Compositional End-to-End SLU

Presentor: Siddhant Arora

siddhana@andrew.cmu.edu

Parts of work from EMNLP 2022 Paper:

Token-level Sequence Labeling for Spoken Language Understanding using Compositional End-to-End Models

Carnegie Mellon University
Language Technologies Institute

Watanabe's Audio and Voice Lab

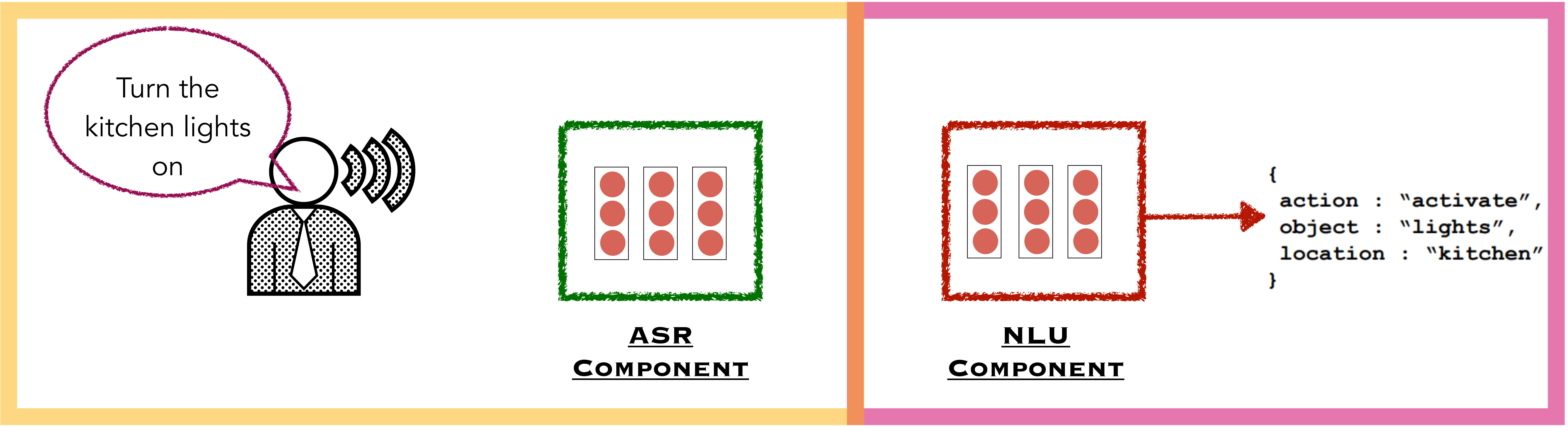
Content

- **Spoken Language Understanding**
- Sequence Labelling
- Current SLU Modelling
- Compositional Models
- Composition model for Sequence Labelling in SLU

Definition: Spoken Language Understanding

- As ASR systems get better, there is increasing interest of using ASR output for downstream NLP tasks.

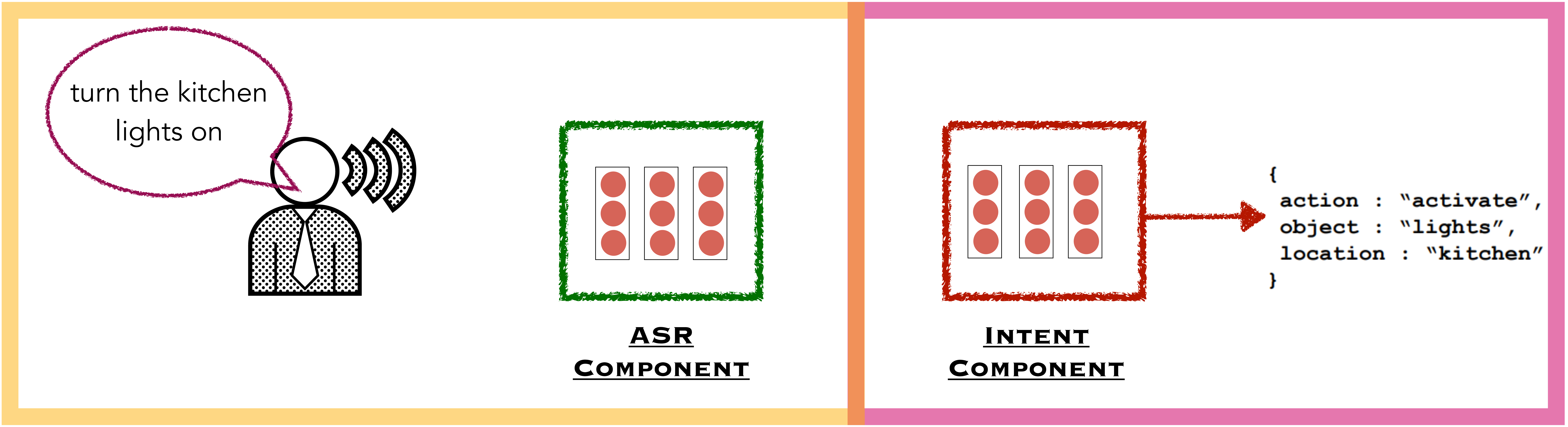
Example: Spoken Language Understanding (SLU) [1] = **ASR** + **NLU**



[1] Lugosch et al., 2021. Speech Model Pre-training for End-to-End Spoken Language Understanding. Interspeech 2019

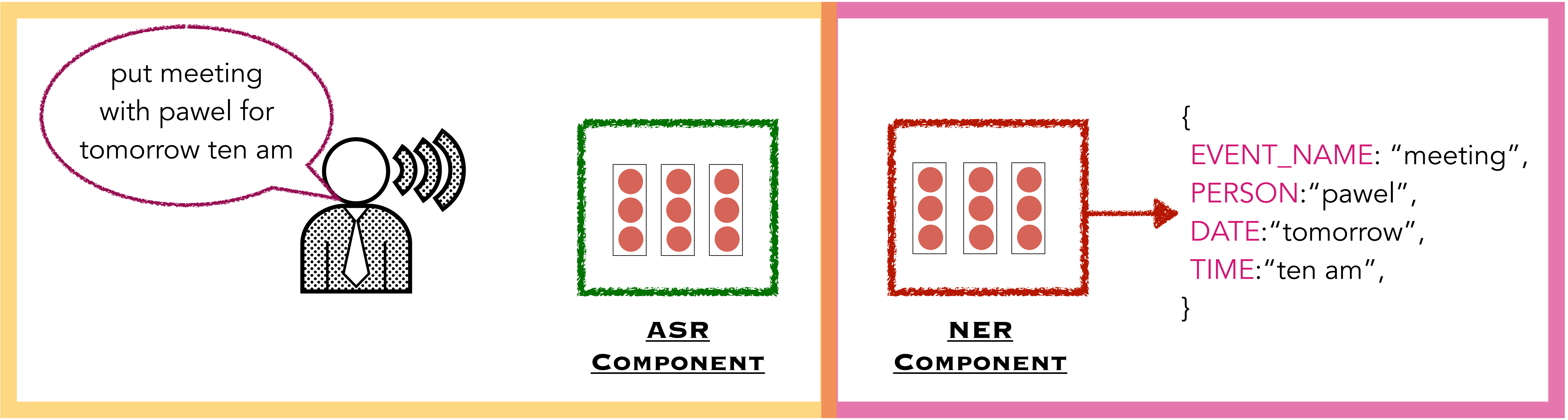
SLU Applications

Intent Classification : Spoken Utterance ➞ Executable Intent



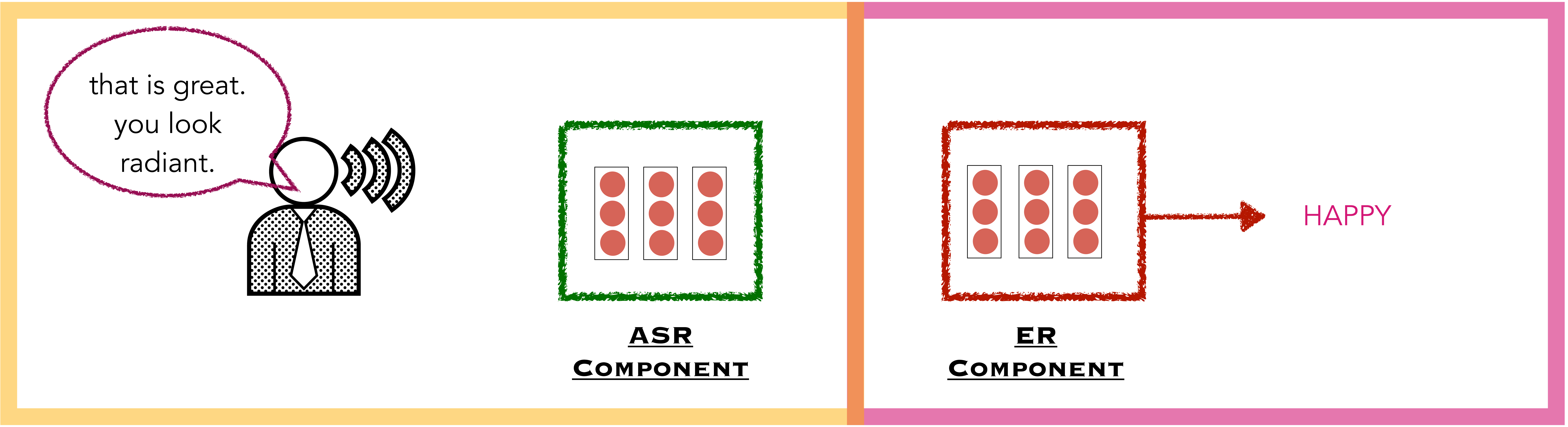
SLU Applications

Slot Filling : User Command ➞ Associated Entities



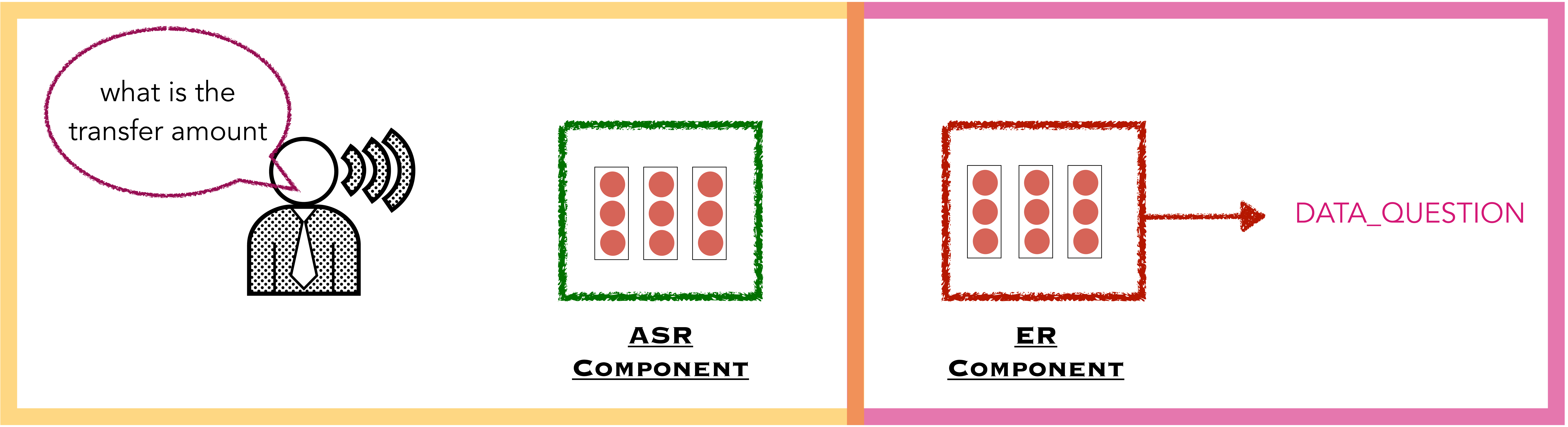
SLU Applications

Emotion Recognition : Understanding the emotion behind a utterance



SLU Applications

Dialogue Act Classification : Modeling the topic of a conversation



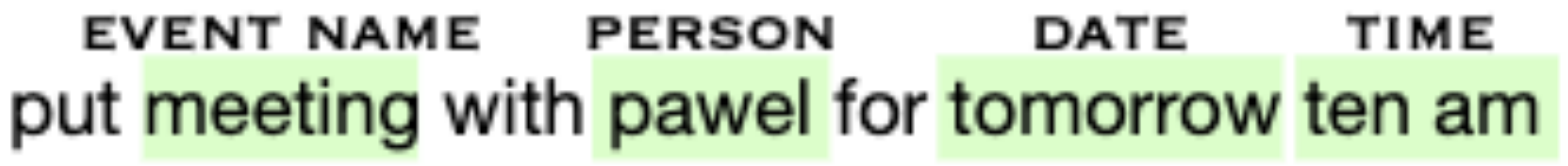
Content

- Spoken Language Understanding
- **Sequence Labelling**
- Current SLU Modelling
- Compositional Models
- Composition model for Sequence Labelling in SLU

Definition: Sequence Labeling

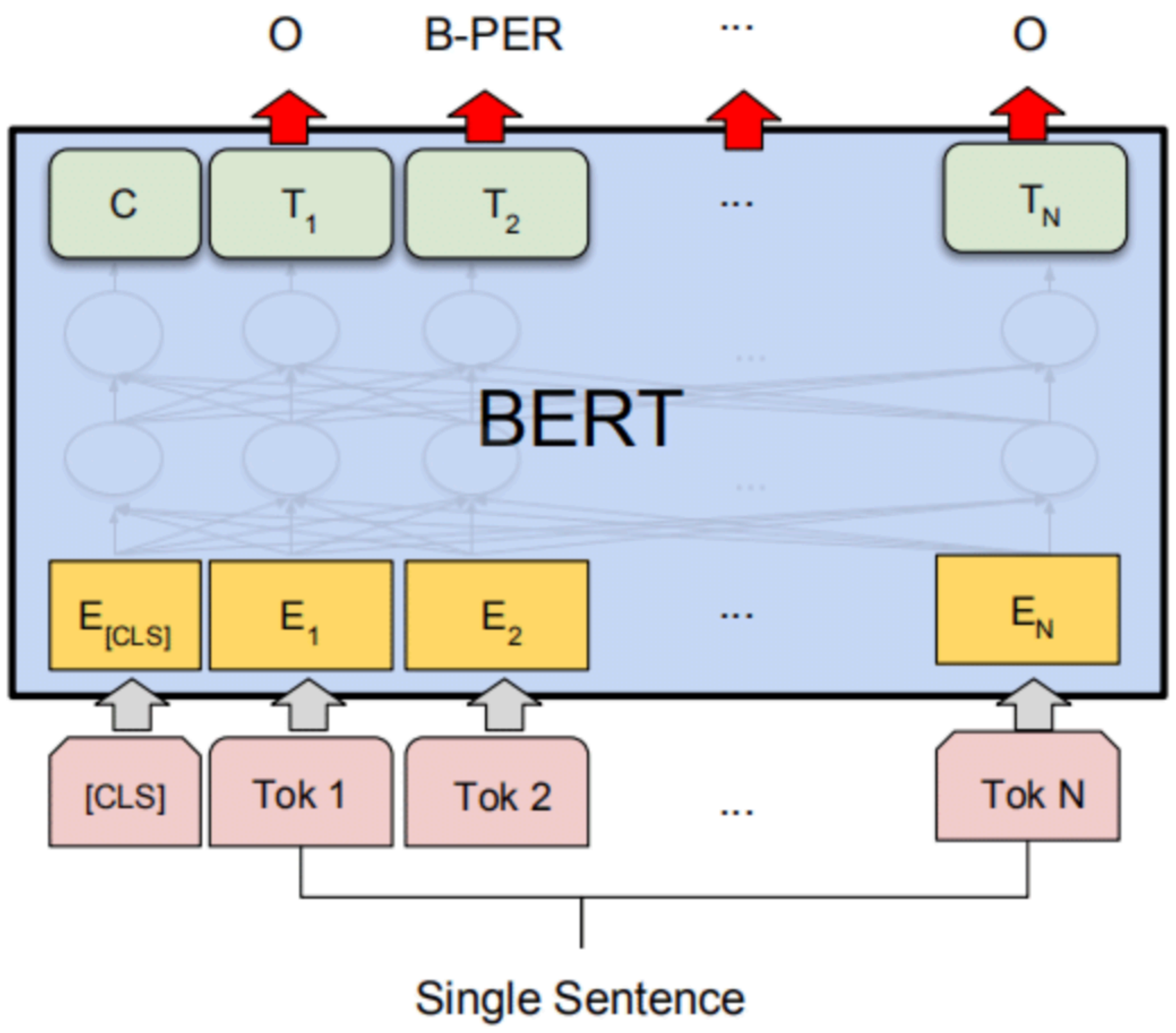
- Sequence labeling (SL) systems **tag** each word in a sentence to provide insights into the sentence structure and meaning

Example: Named Entity Recognition : **Tag** = **Entity Label**



Sequence Labeling for NLU

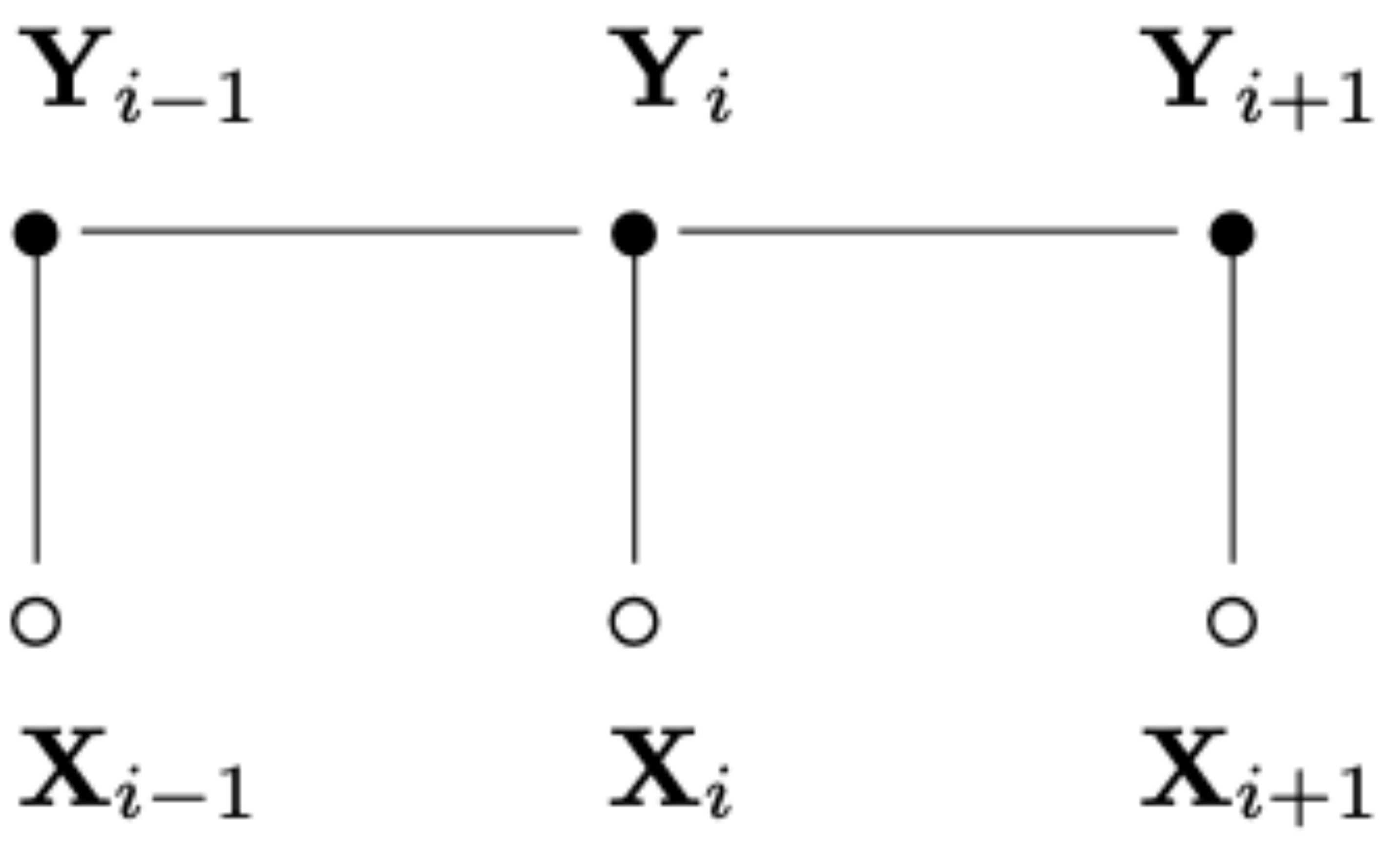
- Token Classification model
  - Current tag is conditionally independent to previous tag
  - BIO Tagging -> Named Entity can span multiple words



[2] Delvin et al., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. EMNLP 2019

Sequence Labeling for NLU

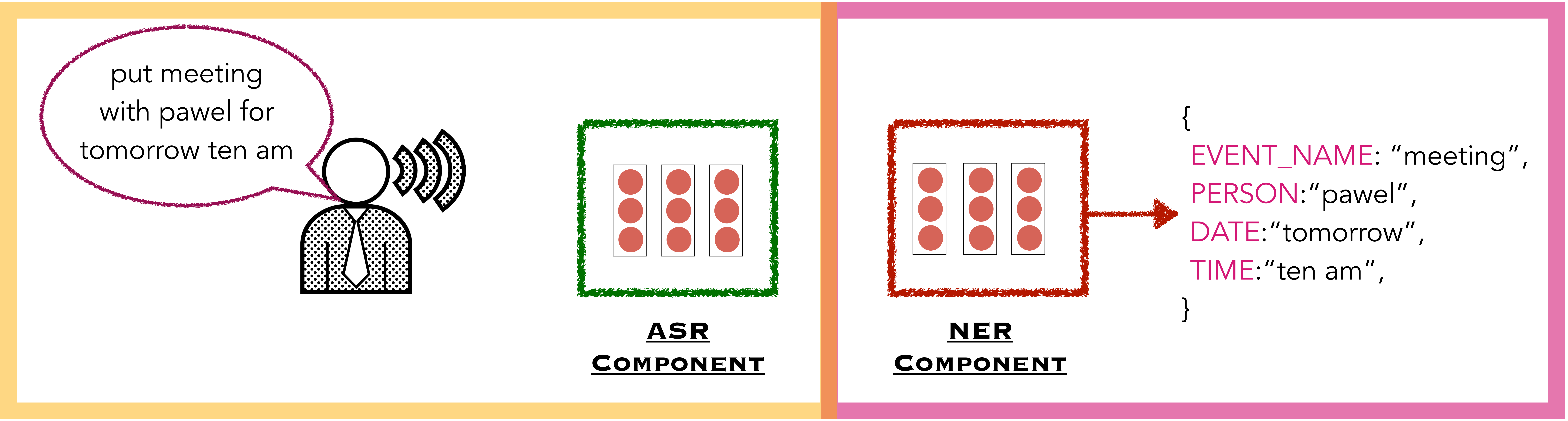
- CRF model
  - Global Normalised Loss $$P(Y|S) = \frac{e^{F(Y,S)}}{\sum_{Y' \in \mathcal{L}^N} e^{F(Y',S)}}$$
  - Global Score = Sum over all words, Emission Score + Transition Score



[3] Lafferty et al., 2001. Conditional random fields:Probabilistic models for segmenting and labeling sequence data. ICML 2001

Sequence Labeling for SLU

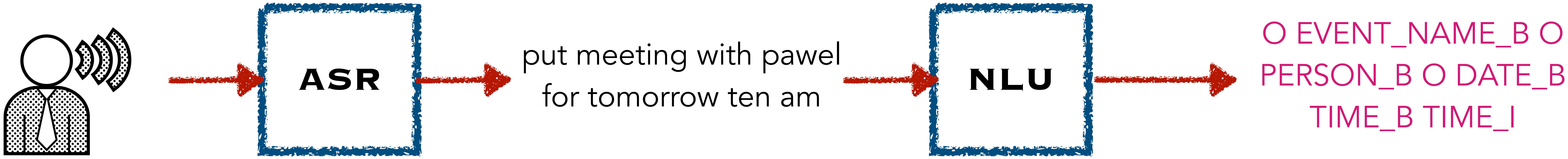
Additional complexity of recognizing the mention of the labels



Content

- Spoken Language Understanding
- Sequence Labelling
- **Current SLU Modelling**
- Compositional Models
- Composition model for Sequence Labelling in SLU

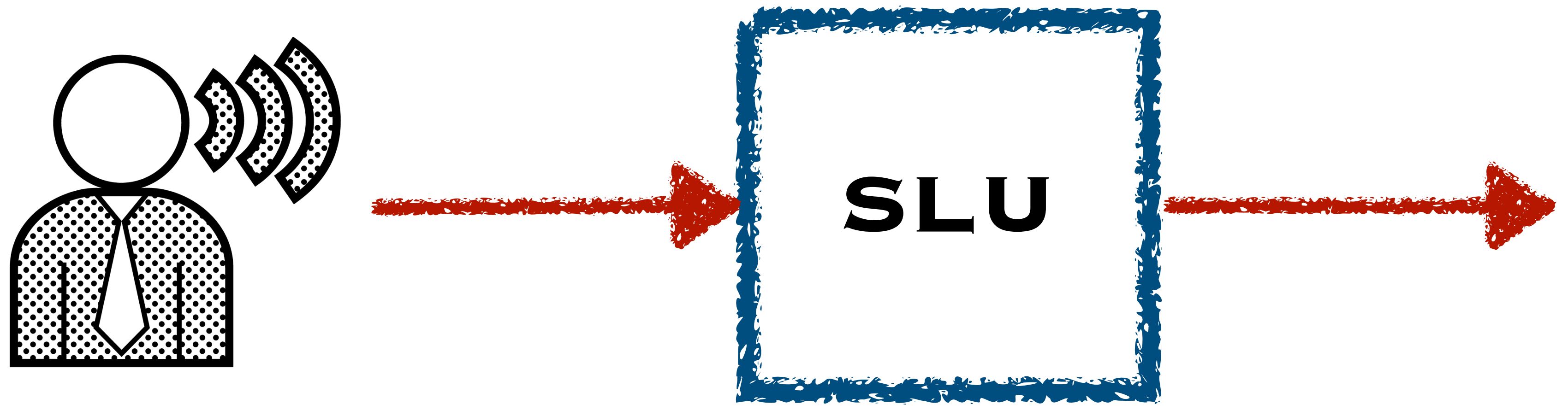
Cascaded SLU Architectures



<u>Cascaded SLU</u>

1. Advantages
    1. Utilise prior ASR and NLU Research
2. Limitations
    1. Error Propagation from ASR

E2E SLU Architectures



put EVENT_NAME FILL meeting SEP with PERSON FILL pawel SEP for DATE FILL tomorrow SEP TIME FILL ten am SEP

E2E SLU

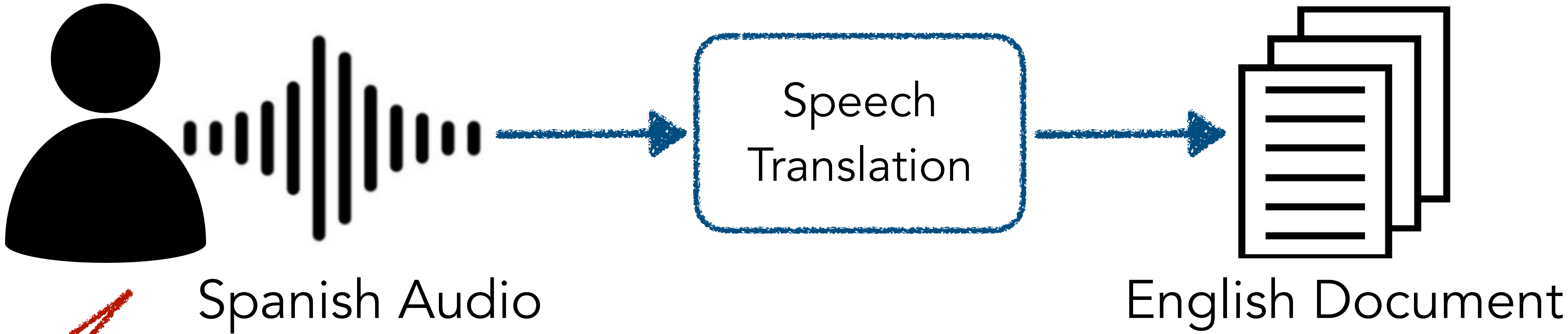
1. Advantages
   1. Avoid drawbacks of the cascaded system
   2. Simplicity
2. Limitations
   1. Cannot utilize the well studied sequence labeling framework
   2. Understanding errors made by system difficult

Content

- Spoken Language Understanding
- Sequence Labelling
- Current SLU Modelling
- **Compositional Models**
- Composition model for Sequence Labelling in SLU

What is Compositionality?

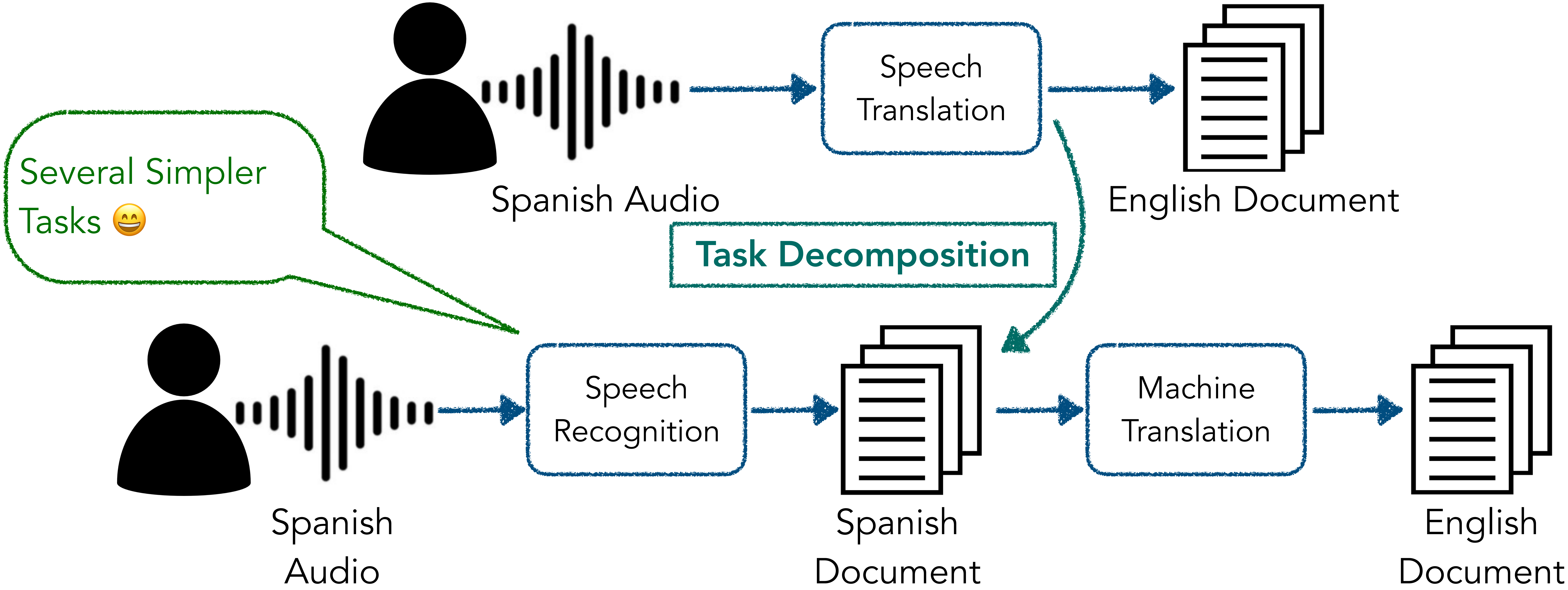
- Compositionality is the principle behind building complex systems by composing together simpler sub-systems.



Single Complex Task 😕

What is Compositionality?

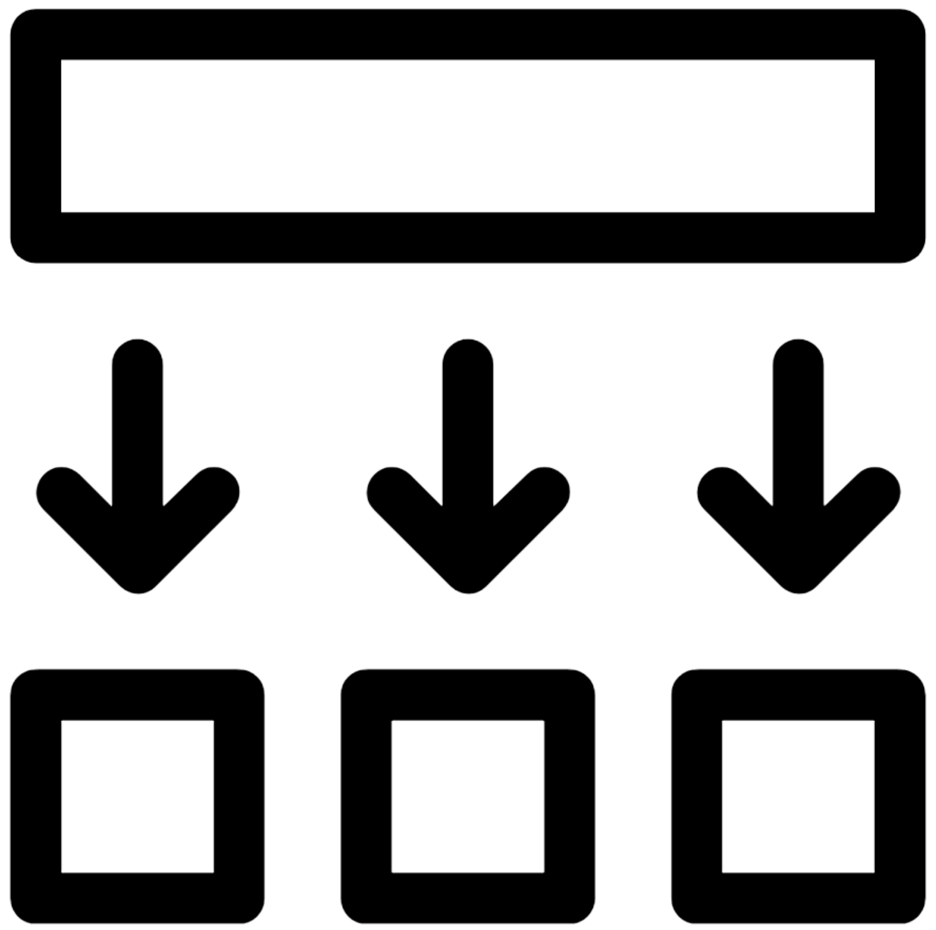
- Compositionality is the principle behind building complex systems by composing together simpler sub-systems.



Compositionality in System Building

- Compositionality takes a practical approach to system building, going from creating stand-alone systems to their large-scale development.

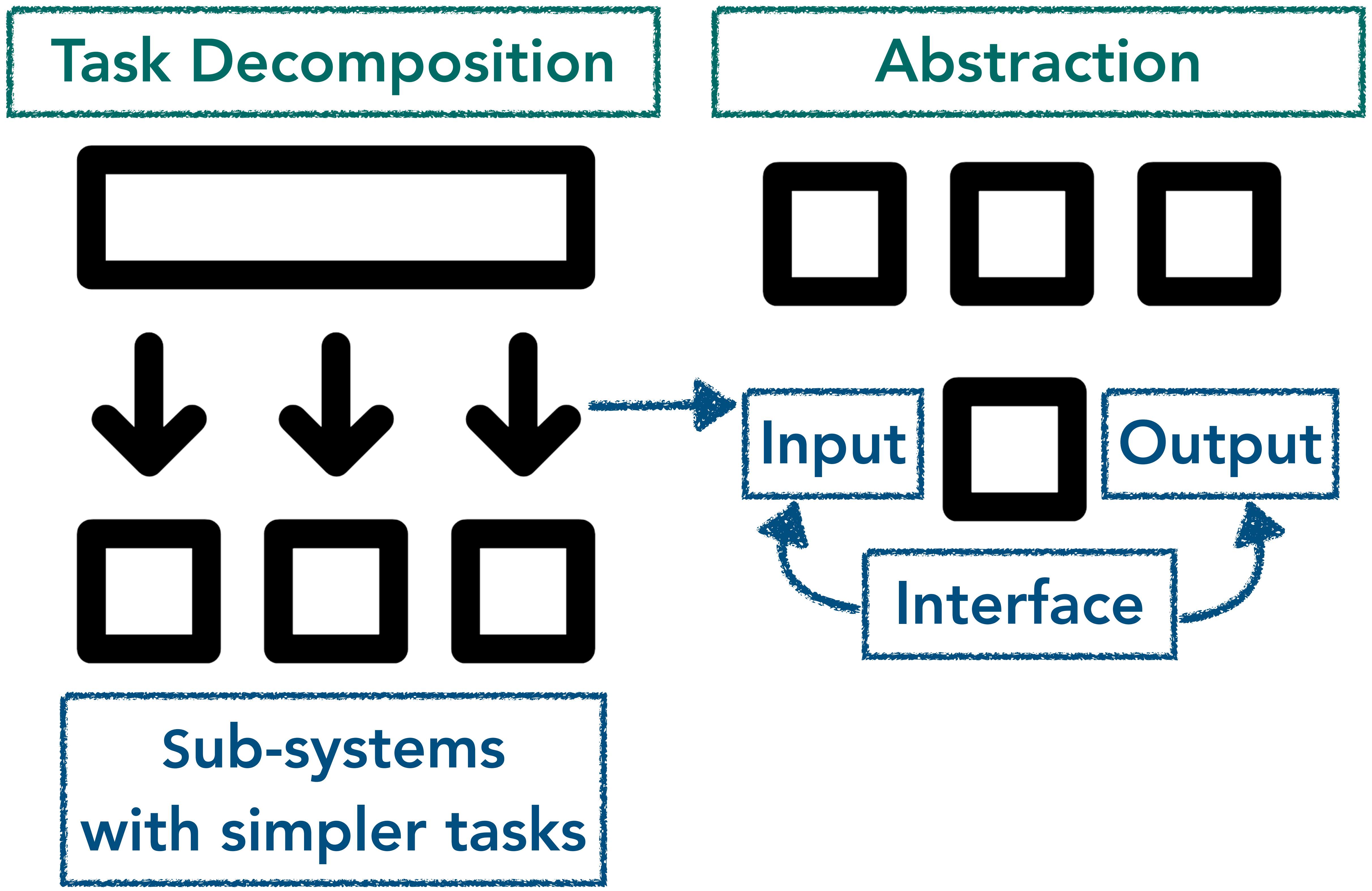
Task Decomposition



Sub-systems
with simpler tasks

Compositionality in System Building

- Compositionality takes a practical approach to system building, going from creating stand-alone systems to their large-scale development.



Compositionality in System Building

- Compositionality takes a practical approach to system building, going from creating stand-alone systems to their large-scale development.

Compositional E2E Model with Searchable Intermediates

- General end-to-end framework to exploit natural decomposition in sequence tasks.
  - A sequence task, A ➞ C is decomposable, if there is an intermediate sequence B for which A ➞ B sequence transduction followed by B ➞ C prediction achieves the original task.
    - For instance, Speech Translation or Spoken Language Understanding using ASR intermediates

Dalmia et. al., 2021

Compositional E2E Model with Searchable Intermediates

- Compositional E2E Models learns $P(\text{C} \mid \text{A})$ through decomposition;
  - $P(\text{C} \mid \text{A}) = \sum_{\text{B}} (P(\text{C} \mid \text{A}, \text{B}) P(\text{B} \mid \text{A}))$, using Sum Rule.
  - $P(\text{C} \mid \text{A}) \approx \max_{\text{B}} (P(\text{C} \mid \text{A}, \text{B}) P(\text{B} \mid \text{A}))$, approximated with Viterbi.
  - This allows the use of traditional formulations for building $P(\text{B} \mid \text{A})$ and $P(\text{C} \mid \text{B})$.

Compositional E2E Model with Searchable Intermediates

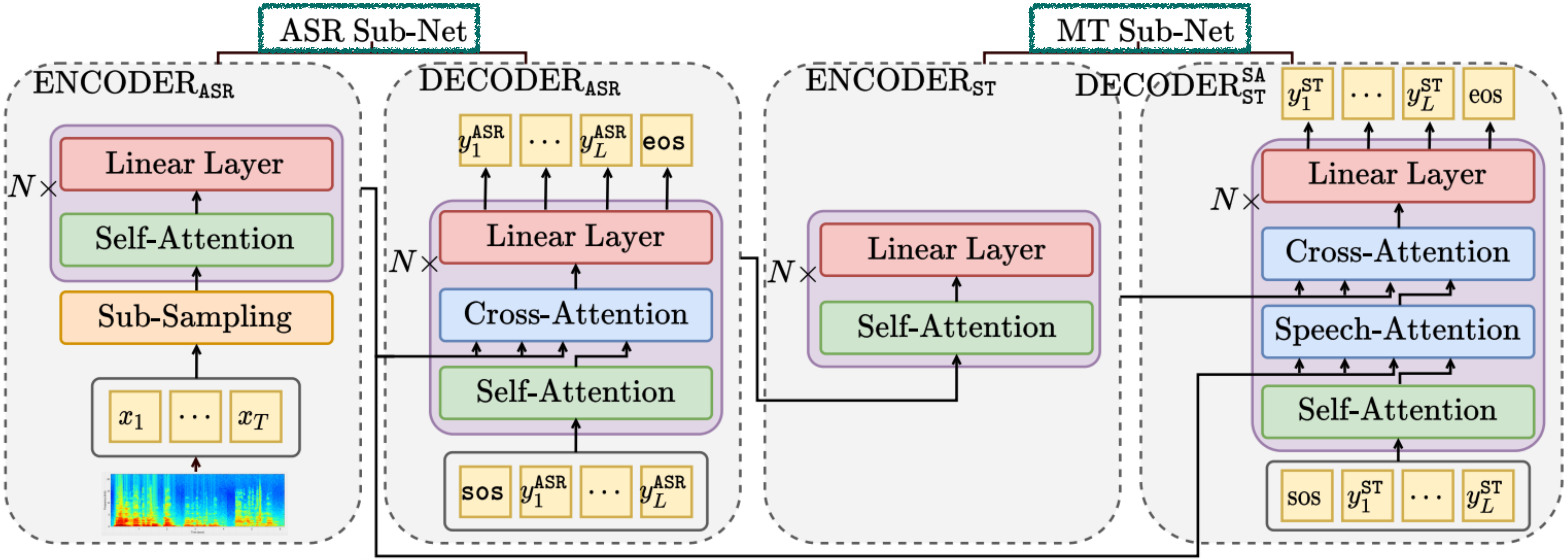
- Compositional E2E Models learns $P(\text{C} \mid \text{A})$ through decomposition;
    - $P(\text{C} \mid \text{A}) = \sum_{\text{B}} (P(\text{C} \mid \text{A, B})P(\text{B} \mid \text{A}))$, using Sum Rule.
    - $\boxed{P(\text{C} \mid \text{A})} \approx \max_{\text{B}}(\boxed{P(\text{C} \mid \text{A, B})}\boxed{P(\text{B} \mid \text{A})})$, approximated with Viterbi.
    - This allows the use of traditional formulations for building $P(\text{B} \mid \text{A})$ and $P(\text{C} \mid \text{B})$.

Speech Translation	Text-Based Machine Translation Formulation	Speech Recognition Formulation
ST	MT	ASR

Compositional E2E Model with Searchable Intermediates

- The Compositional E2E Model with Searchable Intermediates has three main focus -
  - Simplify learning process by decomposing tasks, while maintaining end-to-end differentiability.
  - Utilize existing and well-studied Speech and NLP formulations in building complex sequence tasks.
  - Add Component-level Search Capabilities with an Intermediate Decoder.

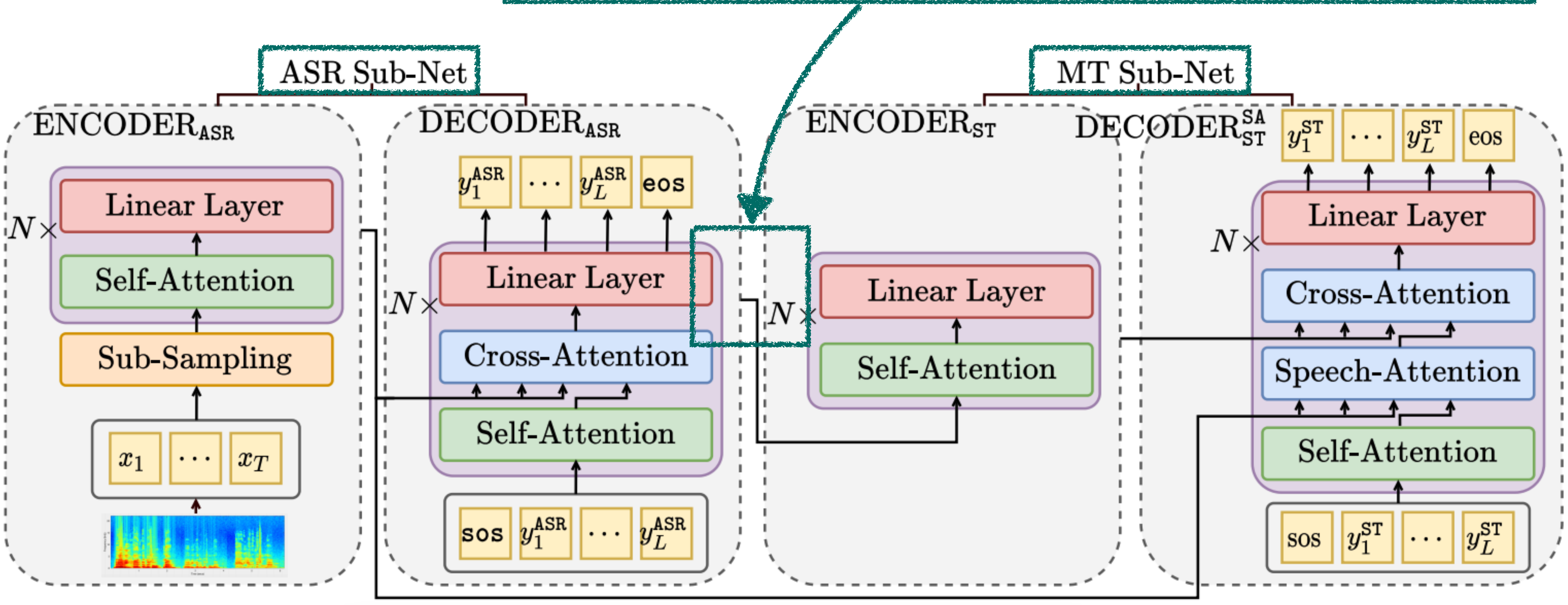
Multi-Decoder Model with Searchable Intermediates



Multi-Decoder Model with Searchable Intermediates

Pass Decoder Hidden Representations:

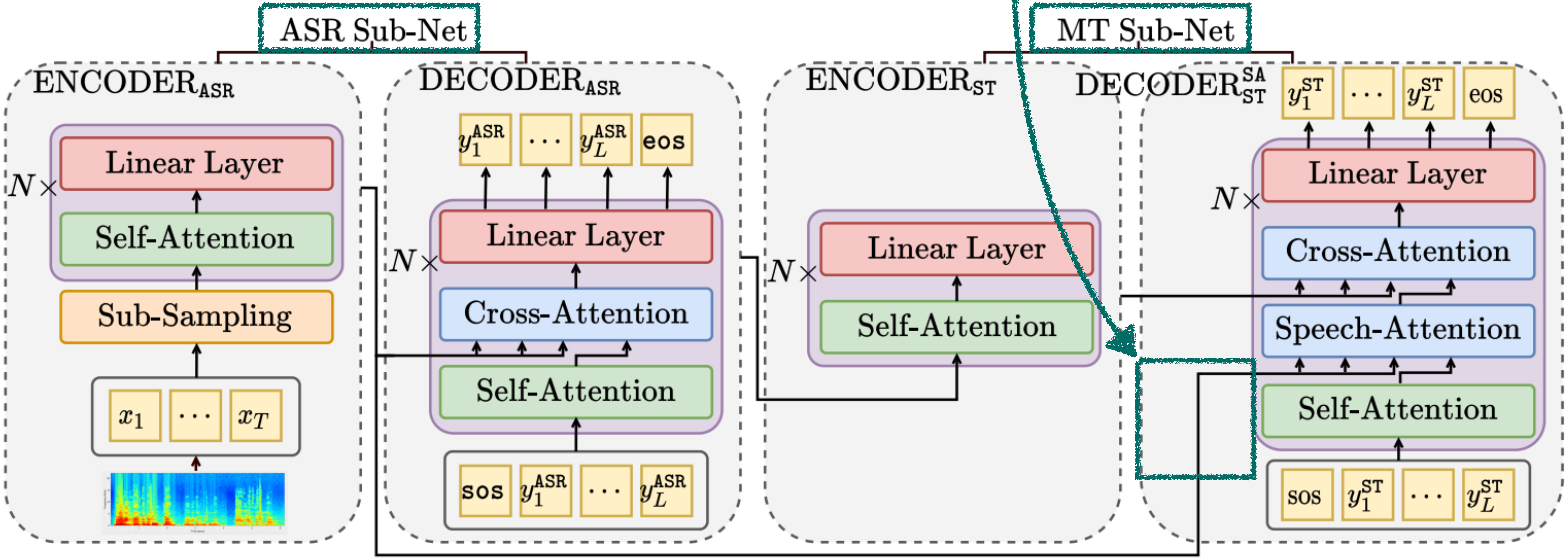
- ASR Sub-Net maps input to sequence of decoder hidden representations $\mathbf{h}^{D_B}$
- MT Sub-Net maps $\mathbf{h}^{D_B}$ to final ST output



Multi-Decoder Model with Searchable Intermediates

Cross Speech Attention:

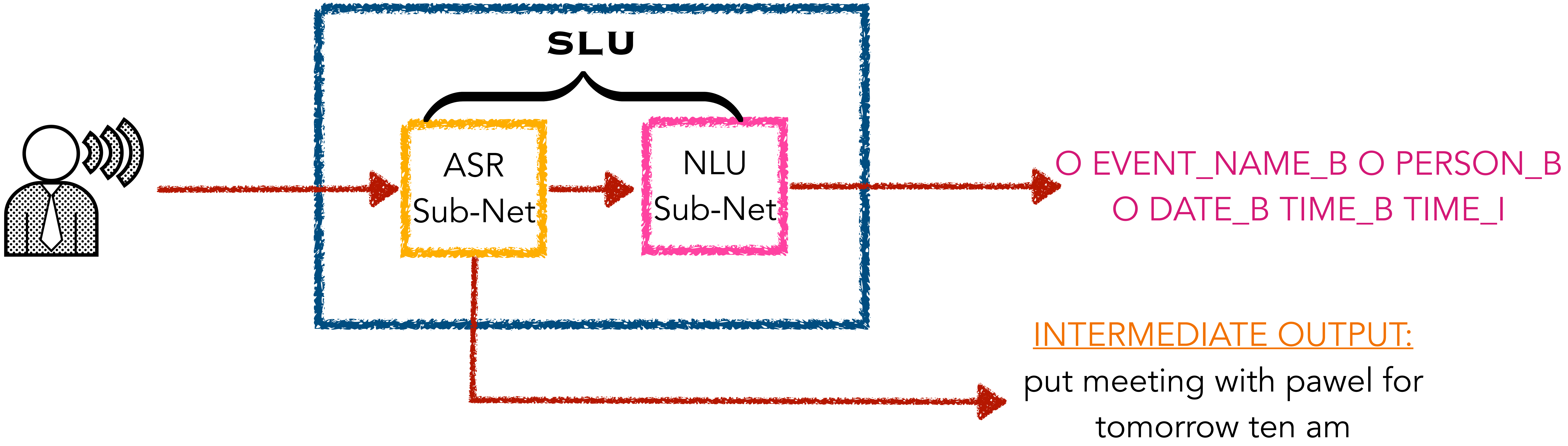
- Conditions on speech information via ASR encoder
- During inference, approximate $\mathbf{h}^{D_B}$ with $\mathbf{h}^{D_B}_{\text{Beam}}$



Content

- Spoken Language Understanding
- Sequence Labelling
- Current SLU Modelling
- Compositional Models
- **Composition model for Sequence Labelling in SLU**

Desired Compositional E2E SLU Architecture



Compositional E2E SLU

Inspired by the principles of task compositionalty in SL for SLU, we seek to bring both schools of thought together

Our Contributions-

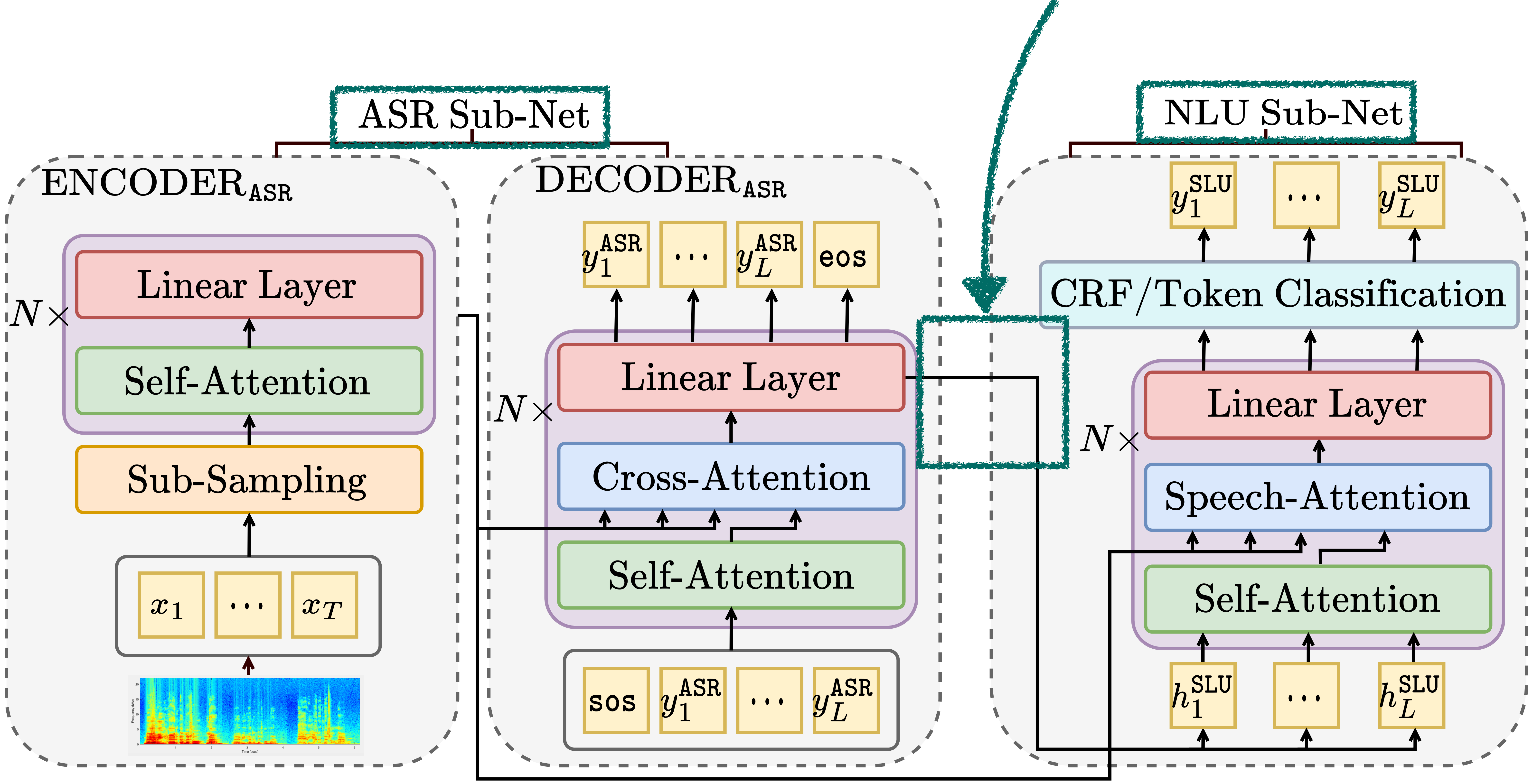
1. Build compositional SLU using searchable intermediate framework [2] that
   - Convert spoken utterance to sequence of token representations -> ASR Subnetwork
   - Train token classification network -> NLU Subnetwork
2. Conditioning token-wise classification on speech allows recovery from errors

[4] Dalmia et al., 2021. Searchable hidden intermediates for end-to-end models of decomposable sequence tasks. NAACL 2021

Compositional E2E SLU Model with Searchable Intermediates

The ASR Sub-Net:

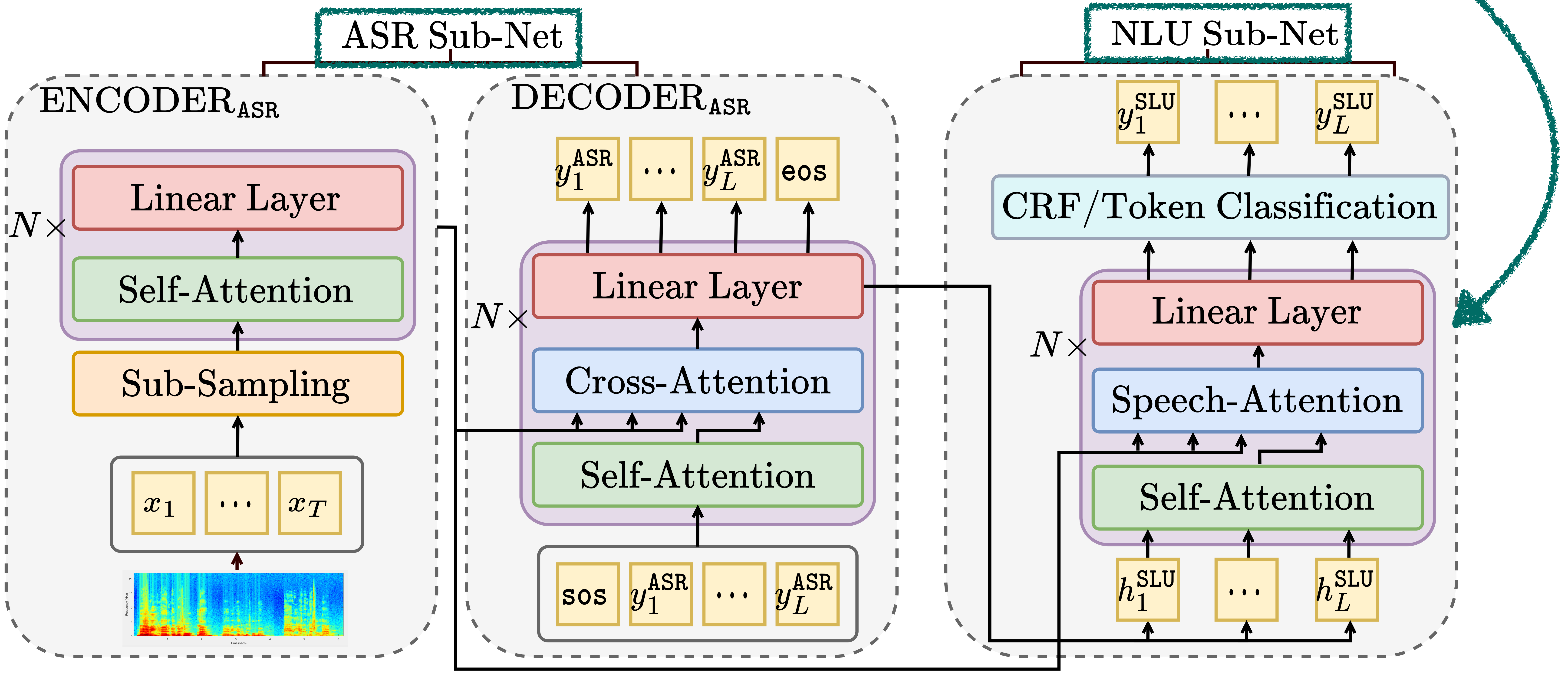
- Maps input to sequence of decoder hidden representations $\mathbf{h}^{D_B}$.
- Passes $\mathbf{h}^{D_B}$ to NLU Sub-Net



Compositional E2E SLU Model with Searchable Intermediates

The NLU Sub-Net:

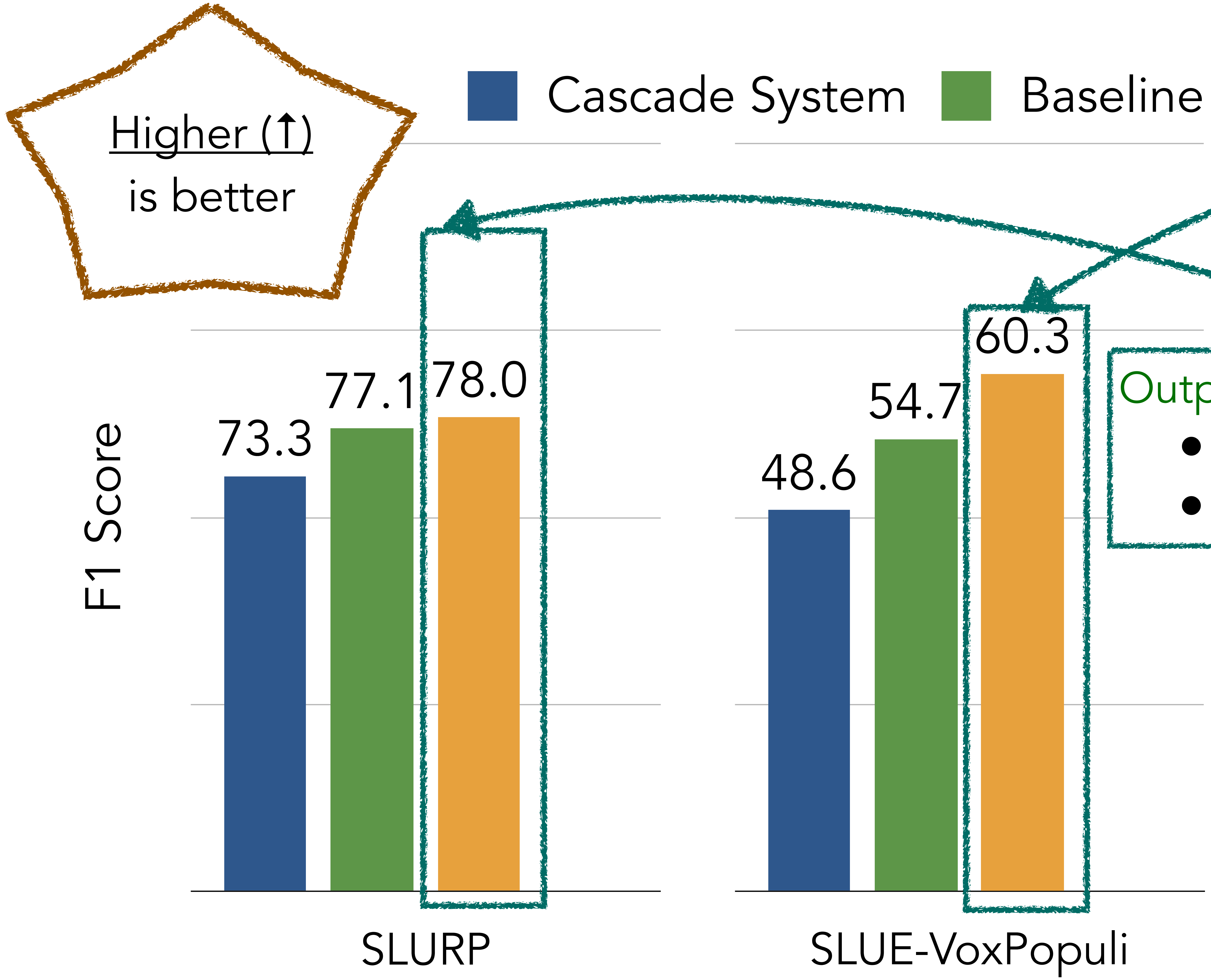
- Uses the sequence of decoder hidden representations $\mathbf{h}^{D_B}$. This makes the length of the NLU sequence known.
- Allowing, token level sequence labeling formulation!
- Can also use globally normalized loss like CRF.



Experimental Setup

1. Task: **Named Entity Recognition**
2. Dataset
    1. SLURP Dataset
    2. SLUE Dataset
3. Models
    1. Baseline
        1. Cascaded SLU
        2. E2E SLU
    2. **Compositional E2E SLU**
        1. **Proposed NLU formulation**
            1. CRF
            2. **Token Classification**
                1. w/o Speech Attention (Ablation)
    3. Pretraining
        1. ASR - Gigapeech dataset
        2. LM - Canine

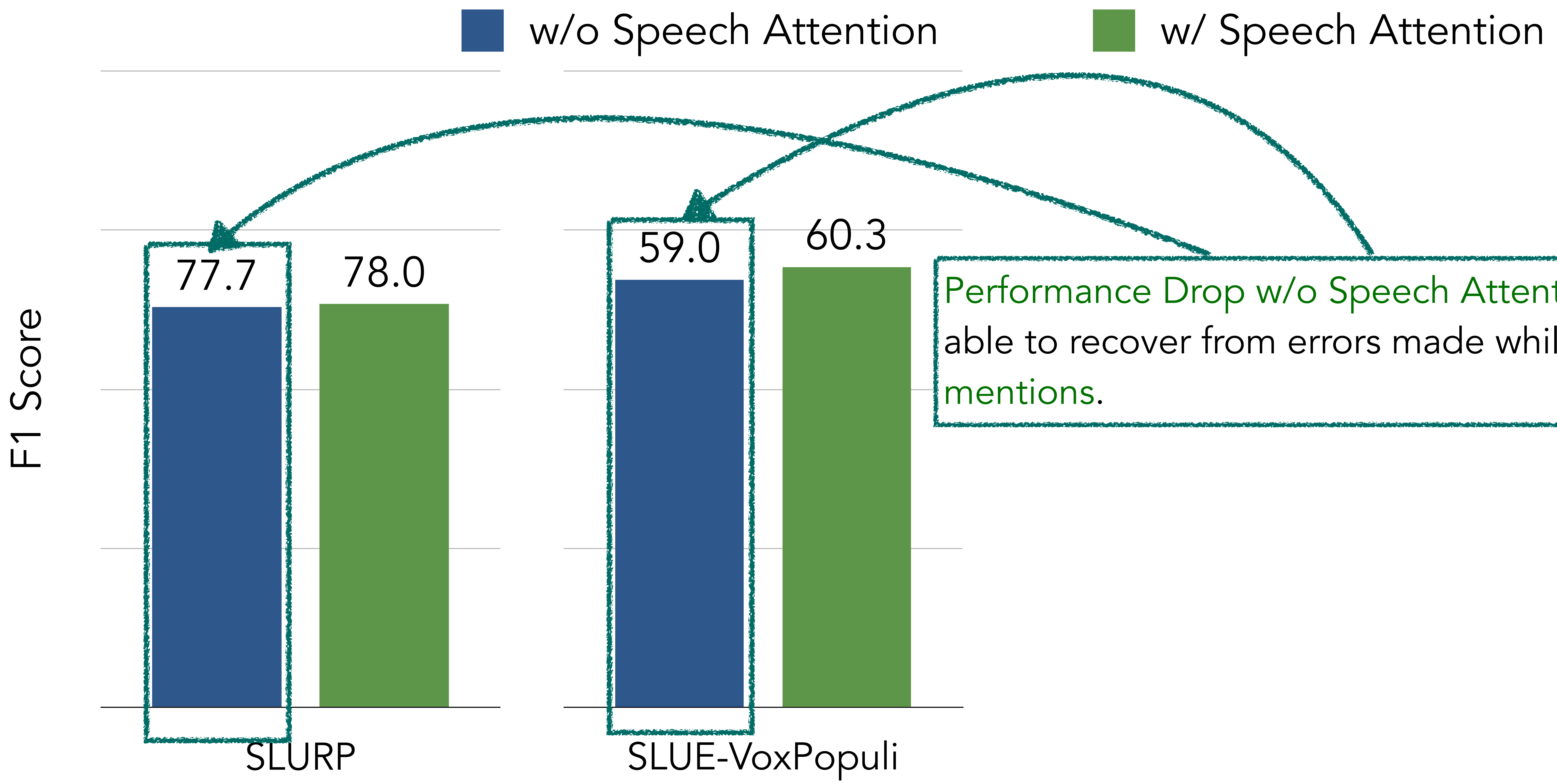
Comparison with Encoder-Decoder



Outperforms both Encoder-Decoder and Cascaded Models -

- +4 F1 and +1 F1 on SLURP
- +12 F1 and +6 F1 on SLUE-VoxPopuli

Using Cross Speech Attention

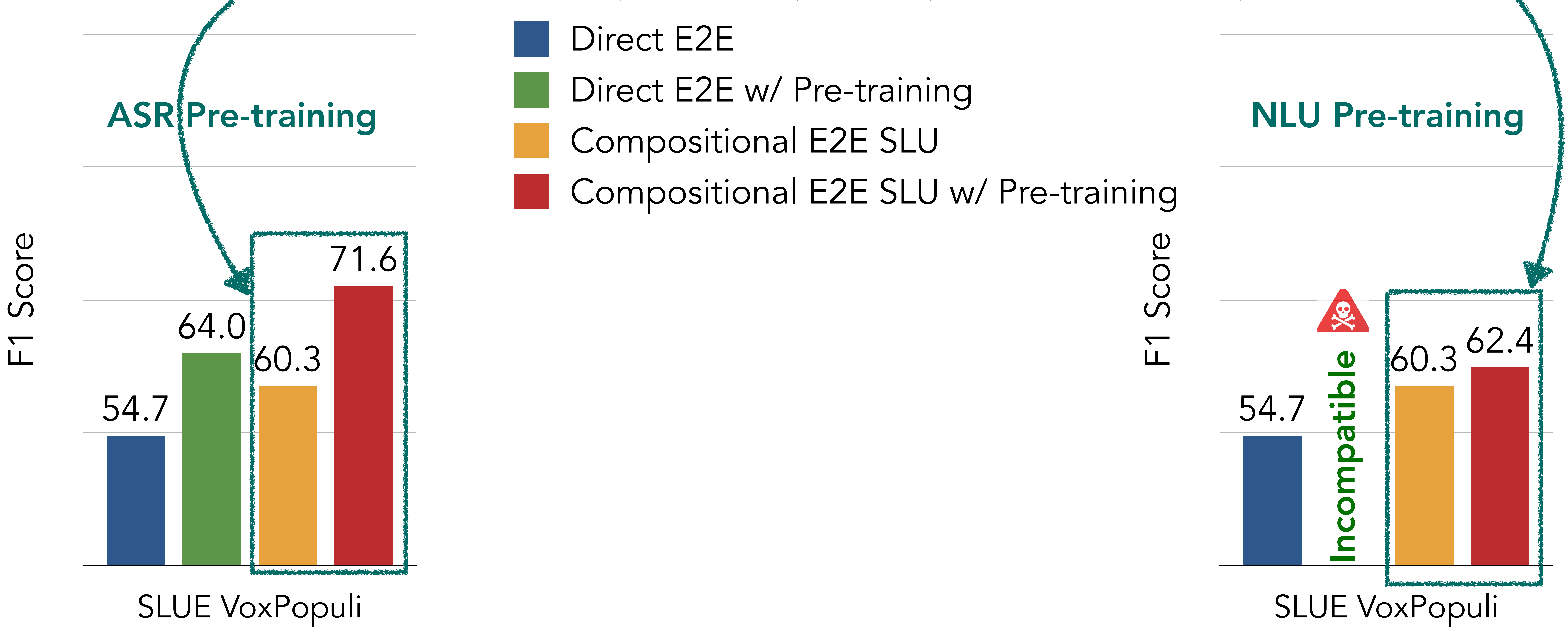


Performance Drop w/o Speech Attention as model is not able to recover from errors made while recognising entity mentions.

Using Pre-trained Subtask Models

Resource Pooling

Each sub-task model can be pre-trained due to the decomposed functionality



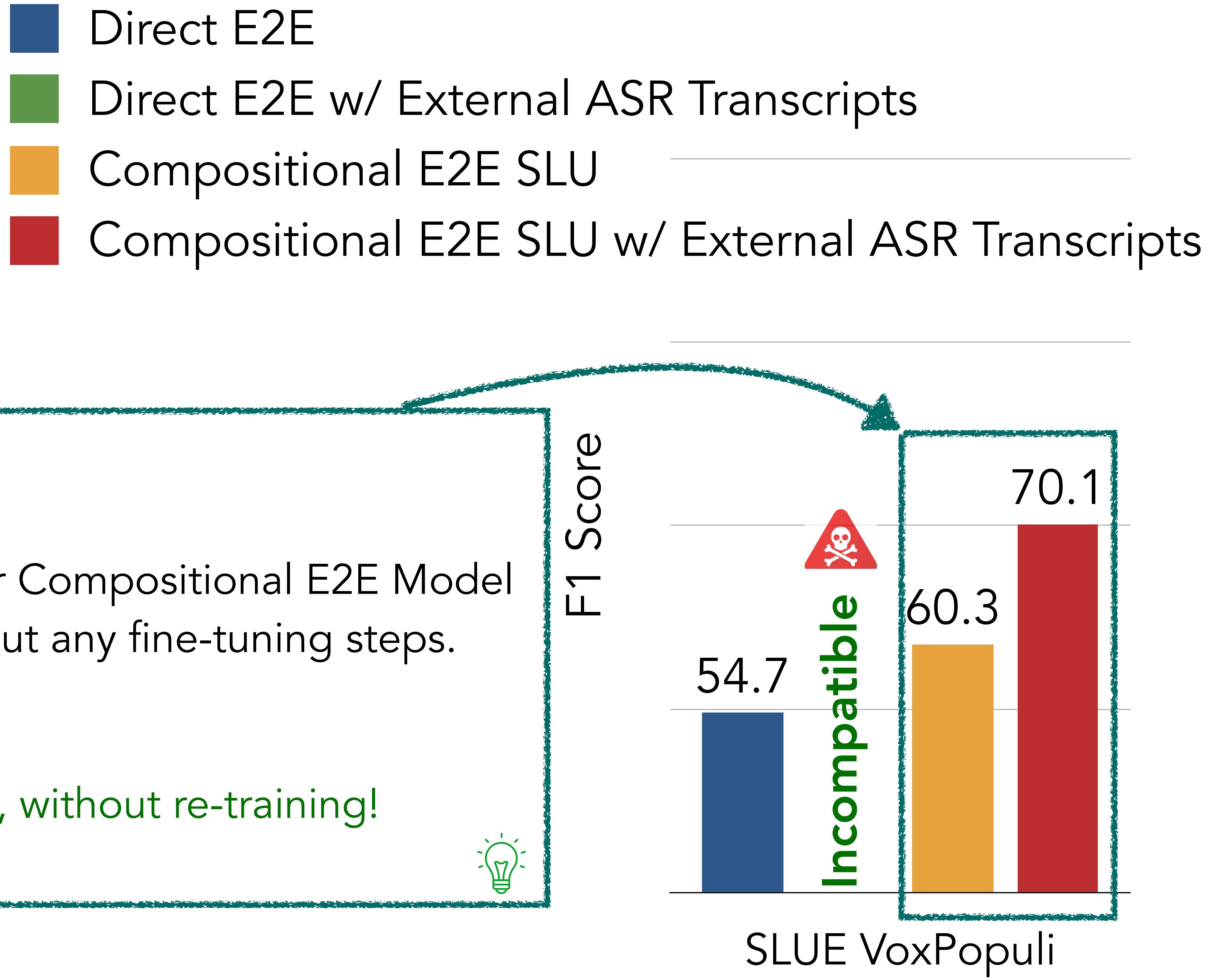
Using Pre-trained Subtask Models

Guiding Intermediate Representations

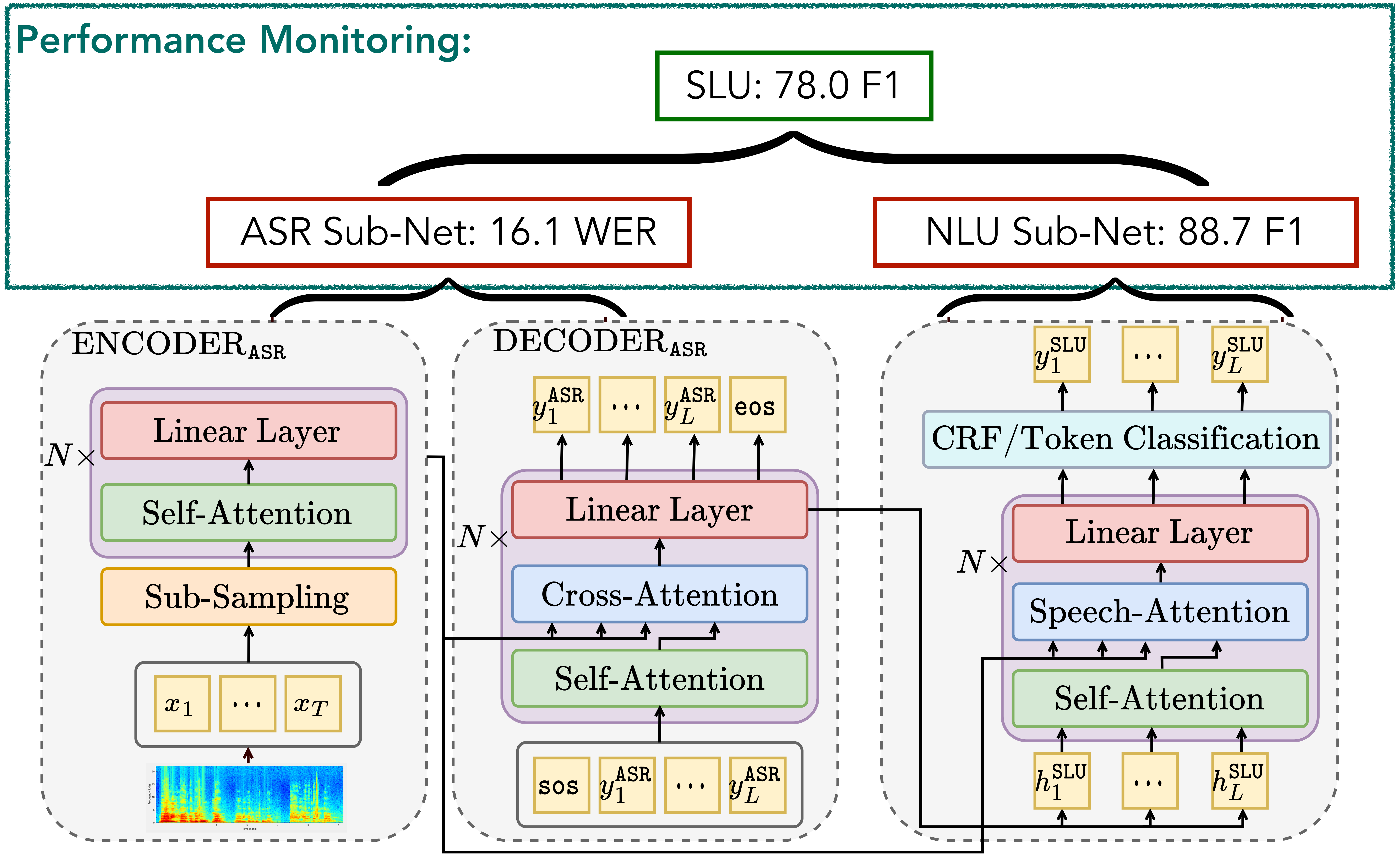
Resource Pooling

We can guide the intermediate representations in our Compositional E2E Model using external sub-net models during inference without any fine-tuning steps.

Performance on SLUE Voxpopuli improves by +10 F1, without re-training!



Performance Monitoring



Error Categorisation

	Hypothesis	Reference
ASR Correct Entity Correct	event reminder mona [EVENT] tuesday [DATE]	event reminder mona [EVENT] tuesday [DATE]
ASR Correct Entity Incorrect	is there anything happening on jazz scene [MOVIE TYPE] around edinburgh [NEWS TOPIC]	is there anything happening on jazz scene [MOVIE TYPE] around edinburgh [PLACE NAME]
ASR Incorrect Entity Correct	create meeting [EVENT NAME] with paul [PERSON] for tomorrow [DATE] at ten am [TIME]	put meeting [EVENT NAME] with pawel [PERSON] for tomorrow [DATE] ten am [TIME]
ASR Incorrect Entity Incorrect	set a birthday [EVENT NAME] event for ninety [DATE]	set a birthday [EVENT NAME] event for martin [PERSON]

One-to-one alignment between ASR and Sequence Labelling help Error Categorisation

- Not possible in E2E Systems

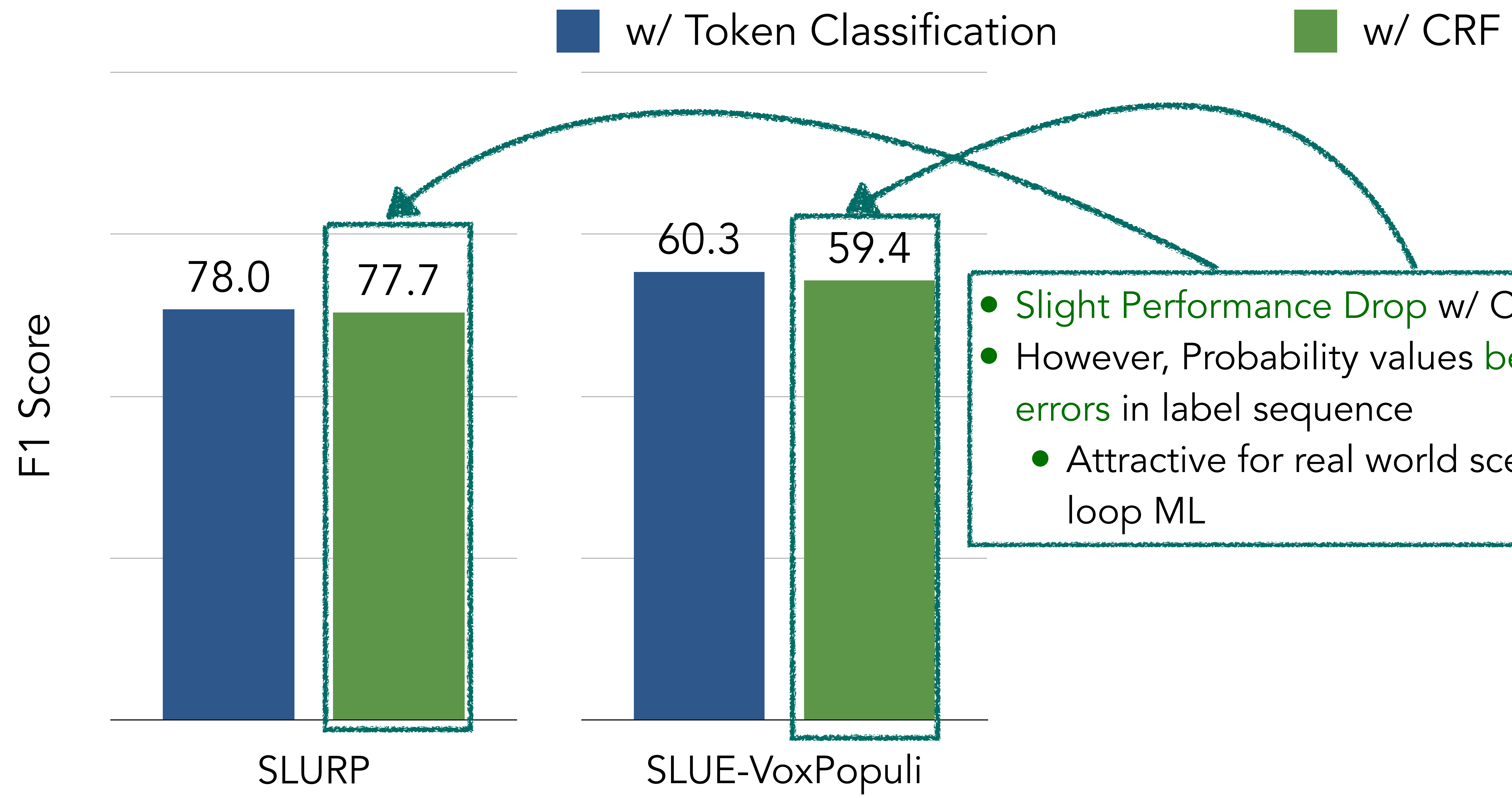
Error Categorisation

	Entity Correct		Entity Incorrect	
	Model	# Examples	Model	# Examples
ASR Correct	w/ SA	8520	w/ SA	465
	w/o SA	8501	w/o SA	474
ASR Incorrect	w/ SA	1568	w/ SA	1343
	w/o SA	1585	w/o SA	1336

Performance Difference w/ Speech Attention caused mainly by the errors where ASR inaccurate, but the NLU module is nevertheless able to recover the correct entity

- Confirms Intuition
- Transparency useful for practitioners to debug model

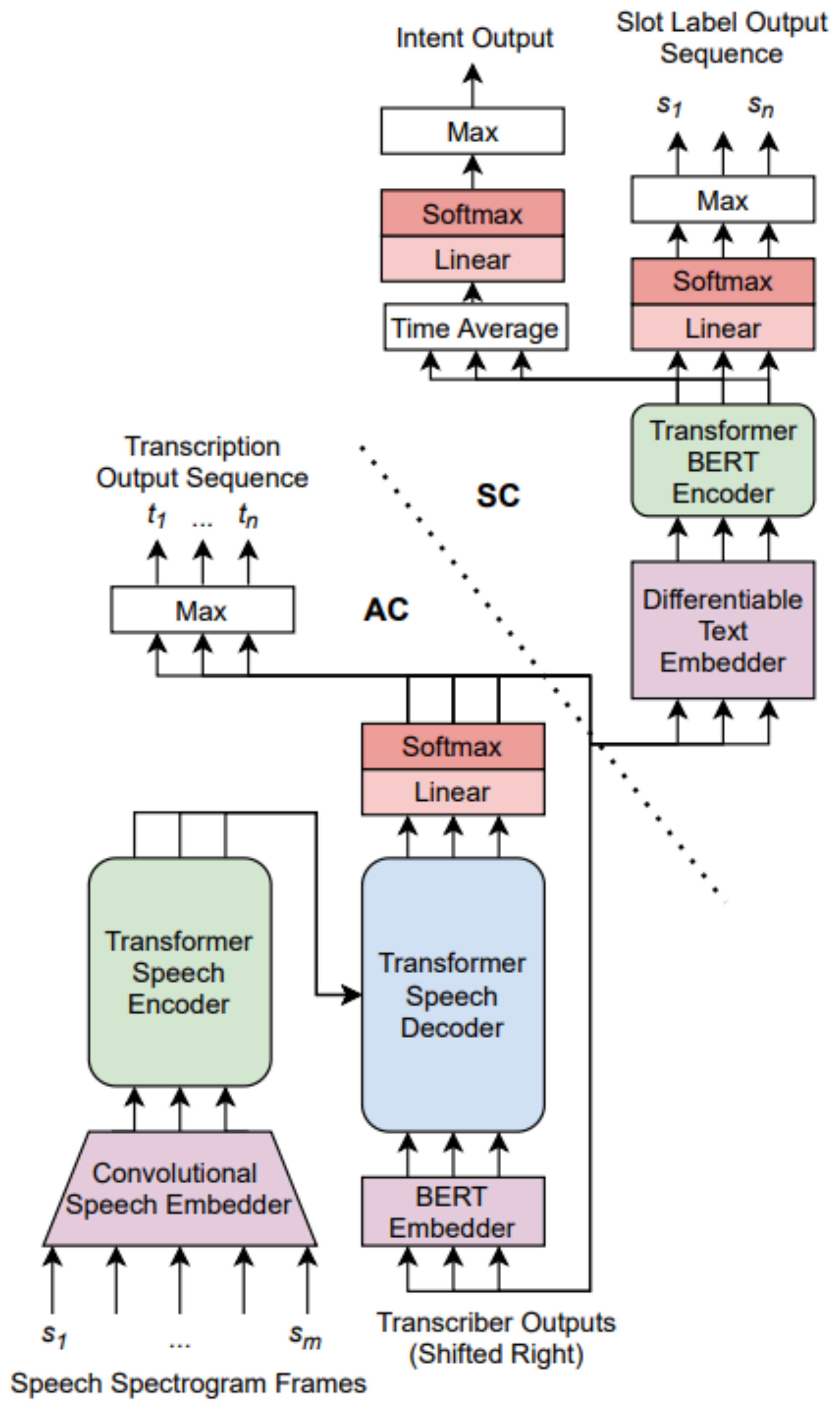
Globally Normalised Losses



- Slight Performance Drop w/ CRF
- However, Probability values better correlated with errors in label sequence
  - Attractive for real world scenarios like human in the loop ML

Related Studies

Discrete outputs from the ASR module that are made differentiable using various approaches like Gumbel-softmax

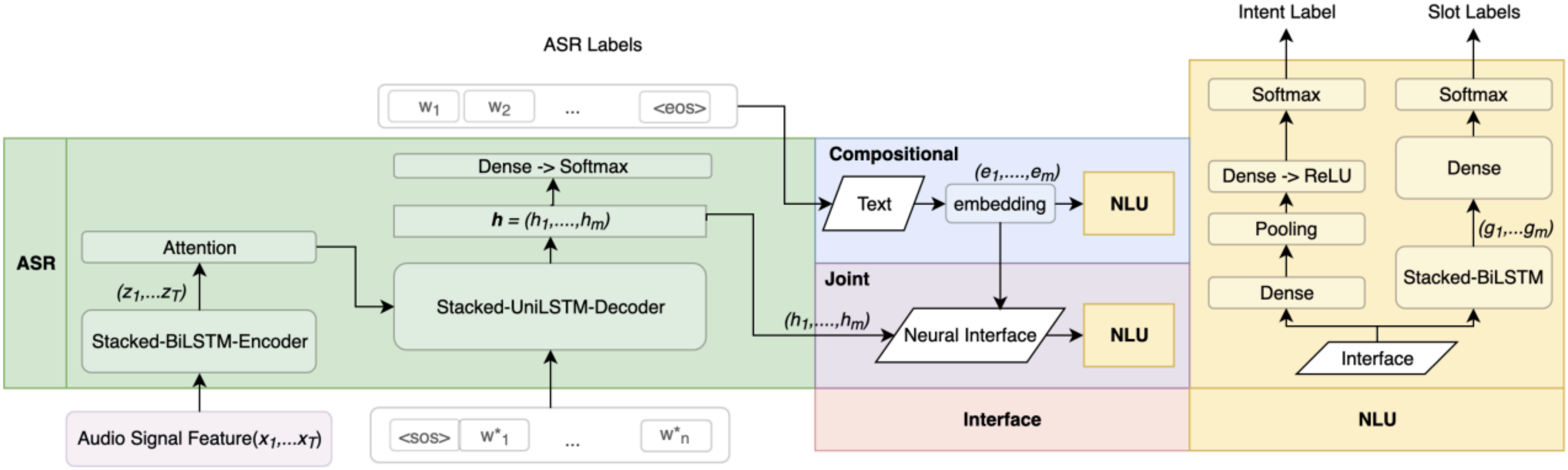


[5] Saxon et al., 2021. End-to-end spoken language understanding for generalized voice assistants. Interspeech 2021.

Related Studies

Uses the ASR decoder hidden representations in the NLU module by concatenating with token embeddings of the ASR discrete output.

- Requires the ASR and NLU submodule to have a shared vocabulary space, limiting usage of pretrained models.



[6] Rao et al. 2020. Speech to semantics: Improve ASR and NLU jointly via all-neural interfaces. Interspeech 2020.

Evaluating End-to-End Systems for Decomposable Tasks

Research Objectives

By exploiting compositionality, can we build benchmark test sets for a dataset that evaluates different portions of end-to-end model?

Case Study: SLU

1. Framework to construct robust test sets using coordinate ascent over sub-task specific utility functions.
2. Given a dataset for a decomposable task, optimally create test sets for each sub-task to individually assess components of the end-to-end model.
    1. One assessing natural language understanding abilities, and
    2. One to test speech processing skills.

[8] Arora et al. 2021. Evaluating End-to-End Systems for Decomposable Tasks. Interspeech 2021.

Conclusion

Compositional model combines the powers of 2 school of thoughts

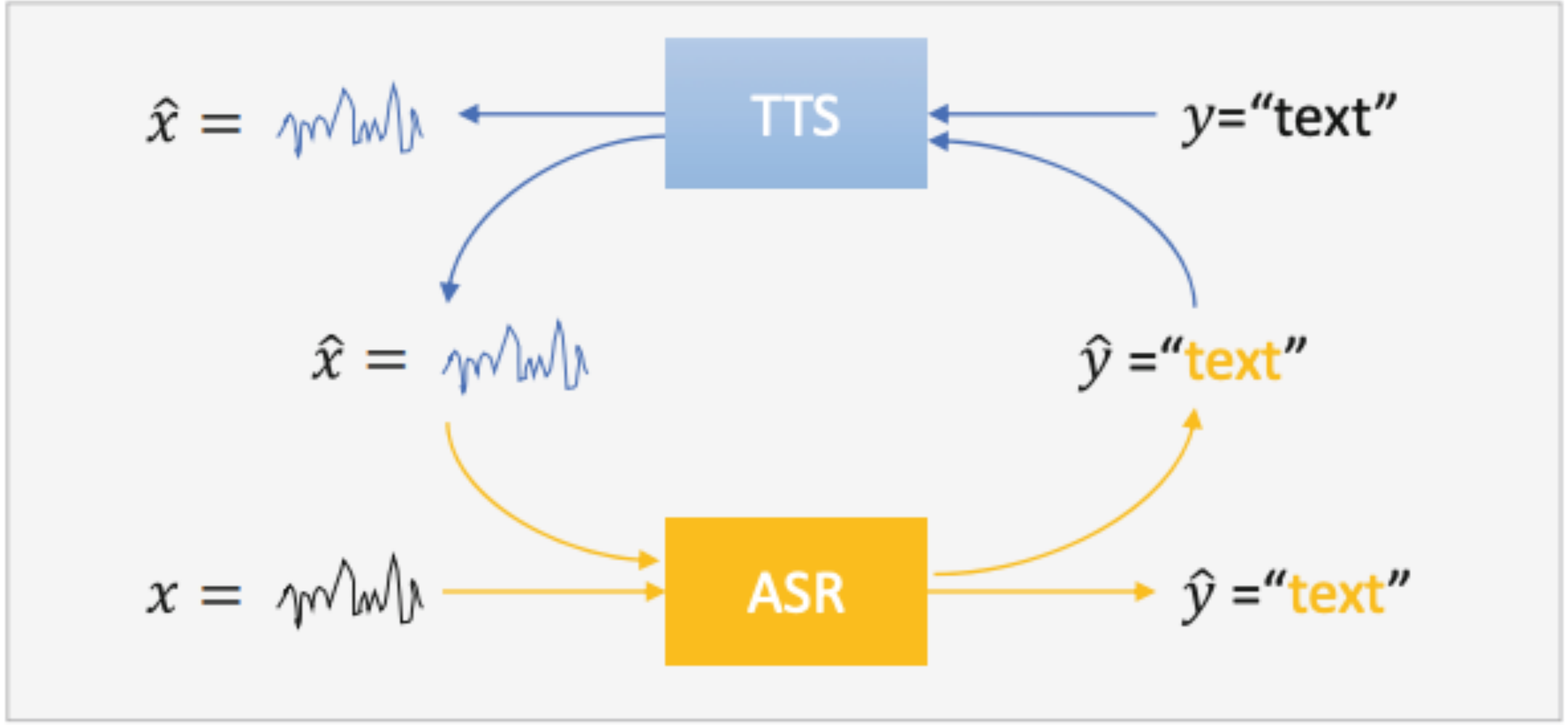
- No Error Propagation
- Better Compatibility with Pretrained Models
- Better Transparency

Higher Performance (↑)

Future Directions

I hope this thesis encourages researchers to -

- Build compositionality inspired neural architectures -
    - Can be extended to other decomposable tasks like Visual QA
    - Can also be used in Dual Learning Framework like Cyclic ASR-TTS Systems
        - Achieve End to End Differentiability using Compositional Model



Future Directions

I hope this thesis encourages researchers to -

- Build flexible tokenization for easy composition of systems -
    - If one token distribution can be converted into another token distribution; for example BPE 100 to 2000,
    - Avoid system interactions in surface text, allowing utilization of additional information like entropy of the prediction!.
- Extend compositional E2E systems to streaming applications

Questions

Thanks for Watching

Dataset & Code : https://github.com/espnet/espnet

(Issues and contribution welcome)