

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

Siddhant Arora\* Siddharth Dalmia\* Brian Yan  
Florian Metze Alan W Black Shinji Watanabe

Language Technologies Institute, Carnegie Mellon University, USA  
{siddhana, sdalmia}@cs.cmu.edu

## Abstract

End-to-end spoken language understanding (SLU) systems are gaining popularity over cascaded approaches due to their simplicity and ability to avoid error propagation. However, these systems model sequence labeling as a sequence prediction task causing a divergence from its well-established token-level tagging formulation. We build compositional end-to-end SLU systems that explicitly separate the added complexity of recognizing spoken mentions in SLU from the NLU task of sequence labeling. By relying on intermediate decoders trained for ASR, our end-to-end systems transform the input modality from speech to token-level representations that can be used in the traditional sequence labeling framework. This composition of ASR and NLU formulations in our end-to-end SLU system offers direct compatibility with pre-trained ASR and NLU systems, allows performance monitoring of individual components and enables the use of globally normalized losses like CRF, making them attractive in practical scenarios. Our models outperform both cascaded and direct end-to-end models on a labeling task of named entity recognition across SLU benchmarks.<sup>1</sup>

## 1 Introduction

Sequence labeling (SL) is a class of natural language understanding (NLU) tasks. These systems *tag* each word in a sentence to provide insights into the sentence structure and meaning (Jurafsky and Martin, 2009). An SL system that processes unstructured text, first encodes the context and relationships of words in the sentence using an encoder and then labels each token (Lample et al., 2016; Dozat et al., 2017; Akbik et al., 2018). However, when dealing with spoken utterances, sequence labeling introduces an additional complexity of

also *recognizing* the mentions of the labels (Kubala et al., 1998; Zhai et al., 2004).

SL in spoken language understanding (SLU) has been approached by two schools of thought, (1) that seek to recognize the spoken words using an Automatic Speech Recognition (ASR) engine and then tag the mentions using an NLU engine in a cascaded manner (Palmer and Ostendorf, 2001; Horlock and King, 2003; Béchet et al., 2004), and (2) that seek to recognize and tag the mentions directly from speech in an end-to-end (E2E) framework (Arora et al., 2022; Ghannay et al., 2018). Prior work has shown that cascaded systems suffer due to error propagation (Tran et al., 2018) from the ASR into the NLU engine, which can be overcome in an E2E framework. However, unlike cascaded models, E2E systems cannot utilize the vast abundance of NLU research (Shon et al., 2022) as they re-define the SL problem as a complex sequence prediction problem where the sequence contains both the tags and its mentions.

Inspired by the principles of task compositionality in SL for SLU, we seek to bring both schools of thought together. Our conjecture is that we can build compositional E2E systems that first convert the spoken utterance to a sequence of token representations (Dalmia et al., 2021), which can then be used to train token-wise classification systems as per the NLU formulation. By also conditioning our token-wise classification on speech, our compositional E2E system allows recovery from errors made while creating token representations. We instantiate our formulation on a popular SL task of named entity recognition (NER) and (1) present the efficacy of our compositional E2E NER-SLU system on benchmark SLU datasets (Bastianelli et al., 2020; Shon et al., 2022) surpassing both the cascaded and direct E2E systems §5.2. (2) Our compositional model consists of ASR and NLU components compatible with pre-trained ASR and NER-NLU models §5.3. (3) Our E2E systems ex-

<sup>1</sup>Our code and models are publicly available as part of the ESPnet-SLU toolkit: <https://github.com/espnet/espnet>. \*Equal Contribution. Siddharth is now at Google.

hibit transparency towards categorizing errors by enabling the evaluation of individual components of our model in isolation §5.4.

The paper first describes the traditional SL formulation (§2), and discusses shortcomings in current SLU formulations (§3). Section §4 presents our compositional E2E model that can overcome these shortcomings. We then evaluate these approaches towards the SL task of NER (§5).

## 2 Sequence Labeling (SL)

SL systems tag each word,  $w_i$ , of a text sequence,  $S = \{w_i \in \mathcal{V} | i = 1, \dots, N\}$  of length  $N$  and vocabulary  $\mathcal{V}$ , with a label from a label set  $\mathcal{L}$ ,  $\{w_i \rightarrow y_i | y_i \in \mathcal{L}\}$ . This produces a label sequence,  $Y = \{y_i \in \mathcal{L} | i = 1, \dots, N\}$  of the same length  $N$ . Using decision theory, sequence labeling models seek to output  $\hat{Y}$  from a set of all possible tag sequence  $\mathcal{L}^N$ ,

$$\hat{Y} = \operatorname{argmax}_{Y \in \mathcal{L}^N} P(Y|S) \quad (1)$$

where  $P(Y|S)$  is the posterior distribution. This posterior can be modeled using various techniques like the traditional HMM (Morwal et al., 2012) and MEMM (McCallum et al., 2000) based modeling and more recently CRF (Ma and Hovy, 2016) and token classification (Devlin et al., 2019) based approaches. We discuss the latter two in detail:

**Conditional Random Field:** Lafferty et al. (2001) aims to directly compute the posterior of the entire label sequence  $Y$  given the sentence  $S$ :

$$P(Y|S) = \frac{e^{F(Y,S)}}{\sum_{Y' \in \mathcal{L}^N} e^{F(Y',S)}} \quad (2)$$

where  $F(Y, S)$  is global score of the tag sequence  $Y$  given  $S$ . This is modeled using a linear chain CRF which computes the global score as a sum of local scores  $f(\cdot)$  for each position in  $Y$  as follows

$$F(Y, S) = \sum_{l=1}^N f(y_{l-1}, y_l, S) \quad (3)$$

Lample et al. (2016) and Yan et al. (2019) use contextualized neural encoders like LSTMs and transformers to model context of the entire sequence  $S$  for every word  $w_l$ . This allows for effective modeling of  $f(\cdot)$  by using encoder representations for each word as the emissions, and maintaining a

separate transition score  $t_{y_{l-1} \rightarrow y_l}$  to give  $F(Y, S)$ :

$$\mathbf{h}_{1:N} = \operatorname{encoder}(w_{1:N}) \quad (4)$$

$$t_{y_{l-1} \rightarrow y_l} = \operatorname{transitionScores}(|\mathcal{L}|, |\mathcal{L}|) \quad (5)$$

$$F(Y, S) = \sum_{l=1}^N (\mathbf{h}_{l,y_l} + t_{y_{l-1} \rightarrow y_l}) \quad (6)$$

**Token Classification Model:** Since the advent of strong contextual modeling using transformer based models, sequence labeling can also be treated as token classification (Devlin et al., 2019), a simplification over MEMM estimations (McCallum et al., 2000), with the assumption that the current tag is conditionally independent to previous tag.

$$P(Y|S) = \prod_{l=1}^N P(y_l | \mathbf{h}_l) \quad (7)$$

These models are still effective as  $\mathbf{h}_l$  is able to model the full context  $S$  for every word  $w_l$ .

In cases like NER, where an entity can span multiple words, these problems are modeled using BIO tags (Ramshaw and Marcus, 1995), where begin (B), inside (I) tags are added for entities and an outside (O) tag for non-entity words, extending the tag set vocabulary from  $\mathcal{L}$  to  $\mathcal{L}' = \{l_B \oplus l_I | l \in \mathcal{L}\} \cup \{O\}$ . When modeled using sub-word tokens the tags can be aligned to the first sub-word token of the word and the remaining ones can be marked with a special token  $\emptyset$  giving  $\mathcal{L}'' = \mathcal{L}' \cup \{\emptyset\}$ .

## 3 Sequence Labeling in SLU

Sequence Labeling in SLU introduces an added complexity of recognizing mentions on top of text-based SL tasks (§2) as they aim to predict the tag and its mentions directly from a spoken sequence. Given a sequence of  $d$  dimensional speech feature of length  $T$  frames,  $X = \{\mathbf{x}_t \in \mathbb{R}^d | t = 1, \dots, T\}$ , these systems seek to estimate the label sequence  $\hat{Y}$  from

$$\hat{Y} = \operatorname{argmax}_{Y \in \mathcal{L}^*} P(Y|X) \quad (8)$$

where  $P(Y|X)$  have been modeled as:

**Cascaded SLU** (Béchet et al., 2004; Parada et al., 2011; Zhou et al., 2015) models  $P(Y|X)$  from  $P(Y|S)$  using an NLU framework (§2) and  $P(S|X)$  using an ASR model (Povey et al., 2011; Chan et al., 2016; Graves, 2012), assuming condi-

tional independence of  $Y|S$  from  $X$ ,

$$P(Y|X) = \sum_S P(Y|S, X)P(S|X) \quad (9)$$

$$\approx \max_S P(Y|S)P(S|X) \quad (10)$$

$$\approx P(Y|\hat{S}) \max_S P(S|X) \quad (11)$$

$$\hat{S} = \underset{S \in \mathcal{V}^*}{\operatorname{argmax}} P(S|X) \quad (12)$$

Once  $\hat{S}$  is estimated,  $\hat{Y}$  can be estimated using Eq 1. Although this enables realizing  $\hat{Y}$  using two well studied frameworks, the independence assumption doesn't allow recovery from errors in estimating  $\hat{S}$ .

**Direct End-to-End SLU** (Arora et al., 2022; Shon et al., 2022; Ghannay et al., 2018) systems avoid cascading errors by directly modeling  $P(Y|X)$  in a single monolithic model. To achieve this while being able to recognize the spoken mentions, these systems enrich  $Y$  with transcripts  $S$ ,  $Y^e = \{y_i^e \in \mathcal{V} \cup \mathcal{L} | i = 1, \dots, N'\}$ , where  $N'$  is the length of  $Y^e$ . This can be modeled using an autoregressive decoder as:

$$P(Y|X) = \prod_{i=1}^{N'} P(y_i^e | y_{1:i-1}^e, X) \quad (13)$$

However this new formulation cannot utilize the well studied sequence labeling framework §2. Additionally, this applies an extra burden of labeling along with alignment on the decoder and makes understanding the errors made by these systems particularly difficult. For example, Eq 13 gives non-zero likelihood to a corrupt sequence with only labels and no words as  $y^e \in \{\mathcal{V} \cup \mathcal{L}\}$ .

## 4 Compositional End-to-End SLU

We propose to bring the two paradigms together in a compositional end-to-end system, by extending over the cascaded SLU formulation using searchable intermediate framework (Dalmia et al., 2021):

$$P(Y|X) = \sum_S P(Y|S, X)P(S|X) \quad (14)$$

$$\approx \max_S \underbrace{P(Y|S, X)}_{\text{SUBNLU NET}} \underbrace{P(S|X)}_{\text{SUBASR NET}} \quad (15)$$

This system can be realized with two sub-networks as shown in Figure 1, where:

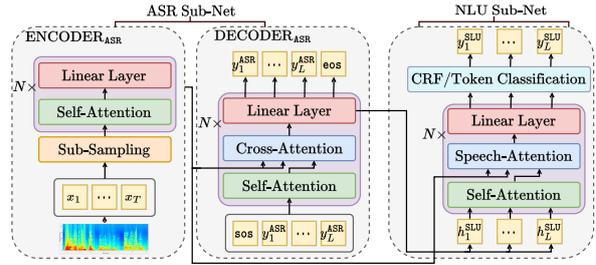


Figure 1: Schematics of our compositional E2E SLU architecture with ASR and NLU sub-nets. The ASR sub-net consists of an encoder and decoder. The NLU sub-net consists of an encoder that conditions on both speech information via  $\text{encoder}_{\text{ASR}}$  and the text information via  $\text{decoder}_{\text{ASR}}$ 's hidden representation  $\mathbf{h}^{\text{ASR}}$  followed by token classification or CRF layer.

**SUBASRNET:** Models  $P(S|X)$ ,

$$\mathbf{h}_{1:T}^E = \text{encoder}_{\text{ASR}}(X_{1:T}) \quad (16)$$

$$\mathbf{h}_l^{\text{ASR}} = \text{decoder}_{\text{ASR}}(\mathbf{h}_{1:T}^E, w_{1:l-1}) \quad (17)$$

$$P(w_l|X, w_{1:l-1}) = \text{softmaxOut}(\mathbf{h}_l^{\text{ASR}}) \quad (18)$$

$$P(S|X) = \prod_{l=1}^N P(w_l|X, w_{1:l-1}) \quad (19)$$

**SUBNLU NET:** Models  $P(Y|S, X)$ ,

$$\mathbf{h}_{1:N}^{\text{NLU}} = \text{encoder}_{\text{NLU}}(\mathbf{h}_{1:N}^{\text{ASR}}, \mathbf{h}_{1:T}^E) \quad (20)$$

$$P(Y|S, X) = \text{CRF}(\mathbf{h}_{1:N}^{\text{NLU}}) \quad \text{OR} \quad (21)$$

$$P(Y|S, X) = \text{TokenClassification}(\mathbf{h}_{1:N}^{\text{NLU}}) \quad (22)$$

The end-to-end differentiability is maintained by using  $\mathbf{h}_{1:N}^{\text{ASR}}$  in Eq 20. During inference, we approximate the Viterbi max of  $S$  using beam search to give  $\hat{\mathbf{h}}_{1:N}^{\text{ASR}}$ . Then  $\hat{Y}$  can be found using Viterbi search with no approximation as the output length is known and the solution is tractable.

This composition allows incorporating the ASR modeling and text-based sequence labeling framework §2. It also brings transparency to end-to-end modeling as we can also monitor performance of individual sub-nets in isolation. Further,  $\text{encoder}_{\text{NLU}}$  can attend to speech representations  $\mathbf{h}_{1:T}^E$  using cross attention (Dalmia et al., 2021) enabling the direct use of speech cues for NLU. This speech attention mechanism can allow the model to recover from intermediate errors made during ASR stage.

Recently, there has been some works (Rao et al., 2020; Saxon et al., 2021) that explore compositional SLU models which utilize the ASR and NLU formulations. Saxon et al. (2021) uses discrete outputs from the ASR module that are made differ-

entiable using various approaches like Gumbel-softmax (Jang et al., 2017). Rao et al. (2020) also uses the ASR decoder hidden representations in the NLU module by concatenating it with token embeddings of the ASR discrete output. However, this approach requires the ASR and NLU submodule to have a shared vocabulary space, limiting the usage of pretrained ASR and LM in this architecture. Moreover, the benefits of our proposed compositional framework are not explored in these works.

## 5 Spoken Named Entity Recognition

To show the effectiveness of our compositional E2E SLU model we build spoken NER systems on two publically available SLU datasets, SLUE (Shon et al., 2022) and SLURP (Bastianelli et al., 2020) (dataset and preparation details in §A.2). We compare our compositional E2E system with cascaded and direct E2E systems. We also compare with another compositional E2E system that predicts the enriched transcript (§3) using a decoder like (Dalmia et al., 2021) instead of label sequence (i.e.  $Y^e$  instead of  $Y$  in Eq. 15) using a token level classification sub-network. We refer to this baseline model as “Compositional E2E SLU with Direct E2E formulation”.

SLURP is evaluated using SLU-F1 (Bastianelli et al., 2020) which weighs the entity labels with the word and character error rate of the predicted mentions and SLUE using F1 (Shon et al., 2022) which evaluates getting both the mention and the entity label exactly right. We also compute Label-F1 for both datasets which considers only the entity label. We report micro-averaged F1 for all results.

### 5.1 Model Configurations

We build all our systems using ESPnet-SLU (Arora et al., 2022) which is an open-source SLU toolkit built on ESPnet (Watanabe et al., 2018), a flagship toolkit for speech processing. We use encoder-decoder based architecture for our baseline E2E system. We use Conformer encoder blocks (Gulati et al., 2020) and Transformer decoder blocks (Vaswani et al., 2017) with CTC multi-tasking (Arora et al., 2022). The baseline compositional model with Direct E2E SLU formulation consists of a conformer encoder and transformer decoder in it’s ASR component and transformer encoder and transformer decoder in it’s NLU component. Our proposed compositional model with the NLU formulation, as shown in Figure 1, replaces the

Model	SLURP		SLUE	
	SLU F1	Label F1	F1	Label F1
Direct E2E SLU (Arora et al.)	71.9	-	54.7	67.6
Cascaded SLU (Ours)	73.3	80.9	48.6	63.9
Direct E2E SLU (Ours)	77.1	84.0	54.7	67.6
Compositional E2E SLU				
w/ Direct E2E formulation (§3)	77.2	84.6	50.0	68.0
w/ Proposed NLU formulation (§4)				
CRF w/ Speech Attention (SA)	77.7	85.2	59.4	73.6
Token Classification w/ SA	<b>78.0</b>	<b>85.3</b>	<b>60.3</b>	<b>73.7</b>
w/o Speech Attention	77.7	84.9	59.0	73.6

Table 1: Results presenting the micro F1 performance of our proposed compositional E2E models using CRF and Token Classification modeling. Cascaded, direct E2E and our compositional E2E with direct E2E formulation are shown for comparison. We also provide an ablation of our model with and without Speech Attention (SA).

NLU component in Direct E2E formulation with a transformer encoder followed by a linear layer. For the cascaded systems, we build systems that have the same size as that of our ASR and NLU sub-networks. All models were tuned separately using validation sets with the same hyperparameter search space. Full descriptions of model and training parameters are in §A.3.

### 5.2 Performance of Compositional E2E SLU

Table 1 shows that our proposed compositional E2E models with the token-level NLU formulation outperform both cascaded and direct E2E models on all benchmarks using both CRF and Token Classification. In order to understand gains of our proposed model, we examine the performance of our compositional system with direct E2E formulation (§3). While being comparable to direct E2E models, they still lag behind our proposed models showing the efficacy of modeling SL tasks as a token-level tagging (§2) in an E2E SLU framework.

We further analyze our compositional systems that don’t attend to speech representations. We observe a performance drop as these models are not able to recover from errors made while “recognising” entity mentions. For example, in an utterance that says “change the bedroom lights to green”, though the ASR component incorrectly predicts the transcript as “change the color of lights to green”, the NLU component w/ Speech Attention is able to recover the entity type HOUSE\_PLACE.

### 5.3 Utilizing External Sub-Net models

Components of our compositional E2E SLU model have functions similar to an ASR and NLU model (Eq 16-22). This allows fine-tuning our models using sub-systems, pre-trained on large amounts

Model	SLURP		SLUE	
	SLU F1	Label F1	F1	Label F1
Direct E2E SLU	77.1	84.0	54.7	67.6
w/ NLU fine-tuning		Incompatible		
w/ ASR fine-tuning	73.5	81.2	64.0	80.6
Compositional E2E SLU (w/ SA)	78.0	85.3	60.3	73.7
w/ NLU finetuning (w/o SA)	77.7	84.9	62.4	76.4
w/ ASR finetuning (w/ SA)	<b>81.4</b>	<b>88.8</b>	<b>71.6</b>	<b>85.2</b>
Compositional E2E SLU (w/ SA)	78.0	85.3	60.3	73.7
w/ External ASR Transcripts ( $S^{\text{ext}}$ )	<b>81.0</b>	<b>88.1</b>	<b>70.1</b>	<b>81.2</b>

Table 2: Results presenting the compatibility of our models with pre-trained ASR and NLU systems by (1) finetuning pre-trained components and (2) directly utilizing transcripts from an external ASR model.

	SLURP		SLUE	
	ASR (%WER ↓)	NLU (SLU-F1 ↑)	ASR (%WER ↓)	NLU (F1 ↑)
Pure ASR & NLU models	16.1	82.4	30.4	58.1
Compositional E2E SLU				
CRF w/ Speech Attention (SA)	16.3	88.3	27.4	75.6
Token Classification w/o SA	16.0	87.9	27.6	74.1
Token Classification w/ SA	16.1	88.7	27.5	75.6

Table 3: Results showcasing the transparency of our compositional E2E models by evaluating the individual sub-networks ASR (%WER) and NLU (F1) in isolation.

of available sub-task data. Table 2 shows that our compositional model has better compatibility with ASR and NLU fine-tuning over direct E2E systems, thereby increasing their performance gap, particularly for SLUE, an under-resourced SLU dataset.

Further our models have the ability to use transcripts from a strong external model ( $S^{\text{ext}}$ ) directly during inference, by instantiating our models with these transcripts to produce  $\mathbf{h}^{\text{ASR}}$  and then evaluate  $P(Y|S^{\text{ext}}, X)$ . Table 2 shows using transcripts from an external ASR *with no fine-tuning steps* can achieve similar performance to ASR fine-tuning.

#### 5.4 Transparency in Compositional E2E SLU

Following Eq 15, we can estimate ASR performance by calculating  $\hat{S}$  using beam search and NLU performance by estimating  $\hat{Y}$  from  $P(Y|S^{\text{GT}}, X)$ , where  $S^{\text{GT}}$  is the ground truth transcripts. Table 3 shows the performances of individual components of our model along with performances of ASR and NLU only models suggesting that we can effectively monitor the performance of these components, helping practitioners analyze and debug them. For instance, while our models with and without speech attention have comparable performance on ASR, using speech attention improves NLU power. Further the one-to-one alignment of transcripts and sequence labels can provide further categorization of errors, as shown in §A.4.

#### 5.5 CRF vs Token Classification

For practical SLU the likelihoods of our compositional model  $P(Y|S, X)$ , should be correlated with errors in label sequence  $Y$ . We found that in SLURP our compositional E2E SLU, while using locally normalized token classification shows no correlation (Corr=0.13, p=0), using CRF exhibits moderate correlation (Corr=0.43, p=0). This makes globally normalized models attractive for real-world scenarios like automated data auditing and human in-the-loop ML (Mitchell et al., 2018) despite their marginal addition in computation cost.

### 6 Conclusion

We propose to combine text based sequence labeling framework into the speech recognition framework to build a compositional end-to-end model for SLU. Our compositional E2E models not only show superior performance over cascaded and direct end-to-end SLU systems, but also bring the power of both these systems in a single framework. These models can utilise pretrained sub task components and exhibit transparency like cascaded systems, while avoiding error propagation like direct end-to-end systems.

#### Limitations

Our compositional model relies on the availability of transcripts for training. This although is a limitation, it is a safe assumption for sequence labeling tasks for spoken language understanding. We can see from §3 that the task for sequence labeling in SLU also requires the model to recognize the words being spoken along with the sequence labels, implying the need for at least a partial transcript for training direct end-to-end SLU systems.

#### Broader Impact

With our compositional end-to-end SLU model, we strive to bring the research from the text based sequence labeling directly into speech based spoken language understanding. Our aim is to avoid re-invention of the wheel, but rather come up with innovative ways to build end-to-end models by converting a complex problem into simpler ones that have seen substantial research in the past. Additionally we believe the increased capacity for error analysis in our compositional end-to-end system can help towards building better practical systems during deployment. Our compositional end-to-end

systems can effectively utilize pre-trained ASR and NLU systems, thereby avoiding the need for collecting large labeled datasets for SLU. This framework also saves compute by utilizing pre-trained ASR systems directly during inference to improve downstream performances with no fine-tuning.

## Acknowledgement

We thank Aakanksha Naik and the anonymous reviewers for their feedback. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) (Townes et al., 2014), which is supported by NSF grant number ACI-1548562. Specifically, it used the Bridges system (Nystrom et al., 2015), which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

## References

- Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. [Contextual string embeddings for sequence labeling](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Siddhant Arora, Siddharth Dalmia, Pavel Denisov, Xuankai Chang, Yushi Ueda, Yifan Peng, Yuekai Zhang, Sujay Kumar, Karthik Ganesan, Brian Yan, Ngoc Thang Vu, Alan W. Black, and Shinji Watanabe. 2022. [ESPnet-SLU: Advancing spoken language understanding through espnet](#). In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7167–7171. IEEE.
- Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. [SLURP: A spoken language understanding resource package](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Association for Computational Linguistics.
- Frédéric Béchet, Allen L. Gorin, Jeremy H. Wright, and Dilek Hakkani-Tür. 2004. [Detecting and extracting named entities from spontaneous speech in a mixed-initiative spoken dialogue context: How may I help you?](#)<sup>sm, tm</sup>. *Speech Commun.*, 42(2):207–225.
- William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. [Listen, attend and spell: A neural network for large vocabulary conversational speech recognition](#). In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964.
- Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021a. [Gigaspeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio](#). In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 3670–3674. ISCA.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, and Furu Wei. 2021b. [WavLM: Large-scale self-supervised pre-training for full stack speech processing](#). *CoRR*, abs/2110.13900.
- Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. [Canine: Pre-training an efficient tokenization-free encoder for language representation](#). *Transactions of the Association for Computational Linguistics*, 10:73–91.
- Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. [Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces](#). *CoRR*, abs/1805.10190.
- Siddharth Dalmia, Brian Yan, Vikas Raunak, Florian Metze, and Shinji Watanabe. 2021. [Searchable hidden intermediates for end-to-end models of decomposable sequence tasks](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1882–1896, Online. Association for Computational Linguistics.
- Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Żelasko, and Miguel Jetté. 2021. [Earnings-21: A Practical Benchmark for ASR in the Wild](#). In *Proc. Interspeech 2021*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. [Stanford’s graph-based neural dependency parser at the CoNLL 2017 shared task](#). In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30, Vancouver, Canada. Association for Computational Linguistics.

- S. Ghannay, A. Caubrière, Y. Estève, N. Camelin, E. Simonnet, A. Laurent, and E. Morin. 2018. [End-to-end named entity and semantic concept extraction from speech](#). In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 692–699.
- Alex Graves. 2012. [Sequence transduction with recurrent neural networks](#). *CoRR*, abs/1211.3711.
- Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. [Conformer: Convolution-augmented transformer for speech recognition](#). In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 5036–5040. ISCA.
- James Horlock and Simon King. 2003. [Discriminative methods for improving named entity extraction on speech data](#). In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 2765–2768.
- Eric Jang, Shixiang Gu, and Ben Poole. 2017. [Categorical reparameterization with gumbel-softmax](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Dan Jurafsky and James H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International.
- Francis Kubala, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. 1998. [Named entity extraction from speech](#). In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pages 287–292. Citeseer.
- Taku Kudo and John Richardson. 2018. [Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.
- John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. [Conditional random fields: Probabilistic models for segmenting and labeling sequence data](#). In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. [Neural architectures for named entity recognition](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.
- Xuezhe Ma and Eduard Hovy. 2016. [End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.
- Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira. 2000. [Maximum entropy markov models for information extraction and segmentation](#). In *Proceedings of the Seventeenth International Conference on Machine Learning, ICML '00*, page 591–598, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2018. [Never-ending learning](#). *Commun. ACM*, 61(5):103–115.
- Sudha Morwal, Nusrat Jahan, and Deepti Chopra. 2012. [Named entity recognition using hidden markov model \(hmm\)](#). *International Journal on Natural Language Computing*, 1:15–23.
- Minh Nguyen and Zhou Yu. 2021. [Improving named entity recognition in spoken dialog systems by context and speech pattern modeling](#). In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 45–55, Singapore and Online. Association for Computational Linguistics.
- Nicholas A. Nystrom, Michael J. Levine, Ralph Z. Roskies, and J. Ray Scott. 2015. [Bridges: a uniquely flexible HPC resource for new communities and data analytics](#). In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure, St. Louis, MO, USA, July 26 - 30, 2015*, pages 30:1–30:8. ACM.
- David D. Palmer and Mari Ostendorf. 2001. [Improving information extraction by modeling errors in speech recognizer output](#). In *Proceedings of the First International Conference on Human Language Technology Research*.
- Carolina Parada, Mark Dredze, and Frederick Jelinek. 2011. [OOV sensitive named-entity recognition in speech](#). In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, pages 2085–2088. ISCA.
- Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le.

2019. [SpecAugment: A simple data augmentation method for automatic speech recognition](#). In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 2613–2617. ISCA.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.
- Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. [The Kaldi speech recognition toolkit](#). In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.
- Lance A. Ramshaw and Mitch Marcus. 1995. [Text chunking using transformation-based learning](#). In *Third Workshop on Very Large Corpora, VLC@ACL 1995, Cambridge, Massachusetts, USA, June 30, 1995*.
- Milind Rao, Anirudh Raju, Pranav Dheram, Bach Bui, and Ariya Rastrow. 2020. [Speech to semantics: Improve ASR and NLU jointly via all-neural interfaces](#). In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 876–880. ISCA.
- Michael Saxon, Samridhi Choudhary, Joseph P. McKenna, and Athanasios Mouchtaris. 2021. [End-to-end spoken language understanding for generalized voice assistants](#). In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 4738–4742. ISCA.
- Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J. Han. 2022. [SLUE: new benchmark tasks for spoken language understanding evaluation on natural speech](#). In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7927–7931. IEEE.
- John Towns, Timothy Cockerill, Maytal Dahan, Ian T. Foster, Kelly P. Gauthier, Andrew S. Grimshaw, Victor Hazelwood, Scott A. Lathrop, David Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, and Nancy Wilkins-Diehr. 2014. [XSEDE: accelerating scientific discovery](#). *Comput. Sci. Eng.*, 16(5):62–74.
- Trang Tran, Shubham Toshniwal, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Mari Ostendorf. 2018. [Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 69–81. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplín, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. [Espnet: End-to-end speech processing toolkit](#). In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 2207–2211. ISCA.
- Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. 2019. [Tener: Adapting transformer encoder for named entity recognition](#).
- Dian Yu, Michelle Cohn, Yi Mang Yang, Chun-Yen Chen, Weiming Wen, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, Girithija Sreenivasulu, Sam Davidson, Ashwin Bhandare, and Zhou Yu. 2019. [Gunrock: A social bot for complex and engaging long conversations](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019 - System Demonstrations*. Association for Computational Linguistics.
- Lu-Feng Zhai, Pascale Fung, Richard M. Schwartz, Marine Carpuat, and Dekai Wu. 2004. [Using n-best lists for named entity recognition from chinese speech](#). In *Proceedings of HLT-NAACL 2004: Short Papers, Boston, Massachusetts, USA, May 2-7, 2004*. The Association for Computational Linguistics.
- Liyuan Zhou, Hanna Suominen, and Leif Hanlen. 2015. [Evaluation data and benchmarks for cascaded speech recognition and entity extraction](#). In *Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia, SLAM '15*, page 15–18, New York, NY, USA. Association for Computing Machinery.

## A Appendix

### A.1 Applications of SLU

SLU is an essential component of many commercial devices like voice assistants, home assistants (Yu et al., 2019; Coucke et al., 2018) and spoken dialog systems (Nguyen and Yu, 2021) that map speech to executable commands on a daily basis. One of the key applications of SLU is to extract key mentions like entities from a user command to take appropriate actions. As a result, several datasets (Bastianelli et al., 2020; Shon et al., 2022; Del Rio et al., 2021) have been proposed to build understanding systems for spoken utterances.

### A.2 Dataset Description

We evaluated our proposed approach on publicly available SLU datasets, namely SLUE (Shon et al., 2022) and SLURP (Bastianelli et al., 2020) datasets on the task of Named Entity Recognition (NER) from naturally available speech. SLURP is a linguistically diverse and challenging spoken language understanding benchmark that consists of single-turn user conversation with a home assistant, annotated with both intent and entities. Similar to the approach followed in our prior work (Bastianelli et al., 2020; Arora et al., 2022), we bootstrap our train set with 43 hours of synthetic data for all our experiments. We evaluate our approach using SLU-F1 (Bastianelli et al., 2020), a metric for spoken entity prediction, and Label F1, which considers only entity-tag predictions.

SLUE is a recently released SLU benchmark that focuses on Spoken Language Understanding from limited labeled training data. Specifically, it consists of SLUE VoxPopuli dataset that can be used for building systems for ASR and NER. Similar to (Shon et al., 2022), we evaluate our systems using two micro-averaged F1 scores, the first score that evaluates both named entity and tag pairs is referred to as F1, and the second that evaluates only entity-tag phrases is referred to as Label-F1. Note that the released test sets are blind without ground truth labels, and hence we compare different methods using the development set.

The dataset download and evaluation links for SLURP can be found here - <https://github.com/pswietojanski/slurp> and for SLUE here - <https://github.com/asappresearch/slue-toolkit>. The datasets have been processed and prepared using ESPnet, SLURP - <https://github.com/espnet/>

Table 4: Overview of the two publicly available SLU datasets (Shon et al., 2022; Bastianelli et al., 2020) used for our experiments.

Dataset	Size (utterances / hours)		
	Train	Dev	Test
SLURP	11,514 / 40.2	2,033 / 6.9	2,974 / 10.3
SLUE-VoxPopuli	5,000 / 14.5	1,753 / 5.0	1,842 / 4.9

[espnet/tree/master/egs2/slurp\\_entity](https://github.com/espnet/espnet/tree/master/egs2/slurp_entity) and SLUE - <https://github.com/espnet/espnet/tree/master/egs2/slue-voxpathuli>

### A.3 Experimental Setup

Our models are implemented in PyTorch (Paszke et al., 2019), and the experiments are conducted using the ESPnet-SLU toolkit (Arora et al., 2022).

#### A.3.1 Speech Preprocessing

Speech inputs are globally mean-variance normalized 80 dimensional logmel filterbanks using a 16kHz sampling and window of 512 frames and a 128 hop length. We apply speed perturbation for the under-resourced dataset of SLUE of 0.9 and 1.1 to increase the samples. We also apply specaugmentation (Park et al., 2019) on both datasets. We also remove all examples smaller than 0.1 seconds and larger than 20 seconds from the training data.

#### A.3.2 Text Processing

For the cascaded system, we process ASR transcripts  $S$  using bpe tokenization (Kudo and Richardson, 2018) and train ASR models to generate bpe subtokens. We use bpe size of 500 for SLURP and 1000 for SLUE dataset. For the direct E2E models, we predict the enriched label sequence  $Y^e$  using the same bpe size as the ASR models in cascaded sequence. Similarly, compositional models also use the same bpe size to generate the ASR transcripts.

For creating the BIO tags we modify the data preparation such that we take the entities for each utterance and create a “label utterance”. This consists of one-to-one mapping of the label tags with the words and Begin (B), Inside (I) and Outside (O) marked for each label. After performing BPE tokenization we add  $\emptyset$  for every subtoken of the word. We have attached the data preparation code.

#### A.3.3 Model and Training Hyperparameters

We run parameter search for both direct end-to-end and our compositional end-to-end systems using the same model search space (Table 5). In this

section, we will describe our best architecture for both direct and compositional E2E systems.

**Direct E2E SLU systems** After searching through hyperparameter space, our Direct E2E SLU systems consists of 12-layer Conformer (Gulati et al., 2020) encoder and a 6-layer Transformer (Vaswani et al., 2017) decoder with 8 attention heads for SLURP dataset. We use a dropout of 0.1, output dim of 512 and feedforward dim of 2048, giving a total parameter size of 109.3 M.

For SLUE dataset, we found 12-layer Conformer with 4 attention heads and decoder is a 6-layer Transformer with 4 attention heads to give best validation performance. We use a dropout of 0.1, output dim of 256 and feedforward dim of 1024 in encoder and 2048 in the decoder, giving a total parameter size of 31.2 M.

**Compositional E2E SLU systems** Our Compositional model which uses Direct E2E SLU formulation consists of 12-layer conformer block for encoder, 6-layer transformer block for decoder in it’s ASR component and 4-layer transformer encoder and 6-layer transformer decoder in it’s NLU component. Each of these attention blocks consist of 8 attention heads, dropout of 0.1, output dim of 512, feedforward dim of 2048, giving a total of 153.9M parameters in SLURP dataset. For SLUE dataset, each attention block has 4 attention heads, dropout of 0.1, output dim of 256, feedforward dimension of 1024 in encoder and 2048 in decoder, giving a total parameter size of 46.8M.

Our Composition model with Proposed NLU formulation replaces NLU component in Direct E2E formulation with 8-layer transformer encoder followed by linear layer. All these attention blocks consist of 8 attention heads, dropout of 0.1, output dim of 512, feedforward dim of 2048, giving a total of 142.9M parameters in SLURP dataset. For SLUE dataset, each of these attention blocks have 4 attention heads, dropout of 0.1, output dim of 256, feedforward dimension of 1024 in encoder and 2048 in decoder, giving a total parameter size of 43.8M. Our NLU component can further attend to speech representations using cross attention (Dalmia et al., 2021). We further implement CRF loss using publicly available python library<sup>2</sup>.

The loss from the ASR ( $\mathcal{L}^{\text{asr}}$ ) and NLU ( $\mathcal{L}^{\text{nlu}}$ ) subnet are combined as follows

$$\mathcal{L} = \mathcal{L}^{\text{asr}} + \alpha \mathcal{L}^{\text{nlu}}$$

<sup>2</sup><https://pytorch-crf.readthedocs.io>

Hyperparameter	Value
Output Size	[256, 512]
Attention Heads	[4, 8]
Number of blocks	[4, 6, 8, 12]
Hidden Dropout	[0.1, 0.2]
Attention dropout	[0.1, 0.2]
Position dropout	[0.1, 0.2]
Activation dropout	[0.1, 0.2]
Src Activation dropout	[0.1, 0.2]
Batch size	[ 50, 64]
LR schedule	[inv. sqrt., exp. lr.]
Max learning rate	[0.001, 0.002, 0.003]
Warmup steps	[5000, 15000, 25000]
Number of steps	[50, 70, 100]
Adam eps	1e-9
Adam betas	(0.9, 0.98)
Weight decay	0.000001

Table 5: Model and Training Search for SLU Models.

We search alpha values over [0.3, 0.4, 0.5, 0.6] and found 0.6 to be best for SLURP and 0.3 for SLUE.

Model	SLURP		SLUE	
	SLU F1	Label F1	F1	Label F1
Cascaded SLU (Ours)	76.9	83.9	48.6	63.9
Direct E2E SLU (Ours)	79.2	85.4	54.7	67.6
Compositional E2E SLU				
w/ Direct E2E formulation (§3)	79.3	86.6	50.0	68.0
w/ Proposed NLU formulation (§4)				
CRF w/ Speech Attention (SA)	79.9	87.0	59.4	73.6
Token Classification w/ SA	79.8	86.9	60.3	73.7
w/o Speech Attention	79.7	87.0	59.0	73.6

Table 6: Results presenting the micro F1 performance for all models using CRF and Token Classification modeling on development set for SLURP and SLUE

### A.3.4 Decoding Hyperparameters

We keep the same decoding parameter of beam size and penalty as that of Arora et al. (2022). For direct E2E systems and our models CTC weight of 0.1 worked best. We searched over CTC weight of [0, 0.1, 0.3, 0.5].

### A.3.5 Development Results

We use F1 scores on the validation data to select the best hyperparameters. Table 6 presents the validation performances for our models.

### A.3.6 Compute Infrastructure

Our models were trained using mixed precision training on either a100, v100 or A6000 on our compute infrastructure depending on their availability. Depending on the GPU and the file i/o latency, the training time ranged from 4-7 hours for SLUE, while for SLURP the training time ranged from 12-18 hours.

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

Figure 2: Qualitative examples of our compositional E2E SLU model for various error categories. We can observe that in the first case, the model is correctly able to predict both entity types and mentions even when the name “mona” is not a common name for an event. In the second case, even though it predicts the correct ASR transcript, it mislabels “Edinburgh” as a news topic since the phrase “is there anything happening” usually occurs with news topics. In the third case, even though it makes a mistake in the person name, the model correctly tags it as a person. Finally, the model incorrectly generates the word “ninety,” and this error gets propagated to the NLU component through token representations which then predicts entity type “date”. This analysis shows that the alignment between ASR and NLU outputs can help us gain better insights into model performance.

### A.3.7 External ASR and NLU components

For the experiments in Table 2, we used ASR and NLU models trained on external data. For the ASR fine-tuning we used an ESPnet model<sup>3</sup> trained on the GigaSpeech dataset (Chen et al., 2021a). This model has the same architecture as the baseline direct E2E model on SLURP. We initialize both the encoder and decoder for direct E2E SLU and the ASR sub-net for the compositional E2E SLU model. For NLU fine-tuning we used Canine (Clark et al., 2022), a character based BERT language model, which exhibits strong performance on named entity recognition while being able to model token sizes comparable to our SLU systems.<sup>4</sup> We initialize our NLU sub-network without speech attention with Canine and keep the model parameters fixed during training. For finding the best parameters we only tuned the learning rate and LR schedule from Table 5 and report the best numbers among CRF and Token Classification loss.

For using External ASR Transcripts, we trained an ASR system initialized using GigaSpeech and WavLM (Chen et al., 2021b) respectively. They were then fine-tuned on the respective datasets. These systems achieve 10.0% WER and 9.2% WER on SLURP and SLUE respectively.

## A.4 Error Categorization

The predictions made by our compositional E2E SLU model can be categorized into different buckets on the basis of the errors by ASR or NER component. Table 7 demonstrates this behavior by categorizing the errors of our compositional E2E model

	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

Table 7: Number of examples per error category of our compositional E2E SLU systems with/without Speech Attention on SLURP test set. There are four categories depending on whether mistakes are made by ASR or NLU component. Note that the **first quadrant** lists # of **correct** examples, while the **rest** list **incorrect** ones. Direct E2E systems cannot offer such categorizations particularly for incorrect entities as there is no alignment between ASR and NLU outputs.

trained with and without speech attention. Most of the performance differences between compositional E2E SLU model w/ and w/o speech attention are caused by the kinds of errors where the ASR predictions are inaccurate, but the NLU module is nevertheless able to recover the correct entity type from the utterance. This confirms our intuition that cross attention on speech representations can help the NLU module to recover from mistakes made during “recognizing” spoken mentions. We also present anecdotes for each of these error categories in Figure 2. This further emphasizes the transparency in our compositional E2E SLU models. Due to the lack of one-to-one alignment between ASR and Sequence Labeling, such analysis is not possible in direct E2E SLU systems, making it particularly difficult to categorize errors when the entity prediction is wrong.

<sup>3</sup><https://zenodo.org/record/4630406>

<sup>4</sup><https://huggingface.co/google/canine-s>