Comparing Data Augmentation and Annotation Standardization to Improve End-to-end Spoken Language Understanding Models

Leah Nicolich-Henkin
nicolich@amazon.com

Taichi Nakatani
taichina@amazon.com

Zach Trozenski
trozenz@amazon.com

Joel Whiteman
jtwhite@amazon.com

Nathan Susanj
nsusanj@amazon.com

Abstract

All-neural end-to-end (E2E) Spoken Language Understanding (SLU) models can improve performance over traditional compositional SLU models, but have the challenge of requiring high-quality training data with both audio and annotations. In particular they struggle with performance on "golden utterances", which are essential for defining and supporting features, but may lack sufficient training data. In this paper, we compare two data-centric AI methods for improving performance on golden utterances: improving the annotation quality of existing training utterances and augmenting the training data with varying amounts of synthetic data. Our experimental results show improvements with both methods, and in particular that augmenting with synthetic data is effective in addressing errors caused by both inconsistent training data annotations as well as lack of training data. This method leads to improvement in intent recognition error rate (IRER) on our golden utterance test set by 93% relative to the baseline without seeing a negative impact on other test metrics.

1 Introduction

Spoken Language Understanding (SLU) models transcribe, classify, and label human utterances directly from audio input, forming an essential component of digital voice assistants (VAs) such as Amazon Alexa, Google Assistant, and Apple Siri. All-neural end-to-end (E2E) SLU models take in audio and output natural language understanding (NLU) classifications [12][2]. E2E SLU models have a number of advantages over a traditional compositional architecture where the model is composed of separate Automatic Speech Recognition (ASR) and NLU models. These advantages include lower latency, a smaller model size, and higher accuracy, as the model is optimized for NLU and can recover from some ASR errors[13]. However, due to their end-to-end nature, they require their training data to have audio, transcriptions, and annotations, and as a result often struggle to find a sufficient amount of high-quality training data. To train a text-only NLU model we might simply generate additional text inputs, but an E2E model requires that generated data to contain both audio and annotated text.

In particular, we've seen E2E SLU models struggle with high error rates on "golden utterances". Golden utterances are manually selected utterances that represent a cross-section of features and dictate the design of the product and customer experience. They can also be seen as a stand-in for any small set of high stakes utterances.

The need for large amounts of transcribed and annotated audio data is a known challenge of E2E systems. While the use of data synthesis and augmentation to improve ML systems more generally

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

is widespread in both ASR [8][17][9][14] and NLU [16][15][7][4][3][1], there is limited work on audio data augmentation in the context of an end-to-end SLU system. Lugosch et al. [10] find that text-to-speech (TTS) audio can be effectively used to improve an E2E SLU model given a lack of speakers, while Huang et al. [6] experiment with using TTS to create audio for existing text-only training data. However these studies focus on the improving model as a whole, and do not address using synthesized data to target an improvement to specific intents or utterances. They also do not compare to other data-centric AI approaches for improving E2E SLU models.

In this paper we compare two methods for improving E2E model performance on golden utterances: improving data quality and increasing data quantity. To improve data quality we apply a set of rules to standardize golden utterance annotations in our existing training data to a canonical form, while to increase data quantity we include synthesized audio versions of those golden utterances in the E2E model's training data.

2 Data augmentation

Golden utterance representation in our training dataset is limited. Our baseline model is trained on a dataset composed of de-identified utterances used in a live setting to communicate with a voice assistant. Out of 802 golden utterances, only 263 (33%) appear in that baseline training data. Additionally a large portion of goldens that are represented in the training data have very few examples. To overcome the lack of training data, we augmented our training set with synthetically generated text-to-speech (TTS) data, matched with our known ground truth annotations. We hypothesized that adding too much synthesized data might result in a loss in general performance of the model through overfitting to these specific utterances. To measure the relation between augmentation size and recognition performance, we trained the model with several tiers of augmentation. We created synthetic data for each of our golden utterances with 3 sizes: 10, 100 and 1000 samples for each of the 802 golden utterances, each representing male and female voices from various age groups. Each individual synthesized utterance represents a unique TTS voice profile, using up to 1000 unique voice profiles. Our model is targeted at English speakers in the US, and the TTS data likewise reflects that demographic. We combined the resulting augmented data with our existing training data to create the final dataset used for training.

3 Standardizing annotations

While analyzing our training data, we found that a large portion of annotations matching golden utterances were inconsistently annotated, meaning that the transcription matches the golden utterance but the intent and slot labels applied to the transcription vary. Studies have found that incorrect training data can cause decreased model performance [11] [5]. Of the 263 golden utterances that appear in the training data, 215 (81.8%) have some incorrect annotations, and, for 117 (44.5%) of them, more than half of the available training data does not match the golden ground truth. With the hypothesis that consistent annotation improves the ability for the model to learn the goldens, we created a new version of our training dataset where all utterances with transcriptions matching our golden utterances are standardized to the canonical golden annotation. We compared the model trained with standardized annotations to the baseline and also experimented with combining standardized annotations with the augmented data described in Section 2.

4 Model training and evaluation

Our baseline model is trained on a live dataset composed of de-identified human utterances used to communicate with a voice assistant. The entire dataset consists of 494 intents, comprising approximately 13k hours of audio. However, our model use-case targets a subset of 59 intents, with the remainder used for training but considered unsupported and only used for False Accept Rate (FAR) evaluation. The data used for our experiments is described in Table 1.

The SLU model used for our experimentation is a multi-task model that takes an audio signal as input and outputs a transcription, an intent label, and slot labels for each word. The ASR layers of the model take the form of a Recurrent Neural Network-Transducer (RNN-T) model, consisting of an audio encoder network and a wordpiece prediction network, combined in a joint network that

Table 1: Data used in each experimental model

Golden training annotations	Synthetic utterances per golden	Total synthetic utterances
Original	0	0
Original	10	8,020
Original	100	80,200
Original	1000	802,000
Standardized	0	0
Standardized	10	8,020
Standardized	100	80,200
Standardized	1000	802,000

outputs a sequence of predicted wordpieces. The interface layers take embeddings from the predicted wordpieces and the hidden layers from the RNN-T and pass them to the NLU network. Finally the NLU network consists of an intent tagger based on feed-forward networks and a bidirectional long short-term memory (BiLSTM) slot tagger. Although they are distinct, the ASR and NLU networks are jointly trained, which allows the NLU interpretation to make use of, and benefit from, both the predicted transcription and the hidden layers that are output by the ASR subsystem. This architecture improves overall performance by providing the NLU subsystem with more information than simply the ASR 1-best prediction, as would be the case in a traditional compositional architecture. Each of our models starts with a generic ASR RNN-T model that is pre-trained using over 100k hours of data. For training purposes, we use the model and datasets described above to fine-tune the ASR layers of the E2E SLU model alone, then train the NLU layers of the model with ASR layers frozen, followed by jointly training the entire model on all of the data.

Two test sets were used to evaluate our models. A synthetic dataset was used to evaluate golden utterance performance, consisting of one example of each of the 802 golden utterances. Due once again to lack of live data, it was synthesized using a TTS voice from Amazon AWS Polly, a separate TTS system from the one used to generate augmented training data. In addition, a live de-identified dataset was used to evaluate model performance on all supported intents. This was used to evaluate for any potential regression due to overfitting on synthetic golden utterances. This set consists of approximately 100k utterances (72 hours). For both test sets we evaluated ASR performance on word error rate (WER), and NLU performance on intent recognition error rate (IRER), which measures the percentage of utterances that contain any errors, including intent classification, slot value transcriptions, and slot labeling. In addition, for the live set we looked at false accept rate (FAR), which measures the percent of unsupported utterances incorrectly recognized as supported.

5 Results and analysis

Adding synthetic training data led to large performance improvements in both WER and IRER on the golden test set (shown in Table 2). The live set also showed improvements on IRER and FAR, and modest degradation on WER, even with our highest level of data augmentation. Although the change in WER looks concerning, the downstream impact on IRER is negligible, so we consider the tradeoff acceptable. Each level of data augmentation led to an improvement over the previous one, culminating in the model with the largest augmentation amount having the lowest WER / IRER metrics across all experiments. Compared to the baseline, the best performing model, with original annotations and 1000 synthetic utterances per golden, reduced WER by 86.1% relative and IRER by 93.6% relative on the golden test set.

Standardizing existing golden training annotations led to a more moderate improvement on the golden test set of 15.1% IRER relative to the baseline, while also resulting in a degradation of 4.6% IRER relative to the live set. The improvements that we did see were primarily restricted to those goldens that that were well-represented but had errors in the original training data. We believe that more widespread inconsistencies in data annotation contributed to the degradation on the live set, demonstrating that a simple targeted fix cannot make up for other deficiencies.

Table 2: Evaluation relative to baseline with original annotations and no synthetic data

Golden training annotations	Synthetic data per golden	Relative Golden WER (%)	Relative Golden IRER (%)	Relative Live WER (%)	Relative Live IRER (%)	Relative Live FAR (%)
Original	10	-31.9%	-45.3%	+14.9%	+4.4%	-6.7%
Original	100	-72.9%	-85.9%	+14.3%	-7.8%	-9.3%
Original	1000	-86.1%	-93.6%	+14.5%	-7.8%	-9.3%
Standardized	0	+1.7%	-15.1%	+14.4%	+4.6%	-6.7%
Standardized	10	-32.3%	-51.7%	+14.7%	-5.7%	-5.3%
Standardized	100	-69.8%	-88.6%	+14.9%	-5.7%	-8.0%
Standardized	1000	-82.7%	-93.4%	+15.4%	-5.7%	-6.7%

The data suggest the largest augmented set (1000 added utterances per golden) to be the saturation point for IRER and WER improvement. We observe accuracy begin to plateau between experiments using the 100x and 1000x augmented sets, suggesting additional magnitudes of augmentation (10,000x, 100,000x, etc.) may only facilitate incremental reductions in WER and IRER, while introducing unnecessary bias towards the golden utterances in our models.

At low levels of data augmentation, annotation standardization and data augmentation complement each other, with augmentation addressing errors on utterances with little training data while annotation clean-up addresses errors on utterances with large amounts of incorrect training annotations. However, while this approach shows overall improvement on un-augmented data, its impact in reducing IRER is smaller than the impact of training with even the smallest augmented dataset (10 utterances per golden). With large amounts of augmented data, all extra benefit from the annotation standardization is eliminated, demonstrating that augmentation alone is capable of addressing those same errors.

Despite targeting NLU improvements, we also saw a positive impact on ASR errors for golden utterances, with WER decreasing with each level of data augmentation. Models trained on the largest amount of synthetic data were able to reduce the number of golden utterances with ASR errors by 86.1% relative, a significant impact that we see reflected in our downstream IRER metrics.

6 Conclusion

Our results show that both data-centric approaches to improving E2E SLU achieved the desired effect, although data augmentation was much more powerful than annotation standardization. Cleaning the annotated data with one standardized annotation for each golden utterance was limited by the amount of live training data available, therefore it may not be a surprise that adding synthetic data had a positive impact on a broader range of utterances. However, it was surprising that data augmentation in fact superceded data cleaning, and thus combining both techniques turned out to be unnecessary. With even 100 synthetic utterances per golden, the benefit of standardizing the existing training data's annotations disappears. This suggests that, at least for scarce data cases such as golden utterances, it may be preferable to simply overwhelm any annotation inconsistency with new examples that adhere to the proper annotation conventions.

Results from our evaluation show that greater amounts of synthetic data significantly improve intent recognition for these target utterances without degrading overall model performance. Even adding only 10 synthetic examples per golden has a big impact on utterances that are absent from the live training data, and each of our discrete synthetic data increments improves performance further. Results from our live data evaluations show that IRER and FAR remain stable even with the large amounts of augmented data added to the training data. While there is an increase in WER across experiments, this co-occurs with stable IRER metrics, which suggests that these speech recognition errors are not negatively affecting the correct recognition of intents and slot labels. Going forward we plan to expand our data augmentation to support and improve a wider range of use cases.

References

[1] Olga Golovneva and Charith Peris. Generative adversarial networks for annotated data augmentation in data sparse nlu. In *17th International Conference on Natural Language Processing*,

2020.

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

[3] Batool Arhamna Haider, He He, Ajay Mishra, Garima Lalwani, and Mona Diab. Data augmentation for low-resource natural language understanding in dialogue systems. In *Amazon Machine Learning Conference 2019*, Seattle, WA, 2019.

[4] Christopher Hench, Cedric Warny, and Sreekar Bhaviripudi. Data augmentation for intent classification and slot filling of custom semantics. In *The 3rd Shareable NLP Technologies Across Amazon Workshop*, Seattle, WA, 2020.

[5] Sara Hooker, Aaron Courville, Yann Dauphin, and Andrea Frome. Selective Brain Damage: Measuring the Disparate Impact of Model Pruning. *arXiv e-prints*, Nov 2019.

[6] Yinghui Huang, Hong-Kwang Kuo, Samuel Thomas, Zvi Kons, Kartik Audhkhasi, Brian Kingsbury, Ron Hoory, and Michael Picheny. Leveraging unpaired text data for training end-to-end speech-to-intent systems. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7984–7988, 2020.

[7] Robin Jia and Percy Liang. Data recombination for neural semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Berlin, Germany, August 2016. Association for Computational Linguistics.

[8] Aleksandr Laptev, Roman Korostik, Aleksey Svischev, Andrei Andrusenko, Ivan Medennikov, and Sergey Rybin. You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 439–444, 2020.

[9] Duc Le, Thilo Koehler, Christian Fuegen, and Michael L. Seltzer. G2g: Tts-driven pronunciation learning for graphemic hybrid asr. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6869–6873, 2020.

[10] Loren Lugosch, Brett H. Meyer, Derek Nowrouzezahrai, and Mirco Ravanelli. Using speech synthesis to train end-to-end spoken language understanding models. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8499–8503, 2020.

[11] Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:11373–1411, 2021.

[12] Milind Rao, Anirudh Raju, Pranav Dheram, Bach Bui, and Ariya Rastrow. Speech to Semantics: Improve ASR and NLU Jointly via All-Neural Interfaces. In *Proc. Interspeech 2020*, pages 876–880, 2020.

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

[14] Yash Sharma, Basil Abraham, Karan Taneja, and Preethi Jyothi. Improving Low Resource Code-Switched ASR Using Augmented Code-Switched TTS. In *Proc. Interspeech 2020*, pages 4771–4775, 2020.

[15] Alex Sokolov and Denis Filimonov. Neural machine translation for paraphrase generation. In *2nd Conversational AI workshop*, 2018.

[16] Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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)*, pages 6382–6388, Hong Kong, China, November 2019. Association for Computational Linguistics.

[17] Xianrui Zheng, Yulan Liu, Deniz Gunceler, and Daniel Willett. Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end asr systems. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5674–5678, 2021.