On Spoken Language Understanding Systems for Low Resourced Languages

Anonymous ACL submission

Abstract

Spoken dialog systems are slowly becoming and integral part of the human experience due to their various advantages over textual interfaces. Spoken language understanding (SLU) systems are fundamental building blocks of spoken dialog systems. But creating SLU systems for low resourced languages is still a challenge. In a large number of low resourced settings we don't have access to enough data to build automatic speech recognition (ASR) technologies, which are fundamental to any SLU system. Also, ASR based SLU systems do not generalize to unwritten languages. In this paper, we present a series of experiments to explore an extremely low resourced setting - something we refer to as a *true k-shot setting*, where we perform intent classification with systems trained on different values of k. We test our system on English and Flemish and find that even in such granular settings and no language specific ASR technology, we can create SLU systems that can be deployed in the real world.

1 Introduction

Spoken Language Understanding (SLU) systems form an integral part of spoken dialog systems. As shown in figure 1, a traditional SLU pipeline is made up of two modules - a speech to text module which converts input user audio into textual transcripts, and a natural language understanding (NLU) module which aims to understand the semantic content in the user utterance from the textual transcripts (Tur and De Mori, 2011; Lugosch et al., 2019). The conventional two-module SLU pipeline is prone to making ASR errors which propagate throughout the system. To minimize the automatic speech recognition (ASR) errors, a lot of recent research has been focused on creating end-to-end spoken language understanding (E2E-SLU) systems (Qian et al., 2017; Serdyuk et al., 2018). But building these E2E-SLU systems requires an even larger amount of annotated data when compared to two-module split SLU pipelines (Lugosch et al., 2019; Wu et al., 2020).

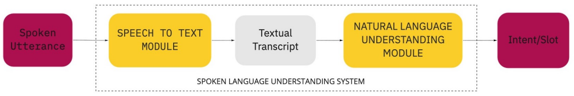
Spoken Utterance
SPEECH TO TEXT MODULE
Textual Transcript
NATURAL LANGUAGE UNDERSTANDING MODULE
Intent/Slot
SPOKEN LANGUAGE UNDERSTANDING SYSTEM

Figure 1: A traditional spoken language understanding system.

While high resourced languages like English are moving towards E2E-SLU, the challenges presented by low resourced languages are very different. Low resourced languages operate in a regime where we have access to tens or hundreds of labelled utterances, which are not enough to build robust E2E-SLU systems. Creating robust ASR systems for low resourced languages is itself a challenge as these require large amounts of manual annotation. For many low resourced languages, we might not even have enough data to create an ASR technologies. Creating ASR technologies for languages that have only a few hundred or a few thousand speakers alive and languages that have no written scripts, is not even a viable option. But can we create spoken dialog systems for such languages?

In this paper, we present a series of experiments to empirically re-create an extremely low resourced setting where each data point becomes valuable. We refer to this as a *true k-shot* setting. In this scenario, we pose an I-class intent classification problem ($I = 2, 4$) where we have a variable number speakers (S) available for recording training data. Each speaker provides only k-utterances per intent for training. In what we call a *true k-shot* setting, we evaluate our system in a granular manner for very small values of k. Specifically, we evaluate our system for each of $k = 1, 2, 3, 4, 5, 6, 7$. Also in such a low resourced setting, we realistically would not have access to an ASR system. Thus we use Allosaurus (Li et al., 2020), a universal phonetic transcription system that creates language independent representations of input speech. Allosaurus has been shown to produce state-of-the-art (SOTA) results for low resource languages like Sinhala and Tamil and reaches close to SOTA for high resourced languages like English Yadav et al. (2021); Gupta et al. (2021). We evaluate our SLU system on robust test sets containing hundreds of utterances collected from multiple speakers which are not present in the training set. We find encouraging performance without using language specific ASR technologies and with very small amounts of training data. Specifically, we find that even with as low as 7 speakers recording 7 audio samples per intent, we can create an SLU system that can be deployed in the real world with simple rule based dialog managers.

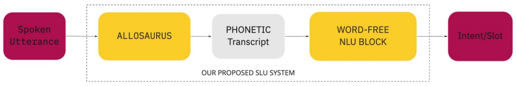

Figure 2: SLU system as proposed in (Gupta et al., 2020a).

2 Related Work

English has been the most widely studied language for creating SLU systems. Various datasets have been released to aid this development (Hemphill et al., 1990; Saade et al., 2018; Lugosch et al., 2019). There have been many previous works on creating SLU systems in a two-module split fashion (Gorin et al., 1997; Mesnil et al., 2014). A typical SLU pipeline, as shown in Figure 1, consists of an ASR system that converts input speech to text and an NLU module that processes the input text to understand the user query. As with any system composed of multiple modules, errors that occur in one part of the system propagate through the system. To prevent this, a large amount of recent work has been focussed on creating E2E-SLU systems (Qian et al., 2017; Serdyuk et al., 2018; Chen et al., 2018). The caveat with making such systems to work is that they require an even larger amount of task specific data.

One of the major bottlenecks in creating SLU systems for low resourced languages is the creation of ASR systems in low data scenario. Previous works have tried to use English-based ASR systems to convert input speech into a vector representation that can be processed by NLU systems (Buddhika et al., 2018; Karunanayake et al., 2019b,a). A series of recent works (Gupta et al., 2020b,a, 2021; Yadav et al., 2021) replace ASR module by a universal phone recognition system called Allosaurus (Li et al., 2020). Their proposed SLU pipeline is shown in Figure 2. Allosaurus provides language and speaker independent phonetic transcriptions and is this able to provide better representations of input audio which can also be used for languages linguistically distant from high resourced languages like English. (Gupta et al., 2021; Yadav et al., 2021) show that using Allosaurus based phonetic transcriptions to encode speech content outperforms previous state-of-the-art methods for Sinhala and Tamil. In our paper, we want to push Allosaurus to the limits and demonstrate performance in extremely low resourced settings, where each data point becomes crucial.

Language	Avg. Utterances per intent in Test Set	No. of Speakers in Test Set
English	135	10
Flemish	54	2

Table 1: Test set statistics.

3 Dataset

In our paper, we work with two languages - English and Belgian Dutch (Flemish). We use two popular SLU datasets for our experiments - the Fluent Speech Commands (FSC) dataset (Lugosch et al., 2019) for the English language and the Grabo dataset (Tessema et al., 2013; Ons et al., 2014; Renkens et al., 2014) for Flemish. FSC is a large and well maintained SLU dataset for the English language. The dataset contains 19 hours of speech data collected from 97 different speakers. The Grabo dataset contains 11 speakers and is much smaller than FSC. We refer the reader to the original papers releasing the datasets for further details.

The primary reason behind the choice of the datasets was that each utterance in the two datasets had clear speaker identities associated with them. Our aim is to test true low resourced settings where getting speaker recordings is extremely hard. Intent recognition datasets in other languages like French (Devillers et al., 2004; Saade et al., 2018), Chinese Mandarin (Zhu et al., 2019; Guo et al., 2021), Sinhala and Tamil (Karunanayake et al., 2019b) do not maintain speaker identities and hence were not suitable for our work. We choose Flemish to demonstrate performance for a low-resourced language setting since Flemish is not used to train

Allosaurus.

Moreover, these datasets also allow us to create large test sets such that the results are robust enough to evaluate the system performance and yet have no overlapping speakers with the training set. We experiment with two different intent classification problems containing I = 2,4 intents. The test set sizes are given in Table 1.

4 System and Model

We use the SLU system proposed in (Gupta et al., 2020a, 2021) for our experiments, as shown in Figure 2. It replaces a language specific ASR system with Allosaurus (Li et al., 2020), which is a universal phonetic transcription system. Allosaurus converts input speech to its phonetic transcriptions. We then build a word-free NLU system from these phonetic transcriptions to perform intent recognition.

The model used in this work is very similar to the model used in (Gupta et al., 2020a). (Gupta et al., 2020a) propose a model which uses Convolutional Neural Networks (CNN) (LeCun et al., 1998) to extract contextual information from phonetic input, and a Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) network to make utterance level decision and account for sequential information.

We reduce the model size to account for the scarcity of data. We use a 256-dimensional embedding layer with just one CNN layer of kernel size 3 and one LSTM layer of hidden dimension 256. Batch normalization (Ioffe and Szegedy, 2015) layer is removed because there are scenarios where we are working with a training set of as low as 2 samples, which are not enough to learn batch statistics and give unstable performance.

5 Experiments

In this paper, we try to emulate a real world data collection scenario for low resourced languages. Data collection is expensive, even more so in extremely low resourced languages. For example, Canadian Indigenious languages like Inuktitut or Siksika have only a few thousand living speakers. Native speakers of such languages are hard to catch hold of for data collection process. This makes every data point collected crucial.

Another challenging aspect of building SLU systems for low resourced languages is having access to language specific ASR systems. We bypass the need for language specific ASR systems by using Allosaurus (Li et al., 2020; Gupta et al., 2020a). We convert input audio to their language independent phonetic transcriptions, and intent recognition is performed using this phonetic transcription. Allosaurus was trained on the English but is not trained on Flemish, thus recreating the scenario where we don't have language specific speech-to-symbol conversion systems.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
$S=1$	61.74	77.10	72.89	71.38	83.73	82.53	85.94
$S=2$	70.18	81.62	82.22	86.14	91.56	87.34	87.75
$S=3$	68.07	85.84	**82.83**	87.65	87.34	90.60	**90.97**
$S=4$	82.22	85.24	81.62	89.75	86.74	90.36	90.60
$S=5$	67.77	82.22	87.04	93.07	91.86	94.57	93.97
$S=6$	84.63	84.63	86.74	92.77	92.77	92.16	92.77
$S=7$	80.12	88.89	**89.45**	92.77	91.86	94.45	**94.27**

Table 2: Two class classification results for the FSC (English) Dataset. Intents - 'activate kitchen lights', 'deactivate bedroom lights'.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
$S=1$	45.83	58.73	47.42	61.11	57.53	54.56	64.88
$S=2$	42.65	55.55	56.15	66.86	73.81	72.42	76.67
$S=3$	52.57	74.01	**73.21**	67.26	78.76	80.35	**85.11**
$S=4$	53.57	75.19	73.81	78.37	76.98	83.13	84.72
$S=5$	59.72	68.25	75.79	80.55	83.53	80.75	83.73
$S=6$	69.45	74.61	78.37	81.54	85.31	82.73	86.31
$S=7$	73.21	72.22	**80.75**	85.51	86.90	84.12	**88.49**

Table 3: Four class classification results for the FSC (English) Dataset. Intents - 'activate kitchen lights', 'deactivate bedroom lights', 'increase washroom heat' and 'decrease volume'.

We pose two I-class intent classification problems, where $I = 2, 4$. The results for English are shown in Table 2, 3 and for Flemish are shown in 4, 5. The columns of each show results for different values of k, where k is the number of utterances recorded by a speaker per intent. This means that if $k = 3$, each speaker provided 3 recordings for each intent, which amounts to a total of $3 * I$ recordings per speaker. In general, each speaker records $k * I$ audios, where k is the number of audios recorded by a speaker per intent, and I is the number of intents. The rows represent the number of speakers (S) involved in creating the dataset.

We want to point the reader to four locations in Tables 2, 3, 4 and 5, which describe four different scenarios of data collection. ($S = 3, k = 3$) refers to a scenario where we only use 3 speakers to create a training set and each speaker records 3 audio samples per intent, which means we need 6 or 12 audio samples per speaker depending on the classification problem. In this scenario, we have just 9 training

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
$S=1$	86.11	86.11	90.74	97.22	83.33	85.18	88.88
$S=2$	91.66	96.29	95.37	96.29	96.29	98.14	94.44
$S=3$	89.81	96.29	**98.14**	96.29	92.59	1.0	**98.14**
$S=4$	97.22	92.59	94.44	98.14	97.22	98.14	97.22
$S=5$	92.59	95.37	94.44	92.59	98.14	97.22	99.07
$S=6$	91.66	96.29	96.29	97.22	99.07	97.22	97.22
$S=7$	93.51	95.37	**98.14**	98.14	97.22	98.14	**1.0**

Table 4: Two class classification results for the GRABO (Flemish) Dataset. Intents - ’approach’, ’lift’.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
$S=1$	56.48	54.16	56.01	55.09	58.33	57.87	57.40
$S=2$	67.12	71.29	75.92	75.92	70.83	72.22	82.40
$S=3$	74.07	72.22	**73.14**	73.14	74.53	77.31	**76.38**
$S=4$	69.90	72.22	72.22	73.14	79.16	73.61	80.09
$S=5$	65.74	75.00	78.70	79.16	78.24	82.87	80.09
$S=6$	73.61	80.09	85.64	84.25	87.03	86.11	89.35
$S=7$	81.48	81.94	**86.57**	86.11	89.81	89.35	**90.27**

Table 5: Four class classification results for the GRABO (Flemish) Dataset. Intents - ’approach’, ’lift’, ’pointer’, ’grab’.

utterances per intent. We see that for both Flemish and English, even ($S=3, k=3$) systems can be deployed in the real world. By systems that can be deployed, we mean that the performance is such that the above SLU system can be incorporated in a spoken dialog system in the real world with a rule based dialog manager, confirming the recognized intent. We also see that a ($S=3, k=3$) system is able to generalize well to a test set with no speaker overlap and hundreds of utterances, which are way less than the utterances used to train the system.

($S=3, k=7$) refers to a more involved recording process where we still only have access to 3 speakers but each record 7 audio samples per intent, which makes it 14 or 28 audio samples recorded per speaker. In this setting, we have 21 training utterances per intent. A comparable case is ($S=7, k=3$), where we again have 21 training utterances per intent, but with a larger number of speakers with reduced load, each having to record only 6 or 12 training utterances in total. We see that for both Flemish and English, increasing the total number of utterances per intent increases performance. Also, keeping the number of utterances per intent constant, increasing the number of speakers provides better results as speaker variability adds to the generalization capability of the model.

Finally, ($S=7, k=7$) is the most exhaustive recording procedure presented in these experiments with 7 speakers, where each speaker still records 7 audio samples each per intent or 14-28 audio samples depending on the classification problem. We see that increasing the number of audio samples per speaker (k) in general increases performance. Increasing the number of speakers provides more variability in the training set and allows the model to generalize better. In a real world data collection setting, getting more individual speakers and getting each speaker to record multiple recordings are both important variables that determine the success of the data collection procedure.

6 Conclusion

In this paper, we provide a series of experiments that empirically recreate a real world setting of building spoken dialog systems for extremely low resourced scenarios - where we don’t even have access to speech to text conversion technologies. To overcome this problem, we use Allosaurus to convert speech into its phonetic transcription which we vectorize and use as inputs to our model. In such a setting, collecting annotated data can be difficult, thus making every collected data point crucial. To see if we can build SLU systems in such settings, we present intent classification results at a granularity where we see the effects changing the number of speakers and utterances recorded by each speaker. We see encouraging results and find that even with as low as 7 speakers recording 7 utterances per intent, we can create real world SLU systems. Through this paper, we want to push the exploration in building spoken dialog systems in extremely low resourced settings and our work is a step in that direction. Note that we haven’t used any data augmentation methods yet which would further boost the performance of our systems.

Allosaurus is a nearly universal phone recognition system creating language and speaker independent representations, thus we can expect similar performance on other languages for tasks of similar complexities, as has been observed for languages like Sinhala and Tamil in previous works (Gupta et al., 2021; Yadav et al., 2021). A possible consideration for the system to work is the task complexity. Usual Intent recognition tasks like the ones in the FSC and Grabo dataset have relatively shorter token sequences, where each utterance is between 2-5 words long. More complex tasks with longer and more confusing utterances might require more data for disambiguation of intents.

References

Darshana Buddhika, Ranula Liyadipita, Sudeepa Nadeeshan, Hasini Witharana, Sanath Javasena, and Uthayasanker Thayasivam. 2018. Domain specific intent classification of sinhala speech data. In *2018 International Conference on Asian Language Processing (IALP)*, pages 197–202. IEEE.

Yuan-Ping Chen, Ryan Price, and Srinivas Bangalore. 2018. Spoken language understanding without speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6189–6193. IEEE.

Laurence Devillers, Hélène Maynard, Sophie Rosset, Patrick Paroubek, Kevin McTait, Djamel Mostefa, Khalid Choukri, Laurent Charnay, Caroline Bousquet, Nadine Vigouroux, et al. 2004. The french media/evalda project: the evaluation of the understanding capability of spoken language dialogue systems. In *LREC*. Citeseer.

Allen L Gorin, Giuseppe Riccardi, and Jeremy H Wright. 1997. How may i help you? *Speech communication*, 23(1-2):113–127.

Zhiyuan Guo, Yuexin Li, Guo Chen, Xingyu Chen, and Akshat Gupta. 2021. Word-free spoken language understanding for mandarin-chinese. *arXiv preprint arXiv:2107.00186*.

Akshat Gupta, Olivia Deng, Akruti Kushwaha, Saloni Mittal, William Zeng, Sai Krishna Rallabandi, and Alan W Black. 2021. Intent recognition and unsupervised slot identification for low resourced spoken dialog systems. *arXiv preprint arXiv:2104.01287*.

Akshat Gupta, Xinjian Li, Sai Krishna Rallabandi, and Alan W Black. 2020a. Acoustics based intent recognition using discovered phonetic units for low resource languages. *arXiv preprint arXiv:2011.03646*.

Akshat Gupta, Sai Krishna Rallabandi, and Alan W Black. 2020b. Mere account mein kitna balance hai?–on building voice enabled banking services for multilingual communities. *arXiv preprint arXiv:2010.16411*.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR.

Yohan Karunanayake, Uthayasanker Thayasivam, and Surangika Ranathunga. 2019a. Sinhala and tamil speech intent identification from english phoneme based asr. In *2019 International Conference on Asian Language Processing (IALP)*, pages 234–239. IEEE.

Yohan Karunanayake, Uthayasanker Thayasivam, and Surangika Ranathunga. 2019b. Transfer learning based free-form speech command classification for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 288–294.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, et al. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. *arXiv preprint arXiv:1904.03670*.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. 2014. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539.

Bart Ons, Jort F Gemmeke, et al. 2014. The self-taught vocal interface. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):43.

Yao Qian, Rutuja Ubale, Vikram Ramanaryanan, Patrick Lange, David Suendermann-Oeft, Keelan Evanini, and Eugene Tsuprun. 2017. Exploring asr-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 569–576. IEEE.

Vincent Renkens, Steven Janssens, Bart Ons, Jort F Gemmeke, et al. 2014. Acquisition of ordinal words using weakly supervised nmf. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 30–35. IEEE.

Alaa Saade, Alice Coucke, Alexandre Caulier, Joseph Dureau, Adrien Ball, Théodore Bluche, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, et al. 2018. Spoken language understanding on the edge. *arXiv preprint arXiv:1810.12735*.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. 2018. Towards end-to-end spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5754–5758. IEEE.

Netsanet Merawi Tessema, Bart Ons, Janneke van de Loo, Jort Gemmeke, Guy De Pauw, Walter Daelemans, et al. 2013. Metadata for corpora patcor and domotica-2. *Technical report KUL/ESAT/PSI/1303, KU Leuven, ESAT, Leuven, Belgium*.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Mike Wu, Jonathan Nafziger, Anthony Scodary, and Andrew Maas. 2020. Harpervalleybank: A domain-specific spoken dialog corpus. *arXiv preprint arXiv:2010.13929*.

Hemant Yadav, Akshat Gupta, Sai Krishna Rallabandi, Alan W Black, and Rajiv Ratn Shah. 2021. Intent classification using pre-trained embeddings for low resource languages. *arXiv preprint arXiv:2110.09264*.

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*, pages 521–525.

A Implementation Details

All models are trained using the NVIDIA GeForce GTX 1070 GPU using python3.7. The training is very quick due to the small dataset sizes, with each epoch taking 1-2 seconds. For each experiment, a validation set identical to the test set was used. For the FSC dataset, the validation set had 10 speakers with no speaker overlap with the training or the test set. Similarly for the GRABO dataset, the validation set had 2 speakers that were not present in the training or the test set. Each experiment in Tables 2-5 was repeated 3 times and the maximum accuracy has been reported.

As mentioned in section 4, we use a CNN+LSTM architecture, as proposed in (Gupta et al., 2020a). We performed a grid search over various parameters of the architecture. The best performing models varied slightly for each experiment. The exact model parameters for the results reported in Tables 2-5 are shown in Table 6. For larger amounts of utterances recorded per speaker, we found better results with 2 LSTM layers instead of one.

Model Parameters	**Value**
Embedding Size	256
CNN kernel size	3
No. of CNN filters	256
No. of LSTM layers	1 (or 2)
LSTM hidden size	256
Batch Normalization	False

Table 6: Model Parameters