SLURP: A Spoken Language Understanding Resource Package

Emanuele Bastianelli[†*], **Andrea Vanzo**[†*], **Pawel Swietojanski**[‡*] **and Verena Rieser**[†]
[†]The Interaction Lab, MACS, Heriot-Watt University, Edinburgh, UK
[‡]Faculty of Engineering, University of New South Wales, Sydney, Australia
{e.bastianelli, a.vanzo, v.t.rieser}@hw.ac.uk
p.swietojanski@unsw.edu.au

Abstract

Spoken Language Understanding infers semantic meaning directly from audio data, and thus promises to reduce error propagation and misunderstandings in end-user applications. However, publicly available SLU resources are limited. In this paper, we release SLURP, a new SLU package containing the following: (1) A new challenging dataset in English spanning 18 domains, which is substantially bigger and linguistically more diverse than existing datasets; (2) Competitive baselines based on state-of-the-art NLU and ASR systems; (3) A new transparent metric for entity labelling which enables a detailed error analysis for identifying potential areas of improvement. SLURP is available at https://github.com/pswietojanski/slurp

1 Introduction

Traditionally, Spoken Language Understanding (SLU) uses a pipeline transcribing audio into text using Automatic Speech Recognition (ASR), which is then mapped into a semantic structure via Natural Language Understanding (NLU). However, this modular approach is prone to error propagation from noisy ASR transcriptions, and ASR in turn is not able to disambiguate based on semantic information. End-to-end (E2E) approaches on the other hand, can benefit from joint modelling. One of the main bottlenecks for building E2E-SLU systems, however, is the lack of large and diverse datasets of audio inputs paired with corresponding semantic structures. Publicly available datasets to date are limited in terms of lexical and semantic richness (Lugosch et al., 2019b), number of vocalizations (Coucke et al., 2018), domain coverage (Hemphill et al., 1990; Dahl et al., 1994) and semantic contexts (Godfrey et al., 1992; Jurafsky and Shriberg, 1997). In this paper, we present the

User: *"Make a calendar entry for brunch on Saturday morning with Aaronson."*
Scenario: Calendar
Action: Create_entry
Entity tags and lexical fillers: [event_name: brunch], [date: Saturday], [timeofday: morning], [person: Aaronson]

Figure 1: Example annotation from SLURP dataset.

Spoken Language Understanding Resource Package (SLURP), a publicly available multi-domain dataset for E2E-SLU, which is substantially bigger and more diverse than existing SLU datasets. SLURP is a collection of ~72k audio recordings of single turn user interactions with a home assistant, annotated with three levels of semantics: Scenario, Action and Entities, as in Fig. 1, including over 18 different scenarios, with 46 defined actions and 55 different entity types as listed on https://github.com/pswietojanski/slurp.[1]

In order to further support SLU development, we propose SLU-F1, a new metric for entity prediction, which is specifically designed to assess error propagation in structured E2E-SLU tasks. This metric has 3 main advantages over the commonly used accuracy/F1 metric, aimed at supporting SLU developers: First, it computes a distribution rather than a single score. This distribution is (1) inspectable and interpretable by system developers, and (2) can be converted into a confidence score which can be used in the system logic (akin to previously available ASR confidence scores). Finally, the distribution reflects errors introduced by ASR and their impact on NLU and thus (3) gives an indication of the scope of improvement that can be gained by E2E approaches. Using this metric, we evaluate 4 baseline systems that represent competitive

[*]Authors contributed equally.

[1]Note that Action & Entities are also referred to as 'Intent'. Entities consist of 'Tags' and 'Fillers', aka. 'Slots' and 'Values'.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7252–7262,
November 16–20, 2020.

pipeline approaches, i.e. 2 state-of-the-art NLU systems and 2 ASR engines. We conduct a detailed error analysis of cases where E2E could have made a difference, i.e. error propagation and semantic disambiguation.

2 Related Work

The first corpora containing both audio and semantic annotation reach as far back as the The Air Travel Information System (ATIS) corpus (Hemphill et al., 1990) and the Switchboard-DAMSL Labeling Project (Jurafsky and Shriberg, 1997). However, it was not until recently when the first E2E approaches to SLU were introduced (Serdyuk et al., 2018; Haghani et al., 2018). Since then, one of the main research questions is how to overcome data sparsity by e.g. using transfer learning (Schuster et al., 2019; Tomashenko et al., 2019), or pre-training (Lugosch et al., 2019b). Here, we present a new corpus, SLURP, which is considerably bigger than previously available corpora. In particular, we directly compare our dataset to the two biggest E2E-SLU datasets for the English language: The Snips benchmark (Coucke et al., 2018) and the Fluent Speech Command (FSC) corpus (Lugosch et al., 2019b). With respect to these resources, SLURP contains ~6 times more sentences than Snips, ~2.5 times more audio examples than FSC, while covering 9 times more domains and being on average 10 times lexically richer than both FSC and Snips, see Section 3.3. SLURP represents the first E2E-SLU corpus of this size for the English language. The only existing comparable project is represented by the CASTLU dataset (Zhu et al., 2019) for Chinese Mandarin.

3 SLURP data

3.1 Data Collection

SLURP was collected for developing an in-home personal robot assistant (Miksik et al., 2020). First, we collected textual data by prompting Mechanical Turk (AMT) workers to formulate commands towards the robot, using 200 pre-defined prompts such as "How would you ask for the time/ set an alarm/ play your favourite music?" etc. We carefully designed the prompts to avoid lexical priming and thus increase linguistic variability of the collected data. This data has been manually annotated at scenario, action and entity level, and released as a text-only NLU benchmark (Liu et al., 2019). The

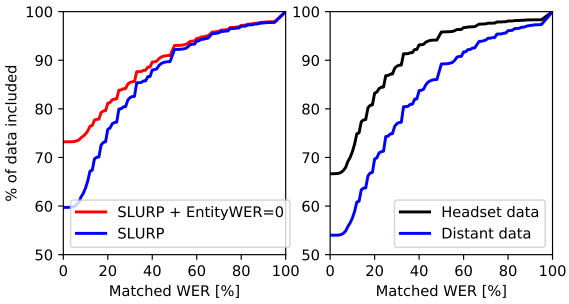


Figure 2: Amounts of data in SLURP matching given WER levels.

textual data also serves as gold standard transcriptions for the audio data.

The audio data was collected in acoustic conditions matched to a typical home or office environment. We asked 100+ participants to read out the collected prompts on a tablet and to provide demographic background information, see Table 1. Speech was captured at distance with a microphone array, but some users were also equipped with a close-talking headset microphone (though, distant and close-talk channels are not synchronised at the sample level). Most recording sessions lasted 1 hour and were split into 4 parts. In each part, the technician changed position of the microphone array in the collection place. Users were encouraged to vary their location in the room from utterance to utterance (seating, standing or walking), and for some utterances not to speak directly to the mic array in order to resemble realistic conditions. These parameters are not logged with the dataset, however, they do pose increased challenges for ASR (Marino and Hain, 2011).

Female	Male	Native	Non-Native	Unk.
37.3%	32.2%	25.5%	44%	30.5%

Table 1: Participants' demographic statistics.

3.2 Audio Data Processing

For quality control of the audio data, we automatically verified i) whether the participant uttered the right / complete SLU query as prompted and ii) if the files were appropriately end-pointed. We used the transcriptions of two ASR systems (referred to as Multi-ASR and Google-ASR, see Sec 5.1). These systems were not estimated from SLURP acoustic data, thus remain unbiased and do not reinforce potential errors. First, we removed all data that failed to force-align to transcripts using Multi-ASR. Then for the remainder we derived

the SLU related confidences based on the matched Word-Error Rate (WER) between textual prompts and the obtained ASR hypotheses (calculated for both utterance and entity fillers), as well as cross-mic validation between close and distant microphones, see Figure 2 (Right). Note that the higher matched WER does not necessarily imply the file lacks the expected content, as simply the file could be more challenging to automatically recognise. At the same time, from SLU perspective, one does not necessarily need grammatically correct utterances, as long as the they carry the information necessary to understand and execute the query. Figure 2 (Left) shows that for nearly 60% of the data at least one ASR system achieved a perfect score (WER=0), and this increases to ~73% after including utterances with imperfect sentence error rates but correct entity fillers (EntityWER=0). After filtering, SLURP comprises ~58 hours of acoustic material. See Table 2 for detailed statistics.

In addition, we provide SLURP-synth following (Lugosch et al., 2019a), where we replace filtered or missing recordings with synthetic vocalisations from Google's Text-to-Speech system[2] using 34 different synthetic English voices.

3.3 Linguistic Analysis and Comparison

In this Section, we compare SLURP with the most recent publicly available E2E-SLU datasets: The Fluent Speech Command (FSC) corpus (Lugosch et al., 2019b) and the Snips benchmark (Coucke et al., 2018), which are also set in the smart-home domain. Snips covers 10 domains. However, only 2 domains have been vocalised, resulting in ~6K audio files. FSC, on the other hand, is considerably bigger than Snips in terms of audio recordings, including ~30k vocalisations. However, the provided semantics only cover a small subset of actions with no more than two fixed entity types as arguments. In the following, we compare these dataset along four dimensions in order to get a first estimate of SLURP's level of complexity.

Audio analysis: Table 2 summarises the audio data for each dataset. Audio files are differentiated in *close* and *far* range microphone. As shown, SLURP has ~1.8× more speakers, more than double the audio files than the biggest dataset FSC, however FSC has an higher audio-per-sentence ratio. Demographic statistics are reported in Table 1.

[2] https://cloud.google.com/text-to-speech

	FSC	Snips	SLURP	SLURP -synth
Speakers	97	69	**177**	34
Audio files	30,043	5,886	**72,277**	69,253
– Close range	30,043	2,943	**34,603**	–
– Far range	–	2,943	**37,674**	–
Audio/Sentence	**121.14**	2.02	4.21	3.87
Duration [hrs]	19	5.5	**58**	43.5
Avg. length [s]	2.3	**3.4**	2.9	2.3

Table 2: Audio file statistics.

Lexical analysis: Table 3 provides an overview of different measures of lexical richness and diversity, following (Novikova et al., 2017), using both lexicalised (LEX) and delexicalised (DELEX) versions of the datasets (delexicalisation is performed by replacing each entity span with the entity label). Note that delexicalisation has a more severe effect on FSC and Snips, which indicates that most of their lexical richness and diversity stems from entity names. On average SLURP has ~100× more tokens, lemmas, bigrams and trigrams than FSC, and ~10× more than Snips. In addition, we compute the following lexicographic measures using the Lexical Complexity Analyser (Lu, 2012). *Lexical Sophistication* (LS2) (Laufer, 1994) is defined as T_s/T, with T_s being the number of sophisticated types of (unique) words[3] and T being the number of types of words in a dataset. The *Corrected Verb Sophistication* (CSV1) (Wolfe-Quintero et al., 1998) is evaluated as $T_{svb}/\sqrt{2N_{vb}}$, with T_{svb} the number of types of sophisticated verbs and N_{vb} the total number of verbs in a dataset. The *Mean Segmental Text-to-Token Ratio* (MSTTR) (Johnson, 1944) is the average Text-to-Token Ratio (TTR – T/N) over all the segments of 10^4 words, with N the number of words in a dataset. The MSTTR is used to capture the variation of classes of words. Again, SLURP shows higher levels of lexical sophistication and richness than the other datasets, especially in the delexicalised case. Note that lexicalised version of Snips contains many names of artists and bands in the music scenario, which contributes to enlarge the set of sophisticated words T_s. The only measure where SLURP doesn't outperform the other datasets is average sentence length. SLURP contains, among others, shorter interactions, such short acknowledgements, elliptic questions and atomic commands, whereas Snips is

[3] Sophisticated words are considered words not in the 2000 more frequent words in English language.

[4] Standard size of a segment for written text is 50, but we are here considering short utterances, so we lowered this number to 10.

	FSC		Snips		SLURP		SLURP-synt	
	LEX	DELEX	LEX	DELEX	LEX	DELEX	LEX	DELEX
Sentences	248	190	2,912	1,437	**17,181**	**15,433**	19,711	16,707
Average Sentence length	4.49	–	**7.48**	–	6.93	–	7.27	–
Distinct Tokens	96	89	2,182	271	**6,467**	**3,774**	5,974	3,553
Distinct Tokens occurring once	31	36	1825	120	**3,007**	**1,778**	2,799	1,676
Distinct Lemmas	102	92	2193	250	**5,501**	**3,080**	5,119	2,920
Distinct Bigrams	218	182	4,004	1,355	**32,303**	**21,724**	28,988	20,308
Distinct Bigrams occurring once	97	97	3,066	698	**21,997**	**14,095**	19,360	12,637
Distinct Trigrams	250	198	5,703	2,408	**50,422**	**37,417**	45,631	35,548
Distinct Trigrams occurring once	131	119	4,499	1,543	**40,184**	**28,393**	34,856	25,553
Lexical Sophistication (LS2)	0.35	0.31	**0.87**	0.41	0.79	**0.69**	0.79	0.68
Corrected Verb Sophistication (CVS1)	0.42	0.38	0.72	0.59	**5.17**	**3.54**	4.58	3.20
Mean segmental TTR (MSTTR)	0.71	0.82	0.78	0.86	**0.92**	**0.96**	0.93	0.96

Table 3: Analysis of Lexical diversity and sophistication.

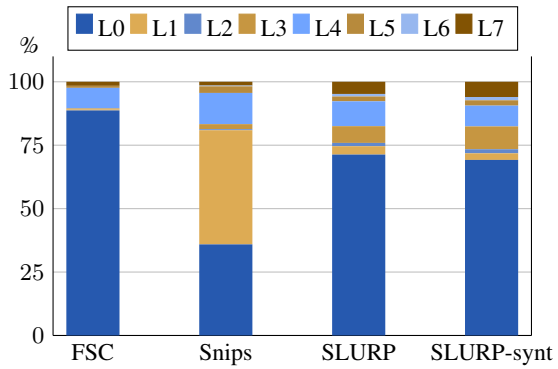


Figure 3: Syntactic complexity on D-Level scale, where higher levels correspond to more complex, deeper syntactic structures.

	FSC	Snips	SLURP	SLURP -synt
Scenarios	2	2	**18**	18
Actions	6	7	**46**	54
Entities	2	4	**56**	56
Tot. Entities	334	2,870	**16,792**	14,623
Entity/Sentence	**1.35**	0.98	0.97	0.65
Unique Entities	16	1,348	**5,613**	4619

Table 4: Semantic analysis of the number of scenarios, actions and entity types, the total number of annotated entities, and the number of unique entities, i.e. entities whose lexical filler appears only once.

mostly composed of commands of similar length, often including multiword named entities.

Syntactic analysis: Next, we use the D-Level Analyser (Lu, 2009) to evaluate the syntactic complexity of user utterances according to the revised D-Level scale (Covington et al., 2006), where higher levels correspond to more complex, deeper syntactic structures, e.g. 0-1 levels include simple sentences, while higher levels presents embedded structures, subordinating conjunction, etc. Figure 3 shows the percentages on the D-Level scale for each dataset. Overall, all the datasets present a majority of Level 0 and 1 sentences. This can be explained with the nature of the application domain, i.e. a smart-home assistant. FSC contains mostly Level 0 sentences (~89%), with some (~9%) Level 4 ones. 89% of Snips sentences fall into Level 0 and 1, against only 74% of SLURP. The remaining 11% of Snips are mostly Level 4 sentences, while SLURP appears more mixed, with even a ~5% of Level 7 sentences.

Semantic Analysis: Finally, we compare the datasets according to their semantic content. SLURP is annotated with three layers of semantics, namely *scenarios*, *actions* and *entities*, where each sentence is annotated with one scenario and one action, see Fig. 1, similar to annotations used in (Budzianowski et al., 2018; Schuster et al., 2019). FSC and Snips contain actions and entities as well, although they do not explicitly annotate the scenarios, however these can be deducted from the dataset file structure. The results in Table 4 show that SLURP's semantic coverage is 9 times wider than other datasets in terms of scenarios, and ~6.5 times in terms of actions, where a higher number of scenarios results in a higher number of actions. FSC has the highest entity/sentence ratio, though it only has 16 unique entities. Snips appears to be the dataset with highest Unique Entities/Total Entities ratio, ~50%, against ~33% of SLURP. Again, this is due to the frequent use of proper names.

4 SLURP Metrics

The standard metric for evaluating E2E-SLU is accuracy, which is defined as "the accuracy of all slots for an utterance taken together – that is, if the predicted intent differs from the true intent in even one slot, the prediction is deemed incorrect" (Lugosch et al., 2019b). However, this notion of accuracy is problematic when it comes to evaluating entities, as it does not account for the interplay between semantic mislabelling and textual misalignment. Nor does it differentiate between entity label and lexical

Gold: [event_name: brunch], [date: Saturday], [timeofday: morning], [person: Aaronson]
SLU: [event_name: brunch], [date: Saturday], [date: morning], [person: Aron's son]

Figure 4: Continued example from Figure 1: Errors in SLU entity tagging.

filler, as in Fig. 4, where lexical filler is defined as span over tokens in the original sentence.[5]

Formally, given a sentence s, let $\mathcal{E}$ and $\hat{\mathcal{E}}$ be the set of gold and predicted entities, respectively. Each $e_i = \langle l_i, f_i \rangle \in \mathcal{E}$ is a tuple where $l_i \in \mathcal{L}$ is the label drawn from the list of available entity labels $\mathcal{L}$, while $f_i = [t_m, \ldots, t_n]$ is the lexical filler, defined as a span of consecutive tokens of s such that $1 \leq m \leq n \leq |s|$. Similarly, predicted entities are of the form $\hat{e}_k = \langle \hat{l}_k, \hat{f}_k \rangle \in \hat{\mathcal{E}}$. In span-based metrics, two entities e_1 and e_2 are identical ($e_1 =:= e_2$) when both labels and lexical fillers are the same ($l_1 = l_2 \wedge f_1 = f_2$). A match is thus found only whenever the gold and predicted entities are identical, i.e. $e_i =:= \hat{e}_k$. This evaluation method holds in NLU because entities are tagged over the same textual sequence. When evaluating E2E-SLU, where entities are identified out of a wave form, this strict coupling with the token sequence may no longer apply. Note that pipeline systems for SLU are affected as well since they operate over ASR transcribed sentences, which can consistently differ from the original gold transcription.

To account for this mismatch, we propose SLU-F1, a new metric which does not overly penalise misalignments caused by ASR errors. In addition, it is able to capture the quality of transcriptions and entity tagging errors at the same time in a single metric. As such, this metric allows to directly compare E2E and pipeline systems. In particular, SLU-F1 combines span-based F1 evaluation with a text-based distance measure *dist*, e.g. WER. The equality property $=:=$ is relaxed by allowing gold and predicted entities (e_i and $\hat{e}_k$) to match ($e_i =:= \hat{e}_k$) when the corresponding labels are identical ($l_i = \hat{l}_k$), even when the fillers are not identical. In this case we increment the True Positives (TPs) by 1. To account for lexical distance/mismatch, we compute the *dist* between gold and predicted fillers ($dist(f_i, \hat{f}_k)$), and increment the False Positives (FPs) and False Negatives (FNs) of this amount, as in Algorithm 1. In the case of a predicted entity label matching with more than one gold entity, e.g. when two or more entities with the same label are present, we opt for a non-conservative approach, selecting the gold annotation minimising the *dist* as a candidate. The assumption is that the pair of entities is most likely referring to the same text span. We use two distance functions to capture different aspects of possible transcription mistakes: WER (Word-F1) and the normalised Levenshtein distance on character level (Char-F1). WER is a strict token-level metric, which outputs errors/null matches whenever a mismatching or misalignment of tokens is observed. The character-based Levenshtein distance, on the other hand, offers the opposite perspective. By computing character-based similarities, it is much less susceptible to small variations of input strings, and thus better accounting for local transcription errors which do not affect NLU tagging. For example, Word-F1 will penalise small morphological differences e.g. singular vs. plural as in *pizza* vs. *pizzas*, which are often seen in transcriptions. This over-penalises NLU outputs, e.g. the tagging of *pizzas* may be semantically correct. Char-F1 on the other hand does not over-penalise NLU, but it also may provide a positive score when two fillers have similar characters, but are semantically and phonetically unrelated. In other words, Word-F1 shows the influence of ASR on NLU, whereas Char-F1 gives an indication of NLU performance despite transcription noise. These *dist*-F1 metrics ($dist =$ Word or Char) metric are similar to the fuzzy matching mechanism proposed in (Rastogi et al., 2020). They fundamentally differ for the adopted string matching schema: any *dist*-F1 considers string ordering to score string similarity, while the fuzzy mechanism is instead order invariant.

Consider the illustrative entity tagging example in Figure 4. Here, *Aaronson* has been wrongly transcribed into *Aron's son*, and *morning* has been wrongly tagged with date. A $dist-$F1 will score the predicted entities as follows: both [event_name: *brunch*] and [date: *Saturday*] contribute with a $+1$ to the TPs, since both label and filler correspond to gold information. The wrong label associated with *morning* increases the FPs of 1, although it is correctly transcribed. It follows that the entity timeofday is not predicted, increasing the FNs of 1. Finally, [person: *Aron's son*] is correctly labelled, but its filler is

[5] In traditional NLU systems this is identified with pairs of start-end tokens or chars, or token index spans.

partially wrong. It thus contributes to the TPs by 1, but FPs and FNs are both incremented by $dist(Aaronson, Aron's\ son)$.

Algorithm 1 $dist$-F1 for a sentence s

Input $\mathcal{E}, \hat{\mathcal{E}}$,
$TP, FP, FN \leftarrow 0$
$\mathcal{L}_s \leftarrow$ set of gold entity labels in s
$dist \leftarrow$ a text-based distance metric
Output: TP, FP, FN

for each $\hat{e} \in \hat{\mathcal{E}}$ **do**
  if $\hat{e}.label \in \mathcal{L}_s$ **then**
    $\mathcal{P}_l \leftarrow \{(e, \hat{e}) \mid \forall e \in \mathcal{E}.\ e.label = \hat{e}.label\}$
    if $\mathcal{P}_l.size > 0$ **then**
      $(e, \hat{e}) \leftarrow \arg\min_{(e,\hat{e}) \in \mathcal{P}_l} dist(e, \hat{e})$
      TP += 1
      FP += $dist(e.filler, \hat{e}.filler)$
      FN += $dist(e.filler, \hat{e}.filler)$
      $\mathcal{E}.remove(e)$, $\hat{\mathcal{E}}.remove(\hat{e})$
    else
      FP += 1, $\hat{\mathcal{E}}.remove(\hat{e})$
    end if
  else
    FP += 1, $\hat{\mathcal{E}}.remove(\hat{e})$
  end if
end for
for $e \in \mathcal{E}$ **do**
  FN += 1, $\mathcal{E}.remove(e)$
end for

Finally, we combine Word-F1 and Char-F1 in a single number SLU-F1, which evaluates the final performance over the sum of the confusion matrices obtained with Word-F1 and Char-F1.[6]

5 Experiments

We now establish the performance of different baseline systems on the SLURP corpus. As demonstrated in Section 3.1, SLURP is linguistically more diverse than previous datasets, and therefore more challenging for SLU. We first provide an evaluation of two ASR baselines to show the complexity of the acoustic dimension. We then evaluate the semantic dimension, by testing the corpus against state-of-the-art NLU systems. We finally combine ASR and NLU, implementing several SLU pipelines.

Note that so far, the direct comparison of E2E-SLU with pipeline approaches are mainly limited to baselines developed on the same dataset, e.g. a multistage neural model in which the two stages that correspond to ASR and NLU are trained independently, but using the same training data (Desot et al., 2019; Haghani et al., 2018). We follow a different approach, which, as we argue, is closer to the real-life application scenario: We use competitive ASR systems and state-of-the-art NLU systems.

5.1 Acoustic evaluation

We run the analysis of the SLURP acoustic complexity by testing 2 different ASR systems: In-domain ASR trained on SLURP data, and Multi-ASR, which leverages a large amount of out-of-domain data. Both are built with the Kaldi ASR toolkit (Povey et al., 2011). Multi-ASR is a large-scale system estimated from publicly available acoustic data pooled together – Acoustic data including, among others, LibriSpeech (Panayotov et al., 2015), Switchboard (Godfrey et al., 1992), Fisher (Cieri et al., 2004), CommonVoice (Ardila et al., 2019), AMI (Carletta, 2007) and ICSI (Janin et al., 2003),[7] which is further augmented to increase environmental robustness following (Ko et al., 2017). In total, a time-delay neural network acoustic model (Peddinti et al., 2015) is trained on 24,000 hours of augmented audio material with lattice-free maximum mutual information objective (Povey et al., 2016). For decoding, we use a tri-gram Language Model (LM) that is an interpolation of an in-domain LM estimated from 60k voice-command sentences[8] and a background LM estimated from Fisher transcripts. As shown in the first block of Table 5, Multi-ASR offers a competitive performance on this data when compared to the off-the-shelf Google-ASR.[9]

SLURP-ASR shares the overall pipeline with Multi-ASR, except the acoustic model is estimated from the 40 hours of SLURP training data (83 hours when pooled with SLURP-Synth) and bootstrapped from forced-alignments obtained with Gaussian mixture model build for Multi-ASR. Results for this scenario are reported in the second block of Table 5, where adding synthetic data shows 1.6% improvement. For comparison, estimating acoustic models from synthetic data alone (no augmentations) results in 98% WER on Test partition.

Finally, we perform supervised acoustic domain adaptation (Bell et al., 2020) of Multi-ASR with SLURP-Train by a method proposed in (Swietojanski et al., 2016), which achieves the best perfor-

[6] The official script for analysis and evaluation will be released with SLURP at https://github.com/pswietojanski/slurp.

[7] System build while third author was with Emotech LTD.

[8] This includes SLURP-Train and additional 50k sentences that has been collected, but not annotated for NLU purposes.

[9] https://cloud.google.com/speech-to-text/ tested on 20/05/2020 using the command_and_search model. Note, that these systems are not directly comparable as Multi-ASR benefits from speaker adaptation, and an in-domain LM data.

	Dev	Test
Google-ASR	24.0	24.7
Multi-ASR	16.7	17.3
SLURP-ASR (Train)	23.7	23.8
SLURP-ASR (Train + Synth)	22.4	22.2
Multi-ASR + Adapt w/ SLURP	**16.3**	**16.2**

Table 5: SLURP WER for different ASR systems.

mance by around 1% absolute on Test.

In sum, the large out-of-domain Multi-ASR system performs better than the systems trained on in-domain SLURP data. Best results are achieved by using a pre-training approach, i.e. Multi-ASR adapted to SLURP. This shows that, despite SLURP's absolute size, the acoustic data is still too scarce to fully account for its lexical richness and noise conditions. As such, SLURP is a challenging dataset for ASR as well as for SLU.

5.2 Semantic evaluation

System Descriptions: We evaluate SLURP against two state-of-the-art NLU models: HerMiT (Vanzo et al., 2019) and SF-ID (E et al., 2019). Both systems achieved state-of-the-art results on the NLU Benchmark (Liu et al., 2019) and on ATIS/Snips respectively. HerMiT's architecture is a hierarchy of self-attention mechanisms and Bidirectional Long Short-Term Memory (BiLSTM) encoders followed by Conditoinal Random Field (CRF) tagging layers. Its multi-layered structure resembles a top-down approach of Scenario, Action and, Entity prediction, where each task benefits from the information encoded by the previous stages, e.g. Entity detection can benefit from sentence-level encodings.

SF-ID's architecture is also based on attention, using a BiLSTM encoder and CRF tagger. The model defines two subnets that communicate through a reinforce vector. In order to compare with HerMiT's top-down approach, we choose the opposite Entity-first propagation direction for SF-ID, i.e. the entity detection task is executed first and its encodings are used to feed the Intent detection task. Note that while HerMiT uses a multi-layered annotation scheme (Scenario and Action), SF-ID can only handle a single layer of annotation. To this end, we generate another combined semantic layer, Scen_Act, to feed SF-ID with a label composed by the concatenation Scenario and Action.

Scenario and Action Prediction: We split SLURP in train, development and test as in Table 6.

	Train	Dev	Test
Sentences	11514	2033	2974
Audio files	50628	8690	13078
Tot. Entities	11367	2022	2823
Entity/Sentence	0.98	0.99	0.95
Total duration [hours]	40.2	6.9	10.3

Table 6: Data distribution of train, dev and test sets.

	Scenario	**Action**	**Scen_Act**
Gold/HerMiT	**90.15**	**86.99**	**84.84**
Gold/SF-ID	86.48	83.69	82.25
Multi/HerMiT	83.73	79.70	76.68
Multi/SF-ID	81.90	77.72	75.87
Google/HerMiT	81.68	76.58	73.41
Google/SF-ID	78.87	74.31	72.06
SLURP/HerMiT	82.31	78.07	74.62
Multi-SLURP/HerMiT	**85.69**	**81.42**	**78.33**

Table 7: System accuracy of Scenario and Action.

We first evaluate accuracy for Scenario, Action and a combination of the two. Table 7 summarises the results, where the top two rows are upper bounds based on gold transcriptions. Note that even for the gold transcriptions, both NLU systems perform substantially below their state-of-the-art results on the NLU benchmark (HerMiT=87.55) and Snips respectively (SF-ID= 97.43). This further demonstrates the complexity of SLURP, which also makes it a challenging test bed for future research not only for SLU, but also NLU. When moving on to ASR transcribed data, the results in the middle of Table 7 show the Multi-ASR system in combination with HerMiT achieves top performance for all 3 tasks. Finally, the 3rd block reports HerMiT with ASR from in-domain SLURP audio data (also see Table 5). The results show that our best performing system, HerMiT with Multi-ASR + Adapt w/ SLURP, is only ~5% below the gold standard despite 16% WER. We hypothesise that this is due to robust Scenario and Action encodings, which we will further examine in our error analysis in Section 6.

Entity Prediction: We now analyse the results for entity prediction in more detail using our proposed metric SLU-F1. The results in Table 8 confirm that HerMiT is the stronger NLU system on gold-transcribed data and outperforms the other system combinations for SLU in combination with Multi-ASR. Again, these results suggest that the top-down information flow of HerMiT (i.e. first decoding Scenario, then Action and lastly Entity in a sequence) is better suited for this complex dataset, which we will further demonstrate in the following.

	Word-F1	Char-F1	SLU-F1	F1
Gold/HerMiT	–	–	–	**78.19**
Gold/SF-ID	–	–	–	69.87
Multi/HerMiT	67.78	71.38	69.53	62.69
Multi/SF-ID	65.82	68.92	67.33	60.15
Google/HerMiT	64.01	68.12	66.00	58.00
Google/SF-ID	62.73	65.37	64.02	56.54
SLURP/HerMiT	65.48	68.56	66.99	59.79
Multi-SLURP/HerMiT	**69.34**	**72.39**	**70.84**	**64.16**

Table 8: System performance on entity prediction

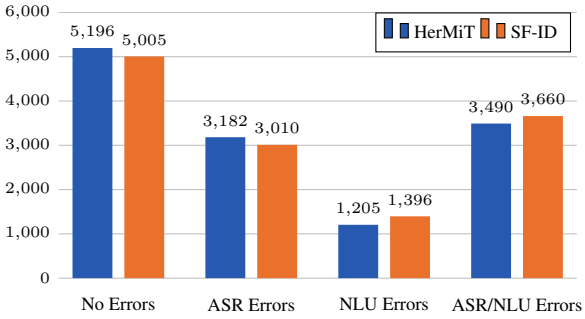


Figure 5: Error propagation: *No Errors* refer to the number of predicted entities that match the gold transcriptions perfectly. *ASR Errors* count the number of predictions where ASR outputs an unmatched candidate but the NLU system is nevertheless able to recover the correct entities from the transcriptions. *NLU Errors* count sentences where transcriptions are correct, but entities do not match. *ASR/NLU Errors* count the sentences where both ASR and NLU errors are present.

6 Error Analysis

6.1 Analysis of Error Propagation for different NLU Approaches

We further describe the types of errors produced by HerMiT and SF-ID for Entity Prediction on noisy ASR data, as shown in Figure 5. Overall, HerMiT has lower error rates for all but ASR errors. Nevertheless, it is able to recover the correct entities from the transcriptions. These results indicate that HerMiT, using a top-down decoding approach – going from the more general Scenario to the more specific Action and Entity Prediction, is more robust to noise propagation than the bottom-up SF-ID system.

6.2 Expressiveness of the SLU-F1 Metric

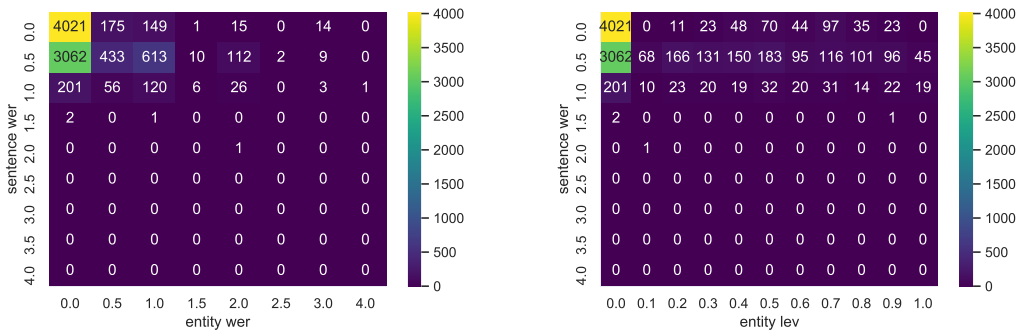
The results in Table 8 show that our proposed metrics Word-F1 and Char-F1 both produce the same ordering as F1. However, a Pearson's correlation between Word-F1 and Char-F1 shows that the two metrics are only weakly correlated ($\rho = 0.2$, $p \ll 0.0001$), which confirms that they are indeed measuring two different aspects despite producing the same final ordering. In addition to an overall performance score, the metrics give us a distribution of value ranges, which can give us insight on system behaviour. Figure 6 shows distributions of entity-level $dist$ value ranges over the WER of the sentence for our top performing system HerMiT/Multi-ASR. For entity-WER (Figure 6a), the distribution shows high density of entities falling between sentence-WER$= [0, 1]$ and entity-WER$= [0, 1]$. When analysing sentences with correct transcriptions, i.e. sentence-WER=0, we find only NLU errors, due to span misalignments. When sentence-WER > 0, most of the entities are scored with a values either in $(0, 0.5]$, or in $(0.5, 1]$. In the first case, we find NLU mistakes caused by shortening entity spans, e.g. "football" instead of "football match". The second range includes span shortening and extensions, e.g. "Saturday morning" instead of "Saturday", as well as many mis-transcribed entities, e.g. due to either morphological errors (singular vs. plural), or transcription errors.

The distribution for entity-level normalised Levenshtein is less spiked, as shown in Fig. 6b. As for WER, all the entries with sentence-WER=0 and entity-Lev>0 correspond to correctly labelled entities, whose span has been shortened or extended. Entities assigned with character-based Lev values falling between $(0, 0.2]$ mostly contain negligible ASR errors, such as morphological errors, compound merging or explosion, or general transcription mistakes, e.g. *Sara* vs. *Sarah*. Entities with Lev$= (0.2, 0.5]$ comprise both ASR errors, as well as including minor NLU errors such as shortened or extended entity spans. When entity-Lev$= (0.5, 0.8]$, we find mostly NLU errors due to wrong span tagging. Finally, two types of NLU errors fall in the range $(0.8, 1.0]$: Either span errors with a substantial mismatch in length with gold annotations, or more severe ASR errors.

7 Discussion

SLURP is not only bigger, but also a magnitude more challenging than previous datasets. The purpose of this new data release is not to provide yet another benchmark dataset, but to provide a use-case inspired new challenge, which is currently beyond the capabilities of SOTA E2E approaches (due to scalability, lack of data efficiency, etc.).

We have tested several SOTA E2E-SLU systems on SLURP, including (Lugosch et al., 2019b) which



(a) (b)

Figure 6: Correlation between sentence-level WER (intervals of 0.5) and entity-level (a) WER values (intervals of 0.5), (b) normalised character-based Levenshtein values (intervals of 0.1).

produces SOTA results on the FSC corpus. However, re-training these models on this more complex domain did not converge or result in meaningful outputs. Note that these models were developed to solve much easier tasks (e.g. a single domain). Developing an appropriate model architecture is left for future work. For this reason, in this work we focus on benchmarking existing approaches.

We show that SOTA modular approaches are able to provide a strong baseline for this challenging data, which has yet to be met by SOTA E2E systems. We also argue that our modular baseline is closer to how real-world applications build SLU systems, nevertheless often overlooked when testing E2E systems. As such, we consider our SOTA modular baseline a major novel contribution.

8 Conclusion

In this paper, we present SLURP, a new resource package for SLU. First, we present a novel dataset, which is substantially bigger than other publicly available resources. We show that this dataset is also more challenging by first conducting a linguistic analysis, and then demonstrating the reduced performance of state-of-the-art ASR and NLU systems. Second, we propose the new SLU-F1 metric for evaluating entity prediction in SLU tasks. In a detailed error analysis we demonstrate that the distribution of this metric can be inspected by system developers to identify error types and system weaknesses. Finally, we analyse the performance of two state-of-the-art NLU systems on ASR data. We find that a sequential decoding approach for SLU, which starts from the more abstract notion of scenario and action produces better results for entity tagging, than an approach which works bottom up, i.e. starting from the entities. Our error analysis suggests that this is due to the former approach being able to better account for noise by priming entity tagging, which is a more challenging task than scenario or action recognition.

In future work, we hope that SLURP will be a valuable resource for developing E2E-SLU systems, as well as more traditional pipeline approaches to SLU. The next step is to extend SLURP with spontaneous speech, which would again increase its complexity, but also move it one step closer to real-life applications.

Acknowledgements

Thanks to Emotech Ltd and H. Zhuang for agreeing to release this data for research purposes. Special thanks to P. Mediano, M. Zhou and X. Chen for help with designing and organising data collection. This research received funding from the EPSRC project MaDrIgAL (EP/N017536/1), as well as Google Research Grant to support NLU and dialog research at Heriot-Watt University.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus.

Peter Bell, Joachim Fainberg, Ondrej Klejch, Jinyu Li, Steve Renals, and Pawel Swietojanski. 2020. Adaptation algorithms for speech recognition: An overview.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, and Milica Gašić. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

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. *ArXiv*, abs/1805.10190.

Michael Covington, Congzhou He, Cati Brown-Johnson, Lorina Naci, and John Brown. 2006. How complex is that sentence? a proposed revision of the rosenberg and abbeduto d-level scale.

Deborah A. Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-smith, David Pallett, Er Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the ATIS task: the ATIS-3 corpus. In *in Proc. ARPA Human Language Technology Workshop '92, Plainsboro, NJ*, pages 43–48. Morgan Kaufmann.

T. Desot, F. Portet, and M. Vacher. 2019. Slu for voice command in smart home: Comparison of pipeline and end-to-end approaches. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 822–829.

Haihong E, Peiqing Niu, Zhongfu Chen, and Meina Song. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5467–5471, Florence, Italy. Association for Computational Linguistics.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, page 517–520, USA. IEEE Computer Society.

Parisa Haghani, Arun Narayanan, Michiel Bacchiani, Galen Chuang, Neeraj Gaur, Pedro J. Moreno, Rohit Prabhavalkar, Zhongdi Qu, and Austin Waters. 2018. From audio to semantics: Approaches to end-to-end spoken language understanding. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 720–726.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The atis spoken language systems pilot corpus. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '90, page 96–101, USA. Association for Computational Linguistics.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Wendell Johnson. 1944. *Studies in Language Behavior: I. A Program of Research*. National Foreign Language Center Technical Reports. American Psychological Association, Psychological Monographs: General and Applied.

Dan Jurafsky and Elizabeth Shriberg. 1997. Switchboard-damsl labeling project coder"s manual.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224. IEEE.

Batia Laufer. 1994. The lexical profile of second language writing: Does it change over time? *RELC Journal*, 25(2):21–33.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. In *10th International Workshop on Spoken Dialogue Systems Technology*.

Xiaofei Lu. 2009. Automatic measurement of syntactic complexity in child language acquisition.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Loren Lugosch, Brett Meyer, Derek Nowrouzezahrai, and Mirco Ravanelli. 2019a. Using speech synthesis to train end-to-end spoken language understanding models. abs/1910.09463.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019b. Speech model pre-training for end-to-end spoken language understanding. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 814–818. ISCA.

Davide Marino and Thomas Hain. 2011. An analysis of automatic speech recognition with multiple microphones. In *INTERSPEECH*.

O. Miksik, I. Munasinghe, J. Asensio-Cubero, S. Reddy Bethi, S-T. Huang, S. Zylfo, X. Liu, T. Nica, A. Mitrocsak, S. Mezza, R. Beard, R. Shi,

R. Ng, P. Mediano, Z. Fountas, S-H. Lee, J. Medvesek, H. Zhuang, Y. Rogers, and P. Swietojanski. 2020. Building proactive voice assistants: When and how (not) to interact. *CoRR*, abs/2005.01322.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely sequence-trained neural networks for asr based on lattice-free mmi.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Kumar Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Schema-guided dialogue state tracking task at dstc8. In *AAAI Dialog System Technology Challenges Workshop*.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805. Association for Computational Linguistics.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. 2018. Towards end-to-end spoken language understanding. *CoRR*, abs/1802.08395.

Pawel Swietojanski, Jinyu Li, and Steve Renals. 2016. Learning hidden unit contributions for unsupervised acoustic model adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(8):1450–1463.

Natalia Tomashenko, Antoine Caubrière, and Yannick Estève. 2019. Investigating Adaptation and Transfer Learning for End-to-End Spoken Language Understanding from Speech. In *Interspeech 2019*, pages 824–828, Graz, Austria. ISCA.

Vanzo, Bastianelli, and Lemon. 2019. Hierarchical multi-task natural language understanding for cross-domain conversational AI: HERMIT NLU. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 254–263, Stockholm, Sweden. Association for Computational Linguistics.

K. Wolfe-Quintero, S. Inagaki, and H.Y. Kim. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy, & Complexity*. National Foreign Language Center Technical Reports. Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.

Su Zhu, Zijian Zhao, Tiejun Zhao, Chengqing Zong, and Kai Yu. 2019. Catslu: The 1st chinese audio-textual spoken language understanding challenge. In *2019 International Conference on Multimodal Interaction*, ICMI '19, page 521–525, New York, NY, USA. Association for Computing Machinery.