The impact of intent distribution mismatch on semi-supervised spoken language understanding

Judith Gaspers, Quynh Do, Daniil Sorokin, Patrick Lehnen

Amazon Alexa AI, Aachen, Germany

{gaspers, doquynh, dsorokin, plehnen}@amazon.de

Abstract

With the expanding role of voice-controlled devices, bootstrapping spoken language understanding models from little labeled data becomes essential. Semi-supervised learning is a common technique to improve model performance when labeled data is scarce. In a real-world production system, the labeled data and the online test data often may come from different distributions. In this work, we use semi-supervised learning based on pseudo-labeling with an auxiliary task on incoming unlabeled noisy data, which is closer to the test distribution. We demonstrate empirically that our approach can mitigate negative effects arising from training with non-representative labeled data as well as the negative impacts of noises in the data, which are introduced by pseudo-labeling and automatic speech recognition.

1. Introduction

Spoken Language Understanding (SLU) is a key task in voice-controlled devices, such as Amazon Alexa and Google Assistant. Typically, two sub-tasks are considered: intent classification (IC) and slot filling (SF). The former identifies a user's intent and the latter assigns a label to each token in the utterance, which is either a slot label or "O", where the latter indicates that the token does not carry relevant semantic meaning for this task. For example, given the user utterance "play music by volbeat", IC should classify *PlayMusic* as intent, while SF should detect "volbeat" as *Artist* and "play" as *O*. In real-world SLU applications, a SLU model is typically applied together with an automatic speech recognizer (ASR) in a pipelined fashion, i.e. a user request is first transcribed via ASR and subsequently fed into the SLU model to obtain semantic information. Since ASR introduces noise in the form of ASR errors, robustness to such errors is important for real-world SLU applications.

Due to the great success of voice-controlled devices, bootstrapping SLU models for new locales and domains in a cost-efficient manner, i.e. using few labeled data, has become an important goal [1, 2]. After initial SLU models trained on comparatively low data amounts, have been deployed to customers, unlabeled data can be collected from the device and recognized via ASR. In a real-world scenario, distributional mismatches may arise among the labeled data, unlabeled data and data used by customers during application. In particular, the labeled dataset may have a skewed intent distribution compared to the customer data during application, while the unlabeled dataset might have a closer distribution. This can be the case, because initial collections of labeled data need to happen before a first model can be deployed to customers, and hence labeled data needs to be collected without having much knowledge about the future customer data distributions, such as which intents will be used frequently. In addition, data distributions in real-world applications are continuously changing over time. Seasonal changes around events, such as Christmas, can cause a high peak of related utterances, like listening to Christmas songs or turning on Christmas lighting, which will quickly disappear again after the event has passed. Therefore, one may expect that the intent distribution used by customers during application is closer to recent unlabeled ASR data than the labeled set. This makes the unlabeled dataset a potential resource to bootstrap SLU performance via a semi-supervised learning (SSL) approach. To the best of our knowledge, the impact of intent distributions on the effectiveness of SSL for SLU has not yet been studied in the literature.

In this paper, we focus on studying the impact of intent distributions on semi-supervised SLU. For SSL, we start from a common teacher-student pseudo-labeling approach, which has been applied in many fields [3, 4]. The main idea is that a teacher model is trained on the labeled data, and in turn used to generate pseudo-labels for the unlabeled data, which are leveraged together with the labeled data for training a student model. Since we study SSL with small amounts of labeled data and noisy ASR traffic, model performance may suffer from low-quality pseudo-labeled data. That is, i) the initial teacher model may not be strong and may generate a large amount of incorrect pseudo-labels, and ii) a comparatively large amount of unlabeled data including noise in the form of ASR errors may decrease the quality of the pseudo-labeled data further. This raises the issue of how to reduce the influence of noises on the SSL performance. Motivated by [5], we use auxiliary learning to feed ASR noisy data indirectly to a DNN-based SLU model, helping our model to avoid the direct impact of different types of noises. In addition, inspired by [4] we perform gradual noising, aiming to make the SLU model more robust to ASR errors.

Our contributions in this paper are that i) we experimentally study the impact of intent distributions on auxiliary learning for semi-supervised SLU when few labeled data are available, and ii) we compare different strategies for leveraging noisy pseudo-labeled data in auxiliary learning for semi-supervised SLU.

2. Related Work

Most modern approaches for SLU use DNN architectures, which model SF and IC jointly [6, 7, 8]. Different SSL approaches for SLU have been studied, including traditional methods which leverage pseudo-labels [9], and methods aiming to improve pseudo-label-based approaches further, e.g. by dual representation learning [10]. However, we cannot compare performance of our approaches directly, as some work was evaluated on internal data [9], and for publicly available datasets the SSL simulations differ. In particular, unlike in our work, in recent work [10] labeled and unlabeled splits were created from the training data, yielding an unrealistically large development set up for small labeled data amounts [11]. Note that our main goal is to shed light on the impact of intent distributions on SSL and investigating the impact of different strategies for leveraging pseudo-labeled

data in teacher-student SSL rather than proposing a new state-of-the-art SSL approach for SLU.

Teacher-student SSL is currently a common approach which has been applied to many different tasks, such as image processing [3] or automatic speech recognition [4]. It has been pointed out that the injection of noise is important w.r.t. performance, and noise may be injected via data augmentation or via model noise [3, 12]. In this work, we study the impact of different strategies for leveraging pseudo-labeled data in teacher-student SSL for SLU. In particular, to mitigate negative impacts from noisy pseudo-labeled data we study indirect injection via an auxiliary task – a method which has been previously shown to mitigate negative effects of using synthetic data in supervised SLU [5]. Notably, in [5] a class-balanced batch generator was applied, as the focus was on improving performance of low-frequency features. Since this approach would remove information about the intent distribution in the unlabeled data, which we aim to leverage, we do not apply a class-balanced batch generator.

Recent work also investigated SSL in the context of end-to-end SLU, i.e. when working directly with speech input for SLU rather than ASR transcriptions [13, 14, 15]. However, while with this approach the propagation of (non-recoverable) ASR errors to the SLU model can be mitigated, pipelined approaches are still the prominent approach in real-world production SLU applications, which are targeted by our work.

Robustness to ASR errors has been explored by leveraging several ASR hypotheses instead of just the 1-best. For instance, using word confusion networks has been proposed [16].

3. Method

In this paper, we explore teacher-student SSL based on pseudo-labeling, in which the student may apply auxiliary training. Inspired by [5], this student model uses a multi-task framework consisting of a primary and an auxiliary SLU task, which share a feature extraction component. While both tasks are trained by alternating across tasks, only the primary task is used for inference. In this work, we train the primary task on labeled data, while the auxiliary task is trained on noisy pseudo-labeled data. Thus, noisy pseudo-labeled data is only indirectly injected into the primary task via the feature extraction component, which helps in our experiments with mitigating the negative impacts of noisy pseudo-labeled data.

3.1. Semi-supervised learning task

We assume a labeled dataset $L = \{(x_i, y_i)\}_{i=1}^{L}$ where $x_i = x_{i_1}, .., x_{i_n}$ is an observed utterance of n tokens. For y_i we assume a sentence-level intent label for IC, and for SF a slot label for each token in x_i. In addition, we assume an unlabeled dataset $U = \{x_j\}_{j=1}^{U}$ where x_j may comprise ASR errors. The distributions of L and U may differ, i.e. $P_L(x) \neq P_U(x)$. Then, we are concerned with building models using both L and U, such that performance is improved on a test set $T = \{(x_i, y_i)\}_{i=1}^{T}$ with potentially $P_T(x, y) \neq P_L(x, y)$ and $P_T(x) \neq P_U(x)$, and T may comprise ASR errors.

3.2. Teacher-student SSL with pseudo-labeling

A common approach to SSL starts by training a teacher model on L which is then used to generate pseudo labels y' for utterances in U yielding $U' = \{(x_j, y'_j)\}_{j=1}^{U}$. U' is then combined with L, and a student model is trained on the augmented data. This process may be repeated N times with setting the student as the next teacher model. Since pseudo-labels are noisy and because

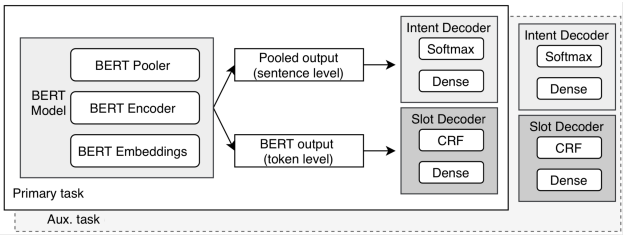

Figure 1: *SLU model architecture (adapted from [5])*

ASR traffic comprises noise in the form of ASR errors, we filter U' before augmenting it into L based on ASR and IC model confidence scores, yielding $U'_{filtered}$. Thresholds for filtering are gradually relaxed with the intuition that student models are becoming stronger over runs and can leverage more noisy data successfully. This intuition is supported by previous work applying gradual relaxing of thresholds in teacher–student SSL for automatic speech recognition [4]. Besides gradual noising with ASR traffic, we inject model noise, e.g. via dropout, as noising has been shown to be important in order to achieve strong performance in teacher-student SSL [3, 12].

3.3. SLU model training

For the teacher model, we start from a common SLU architecture based on BERT, as models based on pre-trained BERT are known to give strong performance in particular in few shot/low data settings. In particular, the model consists of a BERT encoder, an intent decoder and a slot decoder. The BERT encoder's outputs at sentence and token level are used as inputs for the intent and slot decoders, respectively. The intent decoder is a standard feed-forward network including two standard dense layers and a softmax layer on top. Meanwhile, the slot decoder uses a CRF layer on top of two dense layers to leverage the sequential information of slot labels. During the training, the losses of IC (cross-entropy loss) and SF (CRF loss) are optimized jointly with equal weights (1.0:1.0).

For the student, we consider two model backbones:

1. **"Standard" SLU model:** The teacher model described previously is applied, which consists of a single SLU task and is trained by directly injecting the augmented data $L_{aug.} = L + U'_{filtered}$.
2. **Auxiliary SLU model:** A model comprising an auxiliary SLU task is applied. Following [5], our model, as shown in Fig. 1, consists of two SLU tasks, i.e. a main task and an auxiliary task, both having the same model topology and sharing a common feature extraction. The two tasks are optimized jointly during training, but only the main task is kept for the inference phase. While the main task is trained on L, the auxiliary task is trained on $U'_{filtered}$. The overall loss is a weighted sum of four losses, i.e. CRF loss (main SF), cross-entropy loss (main IC), CRF loss (auxiliary SF), and cross-entropy loss (auxiliary IC). Intuitively, the negative effects from using noisy data (i.e. coming both from ASR errors and pseudo-labeling) can be mitigated by the indirect injection of $U'_{filtered}$. This is supported by previous work [5] which has shown that using an auxiliary task can mitigate negative effects of using low-quality synthetic data in supervised SLU [5]. In addition, the model keeps access to each distribution individually rather than having access to the mix up only, and the contribution of labeled vs unlabeled utterances to the overall loss is more balanced during training.

Domain	Labeled	Unlabeled	Test
Music	3,937	74,803	35,210
Video	1,094	9,849	3,666

Table 1: *Number of utterances per internal dataset.*

In our experiments, we use hard labels for SSL, and we include weighting of losses by the teacher's normalized model confidence scores. In sum, we compare the following approaches for leveraging L and $U'_{filtered}$ in teacher-student SSL for SLU:

- **Direct inject:** The "standard" SLU model is applied as student and trained on $L_{aug.}$.
- **Direct inject + weights:** Same as above, but the CRF and cross-entropy loss are weighted with normalized SF and IC confidence scores, respectively. Weights for utterances from L are set to 1.0.
- **Indirect inject:** $U'_{filtered}$ is injected indirectly into a student via an auxiliary task, while the student's main task is trained with L.
- **Indirect inject + weights:** Same as above, but the CRF and cross-entropy loss for the auxiliary task are weighted with normalized SF and IC confidence scores, respectively.

4. Experiments

4.1. Datasets

We conduct experiments using i) the small-scale benchmark dataset ATIS [17] which comprises annotated English data of people making flight reservations, and ii) real-world data samples extracted from a commercial German SLU system; for the latter the data are representative of user requests to voice-controlled devices, and the data were suitably de-identifed and manually annotated with intent and slot labels.

We construct an SSL set up for ATIS by first collapsing the training and development sets and subsequently splitting the data into 10% and 90% to form L and U, respectively. For U, we drop the labels. We collapse training and development datasets first rather than dividing the training data and keeping the development set, as the assumption of having a comparatively large development set together with a small L has been pointed out to be unrealistic [11].

We apply two strategies to select L.

- **Random:** Data are selected randomly. In this case the data distributions should be similar across L, U and T.
- **Skewed:** Data are selected randomly, except that for three of the intents, only 5% of samples are put into L; the three intents are *flight*, *ground_service*, and *airline*. Notably, *flight* is by far the most frequent intent in ATIS. With this selection, there is an artificial mismatch in the intent class distributions across datasets, and the intent class distribution of U is closer to that of T than that in L in the sense that the considered intents are much more over-represented in U and T than in L.

Note that the sizes of L and U are the same in the skewed and the random scenario; just the intent distributions differ. For T we use the standard ATIS test set in all cases.

We include real-world samples from two domains, i.e. *Music* and *Video*, into our experiments, and we construct L and U to follow similar and skewed distributions as described previously. The percentage of labeled data is 10% for *Video* and 5% for *Music*. Unlike for ATIS, data in U are ASR transcriptions which are more noisy due to ASR errors. For testing, we evaluate both on manual transcriptions and on ASR 1-best hypotheses, where the latter represents the application set up in production, while the former is used in offline evaluation. Following [18], to evaluate SF on ASR hypotheses we project slot labels from the manual transcriptions to the corresponding ASR hypotheses if all slot values persisted and otherwise drop an utterances from the ASR test set. By computing the word error rates based on the manual transcriptions and ASR 1-best hypotheses from the test sets, we note that the word error rate is roughly 38% relative higher for *Music* than for *Video* in our test data (note that this number concerns our small data samples only and is not representative of the overall system). Thus, the *Music* domain set up is more challenging in that the percentage of unlabeled data is much higher and these data are moreover noisier. The data amounts are detailed in Table 1. Note that while the internal data amounts are larger than the small-scale ATIS data, the sizes of L are still comparatively low w.r.t. the more complex real-world tasks.

In our experiments, we use 10% and 90% from L for development and training, respectively.

4.2. Settings

We use pre-trained RoBERTa [19] for the English ATIS data and pre-trained multilingual BERT [20] (size 768) for the German in-house data, and max-pooling for sentence representation. Each of our decoders has 2 dense layers of size 768 with gelu activation. The dropout values used in IC and SF decoders are 0.5 and 0.2, respectively. For optimization, we use Adam optimizer with learning rate 0.1 and a Noam learning rate scheduler. We trained our model with a batch size of 32. We conduct three teacher training runs. On ATIS, we use (0.8, 0.6) as IC confidence thresholds, while on internal data we use (0.6, 0.4) and (0.6, 0.1) as IC and ASR confidence thresholds, respectively.

We report results using the standard SLU metrics accuracy for IC and F1 for SF. In addition, following previous work [1], we use a semantic error rate, which measures IC and SF jointly and is defined as follows:

$$\textit{SemER} = \frac{\#(\text{slot+intent errors})}{\#\text{slots in reference + 1}} \quad (1)$$

5. Results

To ensure that we start from a strong baseline model, as a first step we trained and evaluated a "standard" SLU model on the overall ATIS dataset. With 97.65% and 95.69% for IC and SF, respectively, model performance is comparable with performance reported in the literature for other BERT-based state-of-the-art SLU models (e.g. [8]).

Table 2 shows the results on ATIS for the supervised baseline which is trained on L only (few shot) and for the different strategies of additionally leveraging $U'_{filtered}$ via teacher-student SSL for both similar and skewed intent data distributions. As can be seen, training the supervised baseline model (few shot) on data with a skewed intent distribution decreases performance when compared to training on data whose intent distribution is representative of the intent distribution in the test data. In particular, intent accuracy drops form 75.92% to 66.18%, highlighting the impact of the intent class distribution in L on model performance. Note that L and U have the same size for both similar and skewed distribution scenarios.

In the more common SSL scenario in which all datasets have

Distrib.	Method	Slot F1	IC acc.	SemER
Similar	Few shot	83.83	75.92	28.43
Similar	Direct inject	85.95	81.19	23.76
Similar	+ weights	85.52	80.96	24.46
Similar	Aux. inject	**87.49**	87.12	17.72
Similar	+ weights	87.25	**92.05**	**16.83**
Skewed	Few shot	78.96	66.18	36.0
Skewed	Direct inject	82.84	79.84	28.44
Skewed	+ weights	83.27	78.84	28.41
Skewed	Aux. inject	83.61	**82.98**	23.42
Skewed	+ weights	**84.13**	82.64	**22.46**

Table 2: *Results on the ATIS for different strategies of leveraging pseudo-labeled data in teacher-student SSL on two data distribution scenarios (i.e. similar vs. skewed intent distributions).*

Distrib.	Method	Testset	Slot F1	IC acc.	SemER
Music domain					
Similar	Direct	Trans.	+3.05	+1.08	-6.51
Similar	Aux.	Trans.	+2.47	+10.4	-14.5
Similar	Direct	ASR	+1.7	+0.3	-4.17
Similar	Aux.	ASR	+3.13	+13.83	-22.72
Skewed	Direct	Trans.	+2.8	+11.78	-13.46
Skewed	Aux.	Trans.	+1.72	+22.73	-19.58
Skewed	Direct	ASR	+1.71	+19.42	-15.38
Skewed	Aux.	ASR	+3.03	+42.47	-31.5
Video domain					
Similar	Direct	Trans.	+1.85	+2.59	-12.4
Similar	Aux.	Trans.	+2.4	+4.66	-16.85
Similar	Direct	ASR	+1.19	+2.18	-6.43
Similar	Aux.	ASR	+1.87	+4.5	-13.15
Skewed	Direct	Trans.	+1.95	+8.27	-15.38
Skewed	Aux.	Trans.	+0.99	+9.34	-14.7
Skewed	Direct	ASR	-1.01	+10.92	-10.33
Skewed	Aux.	ASR	-1.12	+12.36	-11.25

Table 3: *Results on internal data. Relative change in metrics compared to the few shot baseline is reported per scenario (similar vs skewed) and test set (ASR 1-best hypotheses vs transcriptions). Negative numbers indicate better performance for SemER, while positive numbers indicate better performance for slot F1 and IC accuracy.*

similar data distributions, all methods improve performance over the supervised few shot baseline. However, while the common method of injecting L_{aug} directly into the main task improves performance slightly, injecting $U'_{filtered}$ indirectly via the auxiliary task yields an additional boost in performance, increasing intent accuracy from 75.92% to 92.05% when weighting based on IC confidence scores is applied.

In the scenario with skewed intent distributions, gains in IC accuracy over the few shot baseline are comparatively larger already for injecting L_{aug} directly. In particular, the gain in accuracy is 13.66% vs 5.27% absolute in the skewed vs similar scenario, respectively. This suggests that SSL can be used to mitigate the negative impact of training on non-representative labeled data, and to tailor a model towards an unknown intent distribution using unlabeled data. Again, a further improvement is obtained when pseudo-labeled data is used for auxiliary task training compared to a direct injection. The results for applying weighting are overall mixed, yielding improvements in some cases, but slight drops in performance in others. However, we note that this could potentially be improved by applying calibration of weights.

Since weighting based on IC confidence scores gave mixed results in the previous experiments, for internal data we focus on comparing training on L_{aug} (direct inject) and using $U'_{filtered}$ via auxiliary training without weighting. The results are presented in Table 3. Due to confidentiality reasons, we report the relative change in performance compared to the few shot baseline for each scenario (skewed vs similar distributions) and test set (manual transcriptions vs ASR 1-best hypotheses). We present results in relation to the few shot baseline per scenario rather than across scenarios even though the test sets are the same for reasons of clarity, i.e. to allow for more direct comparison between the two methods for leveraging pseudo-labeled data. We note that for internal datasets, in line with the results on ATIS, performance of the few shot baseline drops significantly for the skewed compared to the similar scenario.

Overall, the results follow a similar trend as the previously reported results on ATIS. In particular, we can again observe clearly that in the scenario with skewed distributions, gains in IC accuracy over the few shot baseline are comparatively larger already for injecting L_{aug} directly. Thus, the results provide further evidence that SSL can be used in this scenario to mitigate the negative impact of training on non-representative labeled data and to tailor models towards the target intent distribution.

The results further indicate that – in line with the observations on ATIS – auxiliary task training with $L'_{filtered}$ improves performance over directly training on L_{aug}. The additional boost in performance when using auxiliary task training is comparatively larger on the *Music* than the *Video* domain. In particular, on the standard SSL scenario with similar data distributions, only small gains are obtained with direct injection, while the gain is rather large with auxiliary task training. Recall that for *Music*, the percentage of labeled data is much lower and that the unlabeled data are noisier as indicated by a higher word error rate for *Music*. Thus, auxiliary task training may be particularly beneficial in cases where rather large and noisy unlabeled data amounts are leveraged in teacher-student SSL. In addition, this method provides a particularly large boost in performance when applied on skewed data for *Music*.

When comparing performance differences on manual transcriptions and ASR 1-best hypotheses, gains are particularly large on ASR test sets. Thus, the results indicate that besides becoming stronger on manual transcriptions, models trained via auxiliary training with gradual noising via ASR traffic might also be more robust to noise, such as ASR errors, during application. This is an important benefit, as in a real-world SLU application the input is typically recognized by an ASR module before being fed into the SLU model.

6. Conclusion

In this paper, we studied the impact of intent distributions for semi-supervised SLU. First, we highlighted on the ATIS dataset that training a model on labeled data with an intent distribution which is not representative of that in the test data decreases the performance significantly, in particular for the intent classification task. Applying SSL can mitigate this effect and auxiliary training further improves the results, as this approach can reduce negative effects from using noisy pseudo-labeled data. In addition, we demonstrated the benefit of SSL and in particular auxiliary training in teacher-student SSL on real-word data for two domains, indicating that this approach also helps with making SLU models more robust to noise, such as ASR errors, yielding relative gains of up to 42.47% in intent accuracy and a decrease of up to 31.5% relative in semantic error rate when tested on ASR hypotheses. This is an important benefit for real-world SLU applications which are applied on the output of an ASR module.

7. References

[1] J. Gaspers, P. Karanasou, and R. Chatterjee, "Selecting machine-translated data for quick bootstrapping of a natural language understanding system," *Proceedings of NAACL-HLT*, 2018.

[2] A. Johnson, P. Karanasou, J. Gaspers, and D. Klakow, "Cross-lingual transfer learning for Japanese named entity recognition," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, 2019.

[3] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] D. S. Park, Y. Zhang, Y. Jia, W. Han, C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition," in *INTERSPEECH*, H. Meng, B. Xu, and T. F. Zheng, Eds., 2020.

[5] J. Gaspers, Q. Do, and F. Triefenbach, "Data balancing for boosting performance of low-frequency classes in spoken language understanding," in *Interspeech 2020*, 2020.

[6] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," in *Proceedings of INTERSPEECH*, 2016.

[7] Q. N. T. Do and J. Gaspers, "Cross-lingual transfer learning for spoken language understanding," *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2019. [Online]. Available: http://arxiv.org/abs/1904.01825

[8] Q. Chen, Z. Zhuo, and W. Wang, "Bert for joint intent classification and slot filling," *arXiv:1902.10909*, 2019.

[9] G. Tur, D. Hakkani-Tür, and R. E. Schapire, "Combining active and semi-supervised learning for spoken language understanding," *Speech Communication*, vol. 45, no. 2, pp. 171–186, 2005.

[10] S. Zhu, R. Cao, and K. Yu, "Dual learning for semi-supervised natural language understanding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1936–1947, 2020.

[11] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 3239–3250.

[12] J. He, J. Gu, J. Shen, and M. Ranzato, "Revisiting self-training for neural sequence generation," 2020.

[13] C.-I. Lai, Y.-S. Chuang, H.-Y. Lee, S.-W. Li, and J. Glass, "Semi-supervised spoken language understanding via self-supervised speech and language model pretraining," 2020.

[14] T. Desot, F. Portet, and M. Vacher, "SLU for voice command in smart home: comparison of pipeline and end-to-end approaches," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Sentosa, Singapore, Singapore, Dec. 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02464393

[15] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," 2019.

[16] D. Hakkani-Tür, F. Béchet, G. Riccardi, and G. Tür, "Beyond asr 1-best: Using word confusion networks in spoken language understanding." *Comput. Speech Lang.*, vol. 20, no. 4, pp. 495–514, 2006. [Online]. Available: http://dblp.uni-trier.de/db/journals/csl/csl20.html#Hakkani-TurBRT06

[17] G. Tür, D. Hakkani-Tür, and L. P. Heck, "What is left to be understood in atis?" in *SLT*, D. Hakkani-Tür and M. Ostendorf, Eds. IEEE, 2010, pp. 19–24. [Online]. Available: http://dblp.uni-trier.de/db/conf/slt/slt2010.html#TurHH10

[18] W. Ruan, Y. Nechaev, L. Chen, C. Su, and I. Kiss, "Towards an asr error robust spoken language understanding system," in *INTERSPEECH*. ISCA, 2020.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423