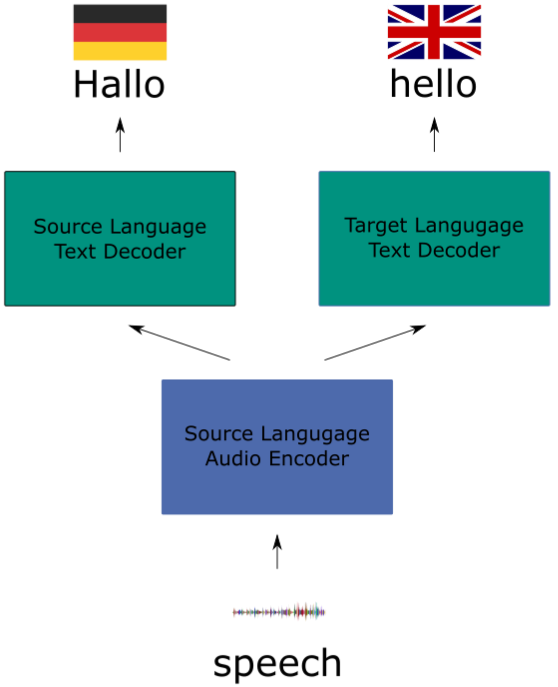


Training

Inference

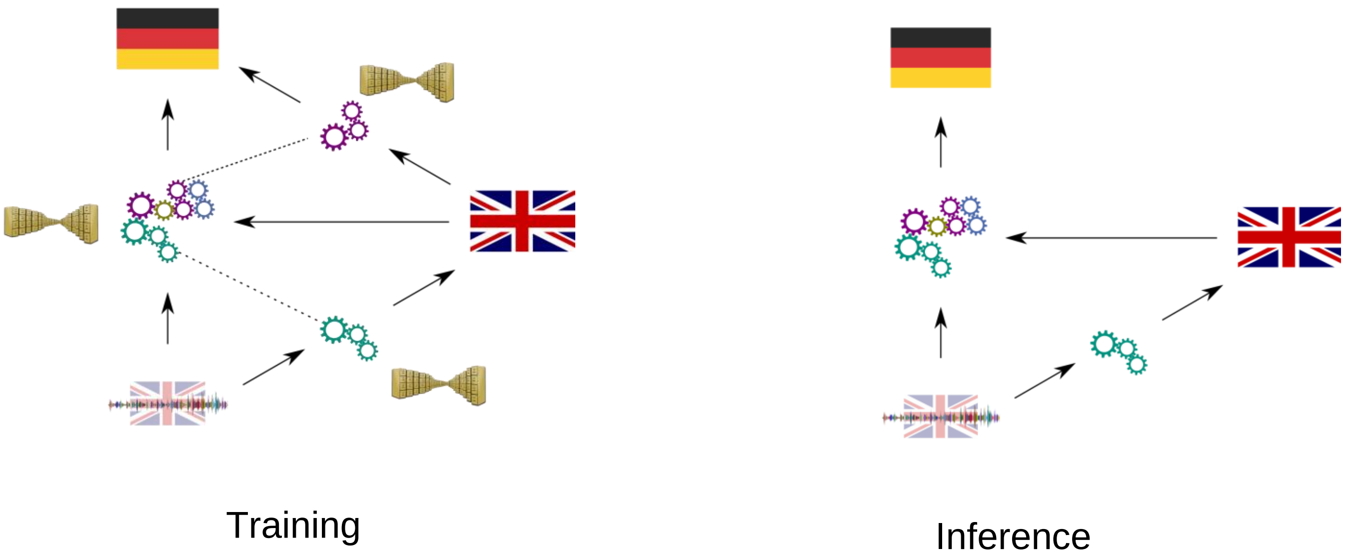
Multi-task learning

- Pre-training (Kano et al., 2018):
  - Train encoder on ASR task
  - Reuse on SLT task
- Multitasking (Weiss et al.,2017):
  - Train SLT and ASR jointly
- Challenge:
  - Data efficiency
  - How much gain from ASR/MT data?



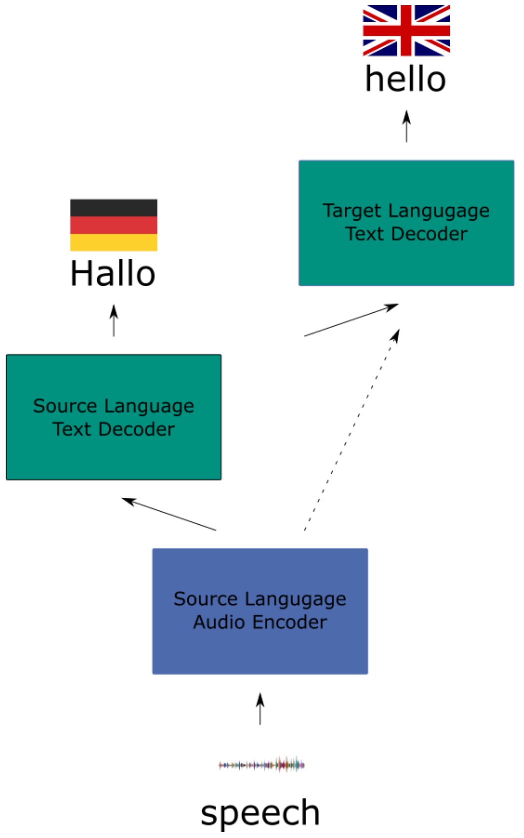
2-stage NN Model

- Intermediate representation in inference



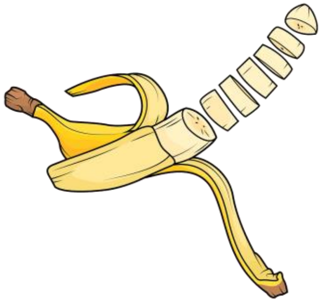
2-stage NN Model

- Intermediate representation in inference
- Stack different decoders
  - Attend to source language decoder hidden states
- Triangle version:
  - Attend to source audio and source text [Anastasopoulos Chiang, 2018]
- Shared context vectors:
  - Ignore hard decisions of source language decoder [Sperber et al;2019]

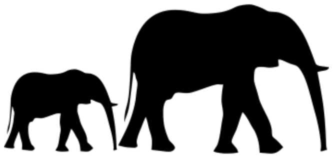


Challenges

- Sentence Segmentation:
  - Text: Sentence-based models
  - Audio: Continuous streams



- Simultaneous Translation:
  - Generate translation while speaking
  - Low-Latency



Challenges – Sentence Segmentation

- Many applications:
  - Continuous audio stream
  - No punctuation in spoken language

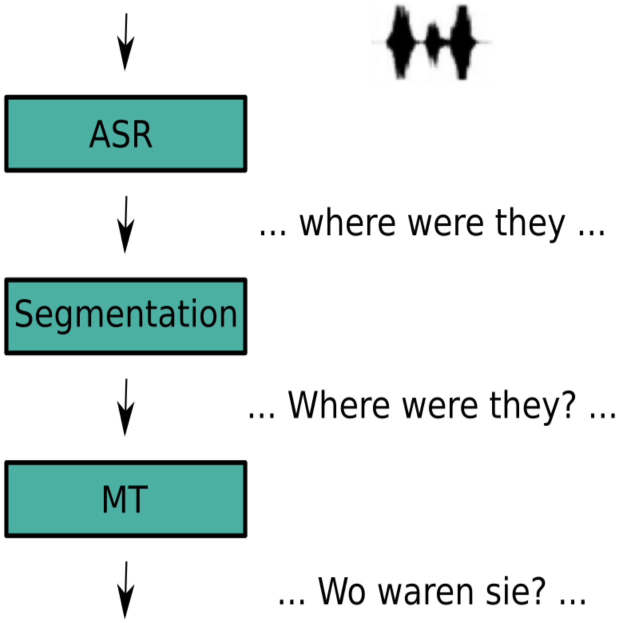
- Automatic segmentation and punctuation needed
  - Readability
  - Semantic
    - Let’s eat Grandpa !
    - Let’s eat, Grandpa !
  - Processing
    - MT often operate on sentence level

Segmentation and Punctuation

- Add segmentation as additional component

- Approaches:
  - Language model-based
    [Stolcke et al, 1998; Rao et al, 2007]
  - Sequence labeling [Lu and Ng, 2010]
  - Monolingual machine translation
    [Peitz et al, 2011;Cho et al, 2012]

- Integration:
  - Between ASR and MT
  - After MT
  - Include into MT



Simultaneous Translation

- Generate translation while speaker speaks
- Tradeoff:
  - More context improves speech recognition and machine 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	zur	Summer	School	an
Gloss	I	register/ cancel	myself	to	summer	School	
English	I	????					

Simultaneous Translation

- Approaches:
  - Learn optimal segmentation strategies
  - Re-translate
    - Update previous translation with better once
  - Stream decoding
    - Dynamically learn when to generate a translation

Optimizing segmentation

- Idea:
  - Create segments that optimizing tradeoff between segment length and translation quality

- Advantages:
  - No changes to the NMT system
- Disadvantage:
  - Shorter context during translation

- E.g.:
  - Oda et al., 2014

Example:

Ich melde mich

zur Konferenz an

Iterative Updates

- Directly output first hypothesis
- If more context is available:
  - Update with better hypothesis

- Example:
  - Ich melde mich
  - I register

  - Ich melde mich von der Klausur ab
  - I withdraw form the exam

- Not only for MT, but for all components [Niehues et al, 2016]
  - No adaptation of the architecture

Update mechanism

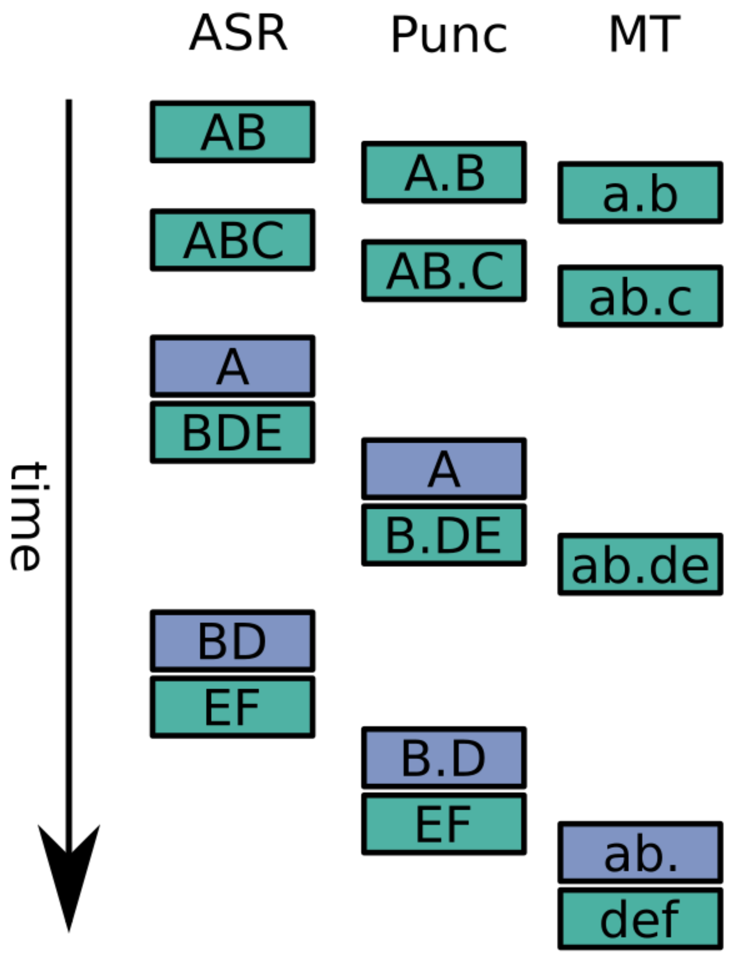
you?

MT

Wo waren Sie?

Where were you?

Iterative Updates - Framework



Latency

- Time in seconds till words appear
  - Brackets:
    - Words do not change anymore

	English- French	German-English
ASR - Static	4.9	5.7
ASR - Updates	1.7 (2.3)	1.6 (2.2)
MT - Static	7.5	8.6
MT - Updates	1.8 (3.3)	2.0 (5.3)

Adaptation to NMT

- Challenge:
  - NMT always tries to generate complete sentence
  - Example:
    - I encourage all of
    - Yo animo a todo el mundo .

  - Train-Test mismatch

Adaptation to NMT

- Idea:
  - Train NMT on partial sentences
  - No parallel data available -> Generate artificial data

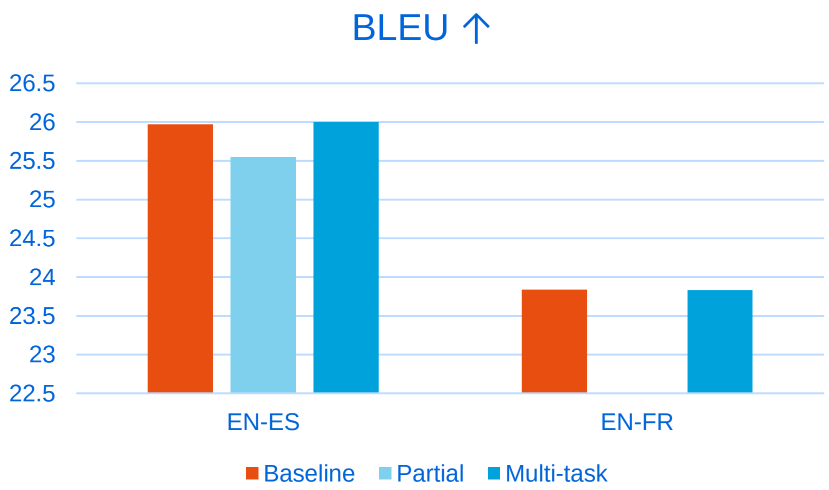
- Source data:
  - Every prefix of the sentence
- Target data:
  - Constraints:
    - As long as possible for low latency
    - Substring of previous prefix for few rewrites
  - Length-based
    - Same ratio of source and target sentence

Adaptation to NMT

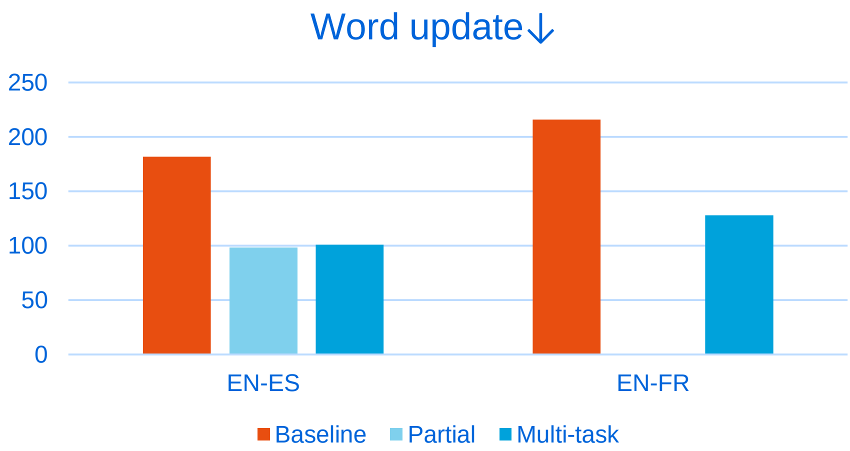
Source	Target
Ich	I
Ich bin	I
Ich bin nach	I went
Ich bin nach Hause	I went
Ich bin nach Hause gegangen	I went home

- Many more prefixes than full sentence
  - Concentrating on prefixes
- Multi-task training
  - Mix partial and full sentences (Ratio 1:1)

Results



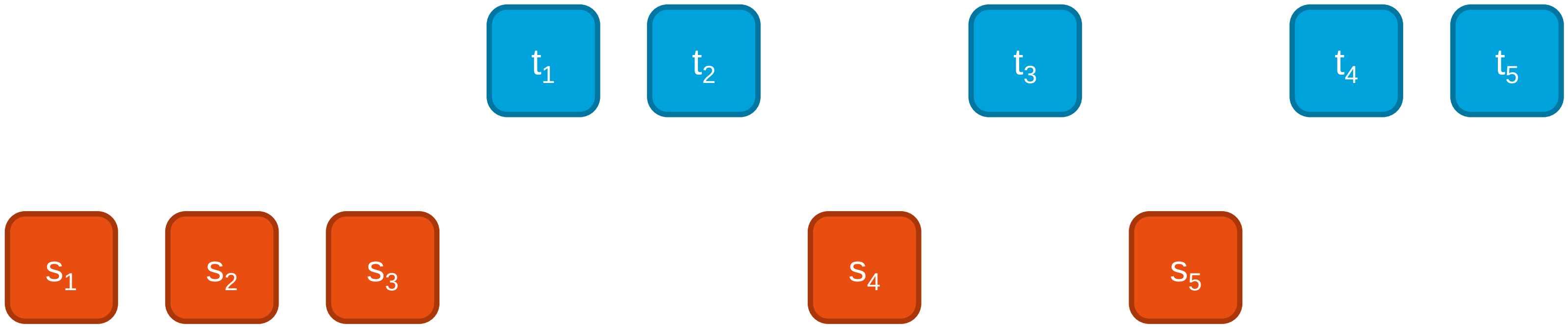
BLEU ↑
26.5
26
25.5
25
24.5
24
23.5
23
22.5
EN-ES
EN-FR
Baseline
Partial
Multi-task



Word update↓
250
200
150
100
50
0
EN-ES
EN-FR
Baseline
Partial
Multi-task

Stream decoding

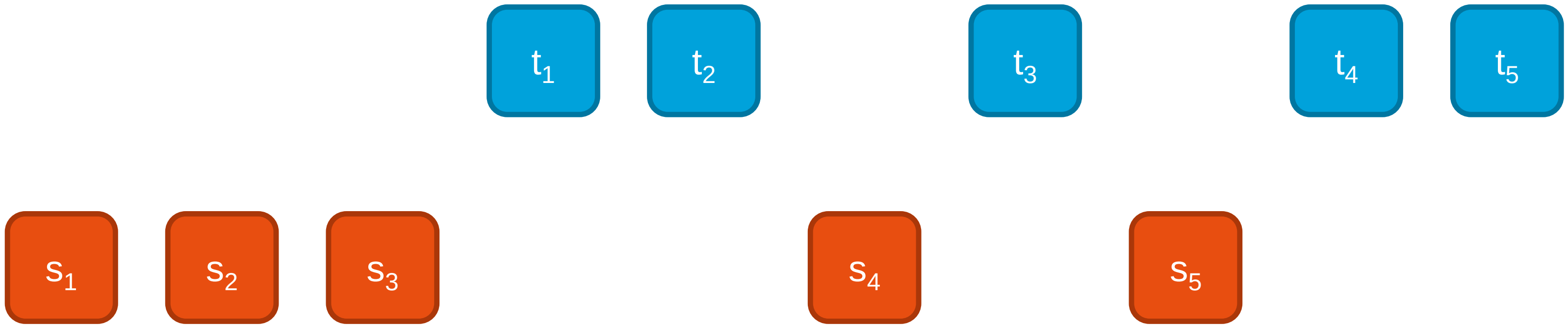
- Idea:
  - At each time step:
    - Decided to output word
    - Wait for additional input
  - (Kolss et at., 2008)



Stream decoding

- Architecture:
  - Encoder-Decoder

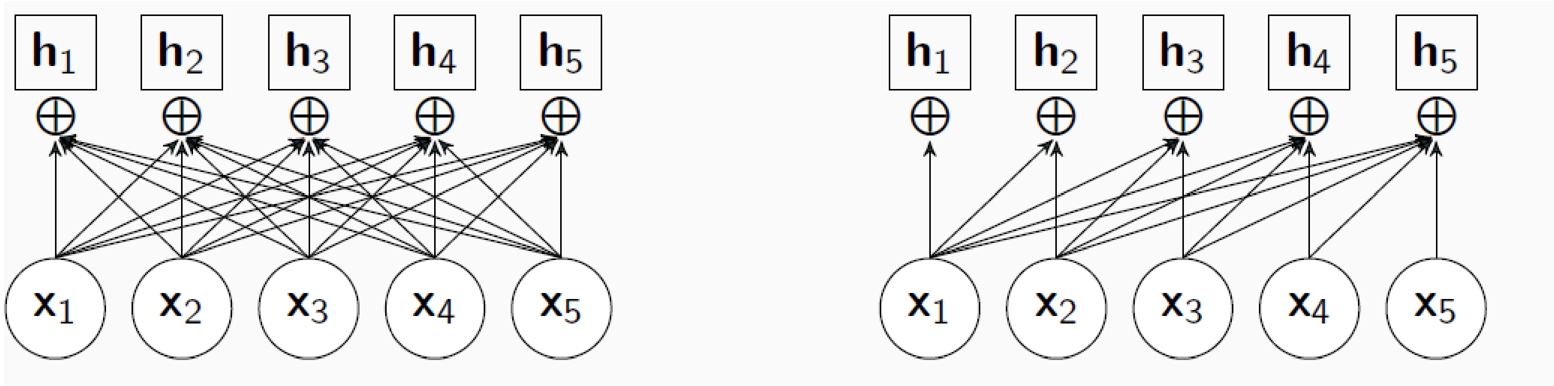
- Challenges:
  - Encoder: Only past input is available
  - Decoder: Wait or Output



Stream decoding - Encoder

- Encoder:
  - No information of the future

  - LSTM:
    - Unidirectional
  - Attention:
    - Only attend to pervious states

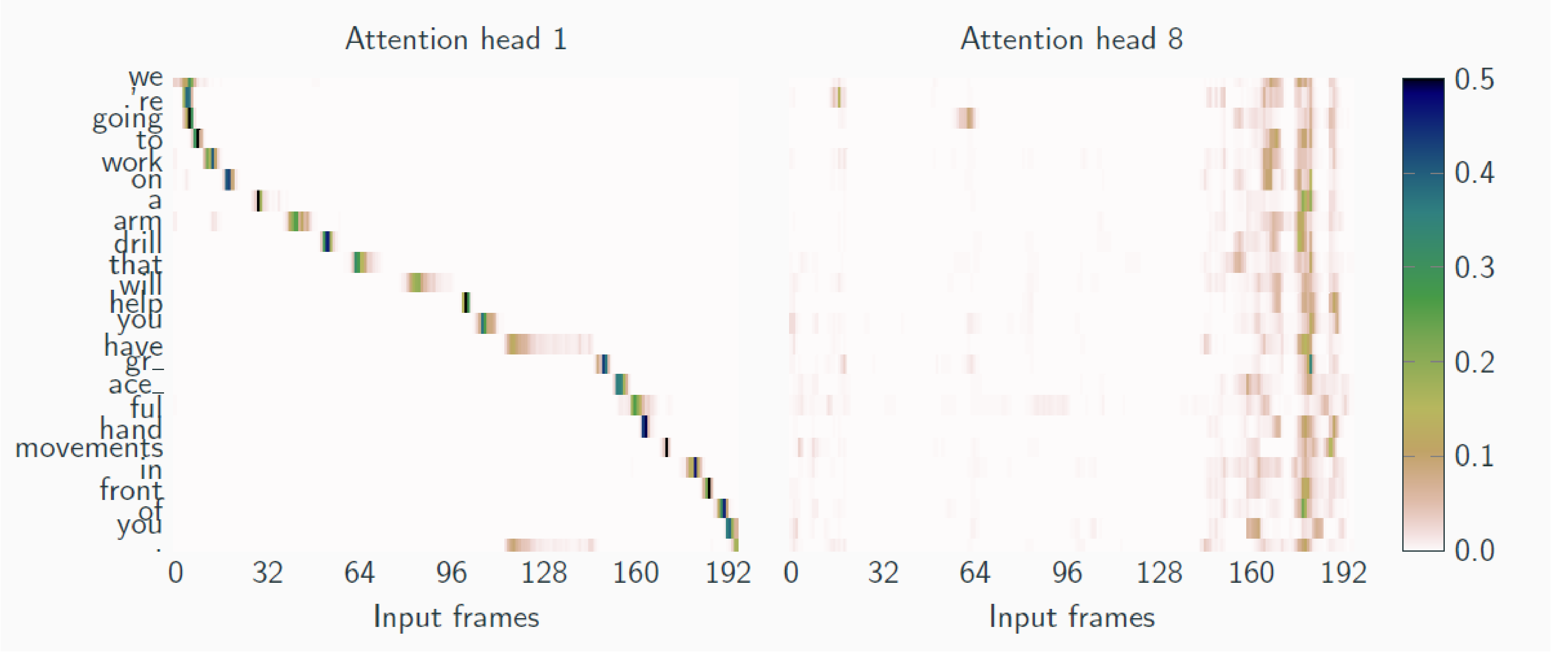


Experiments

- Automatic speech recognition
  - 3 data set
  - Encoder-Decoder Model using 32 Encoding/ 12 Decoder layers
  - Metric:
    - Word Error Rate

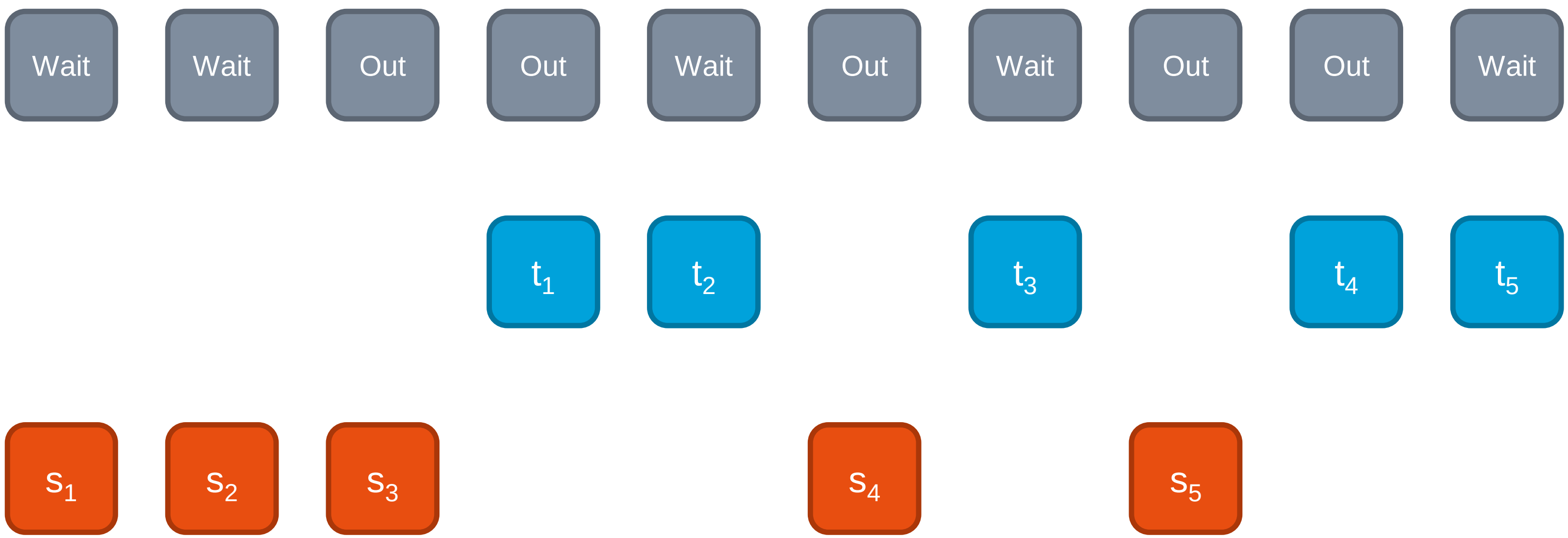
Dataset	Unidirectional	Bidirectional
How2	14.4	14.9
TED	11.1	11.1
LibriVox	9.2	9.7

Attention Matrix



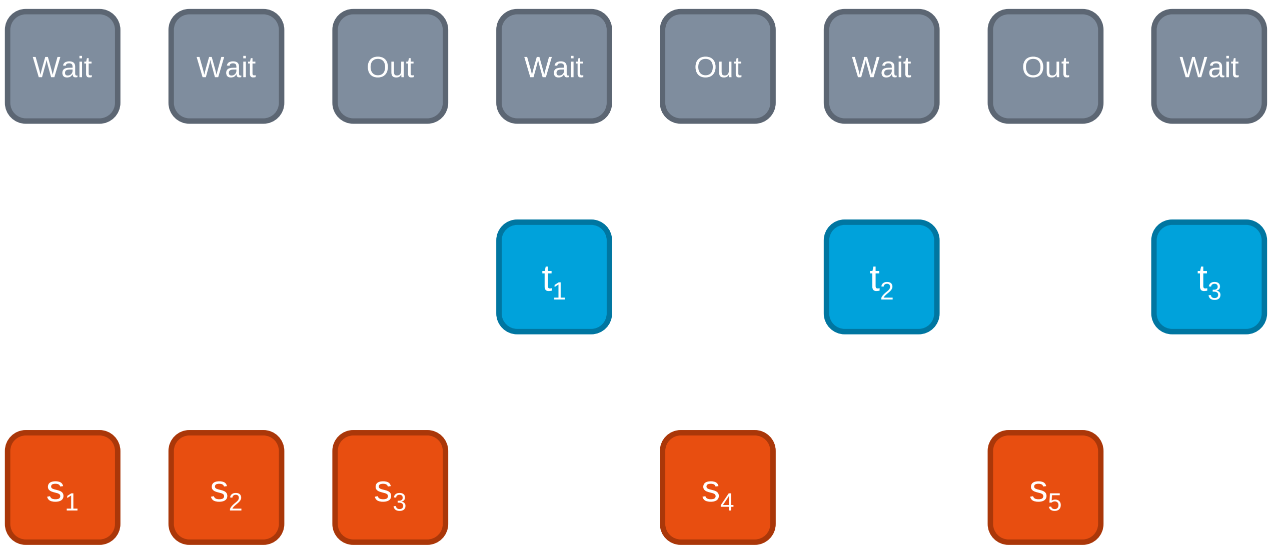
Stream decoding - Decoder

- Methods:
  - Dynamic decision [Cho et al, 2016; Gu et al, 2017; Dalvi et al, 2018]



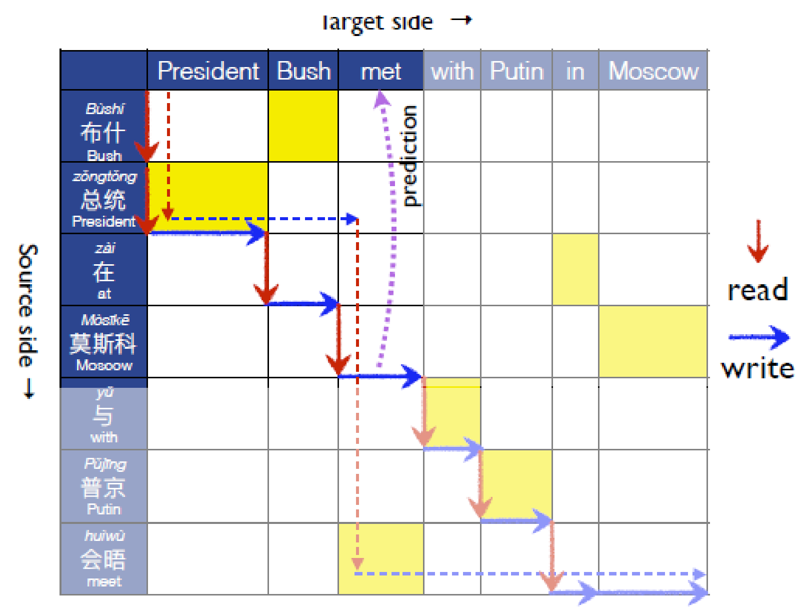
Stream decoding - Decoder

- Methods:
  - Dynamic decision Cho et al, 2016; Gu et al, 2017; Dalvi et al, 2018
  - Fixed schedule (Ma et al, 2019)
    - Wait-k policy



Stream decoding - Decoder

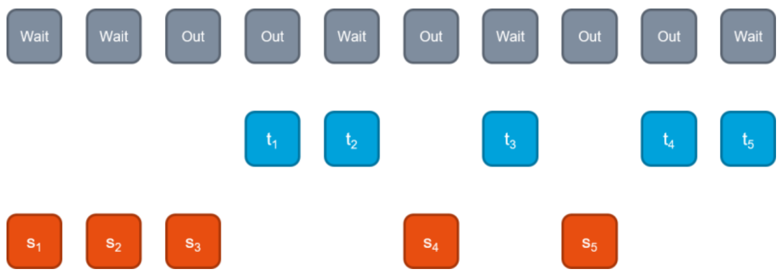
- Methods:
  - Dynamic decision Cho et al, 2016; Gu et al, 2017; Dalvi et al, 2018
  - Fixed schedule (Ma et al, 2019)
    - Wait-k policy



Ma et al., 2019

Relation to iterative Update

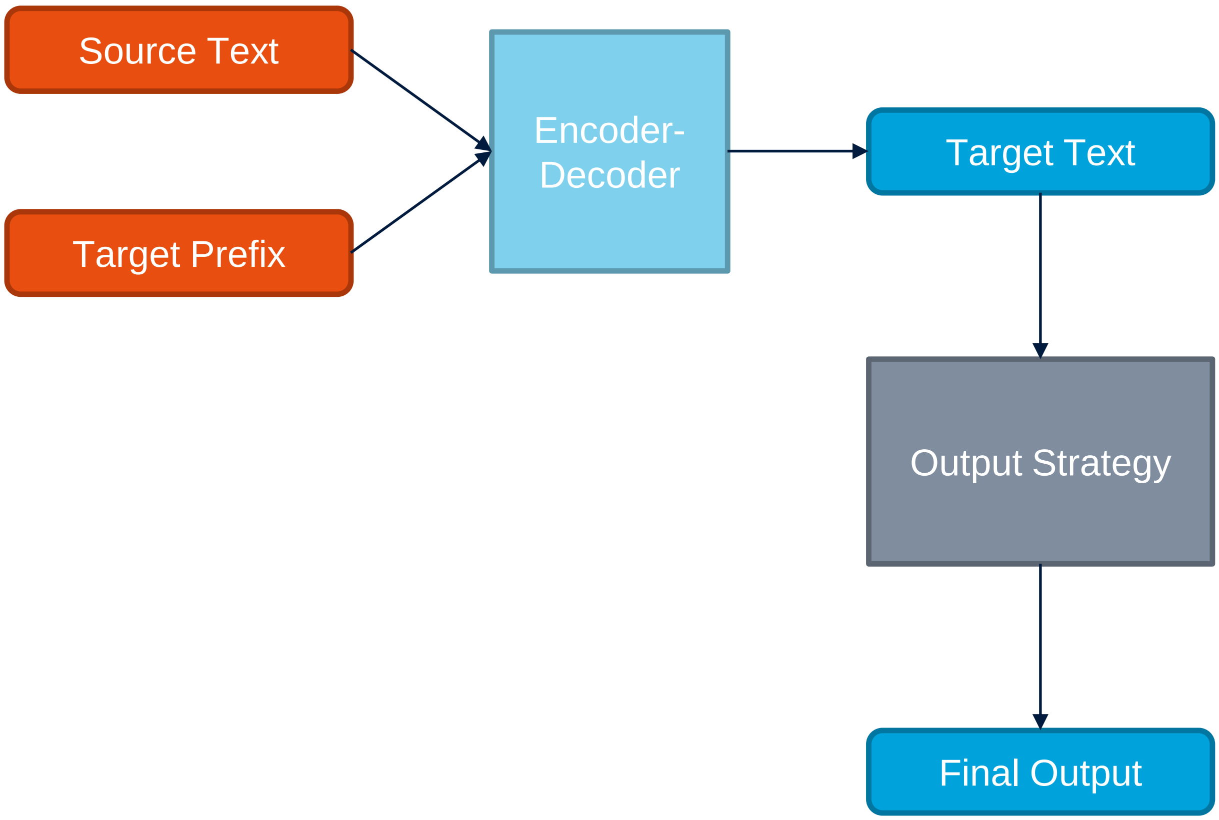
- Decoding with fixed target prefix



Chunks	Displayed	Output
S_1	Ø	Ø
S_1, s_2	Ø	Ø
S_1, s_2, s_3	Ø	t_1, t_2
S_1, s_2, s_3, s_4	t_1, t_2	t_1, t_2, t_3
S_1, s_2, s_3, s_4, s_5	t_1, t_2, t_3	t_1, t_2, t_3, t_4, t_5

Relation to iterative Update

- Decoding with fixed target prefix



Stream decoding strategies

- Wait-k
  - Wait for k seconds
  - Then output with fixed rate

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

Stream decoding strategies

- Hold-n
  - Do not output last n tokens

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

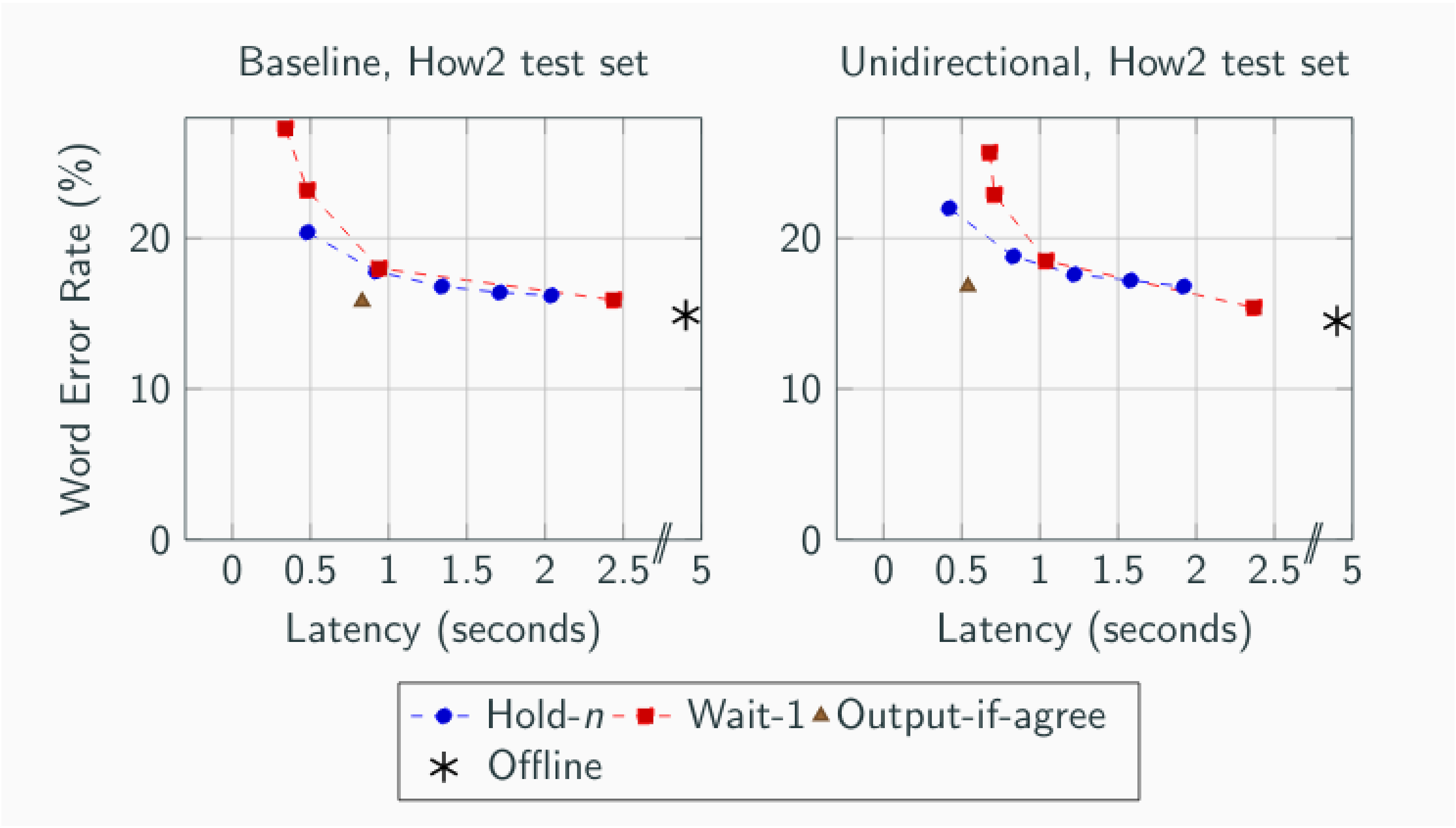
Stream decoding strategies

- 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		
…			

Latency vs. Accuracy

- Speech recognition results



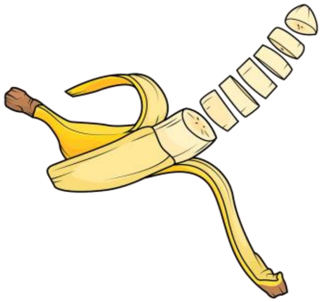
Speech Translation

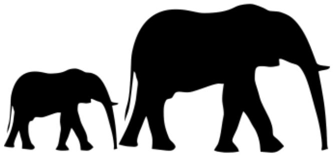
	BLEU	Latency diff.
Offline	44.5	4.36
Hold-2	37.3	0.48
Hold-4	42.2	0.95
Local Agreement	42.1	0.71

Summary

- Speech translation
  - Cascade models
  - End-to-End architecture

- Challenges
  - Segmentation and Punctuation
  - Simultaneous Translation
    - Shorter Segments
    - Stream decoding
    - Iterative updates





Thanks