Learning under Label Noise for Robust Spoken Language Understanding Systems

Anoop Kumar, Pankaj Sharma*, Aravind Illa, Sriram Venkatapathy, Subhrangshu Nandi, Pritam Varma, Anurag Dwarakanath and Aram Galstyan*

Alexa AI, Amazon

{anooamzn, spk, aravini, vesriram, subhrn, spv, adwaraka, argalsty}@amazon.com

Abstract

Most real-world datasets contain inherent label noise which leads to memorization and overfitting when such data is used to train over-parameterized deep neural networks. While memorization in DNNs has been studied extensively in computer vision literature, the impact of noisy labels and various mitigation strategies in Spoken Language Understanding tasks is largely under-explored. In this paper, we perform a systematic study on the effectiveness of five noise mitigation methods in Spoken Language text classification tasks. First, we experiment on three publicly available datasets by synthetically injecting noise into the labels and evaluate the effectiveness of various methods at different levels of noise intensity. We then evaluate these methods on a real-world data coming from a large-scale industrial Spoken Language Understanding system. Our results show that most methods are effective in mitigating the impact of noise with two of the methods showing consistently better results. For the industrial Spoken Language Understanding system, the best performing method is able to recover up to 97% of the loss in performance due to noise.

1. Introduction

Deep neural networks (DNNs) have demonstrated impressive performance on several supervised learning tasks including speech recognition, computer vision, natural language understanding and have been adopted in large-scale industrial systems. One of the most striking phenomena about DNNs is it's ability to generalize well in the over-parameterized regime where the number of model parameters is much larger than the number of available labeled data instances. However, such over-parameterized models are susceptible to label noise. Indeed, recent work [1] has shown that DNNs are able to achieve perfect training accuracy even when the labels are randomly permuted. This undesirable behavior is due the ability of DNNs to memorize noisy examples and thus recent work has focused on mitigating the effects of memorization through various approaches.

While memorization in DNNs has been studied extensively in computer vision tasks, the impact of noisy labels in Spoken Language Understanding (SLU) tasks such as text classification is largely under-explored. Despite some notable exceptions [2], we lack an understanding of the severity of memorization in text classification and good strategies for limiting the negative impact of noise in labels.

In this paper, we address this gap by conducting an exhaustive study examining the role of label noise in several multi-class classification tasks. Our contributions are as follows: i) we demonstrate extreme overfitting by DNNs in the presence of label noise under a uniform noise model; ii) we explore five different label noise mitigation strategies and adopt them for text classification settings; iii) through extensive experimentation on three publicly available datasets and an internal dataset from a large-scale SLU system, we observe that different methods are indeed able to limit the impact of noise to varying degree of effectiveness.

In addition to the experiments with synthetically injected noise, we conducted experiments on an internal data from a large-scale SLU system. While the experiments on public data focuses on model performance under synthetically injected label noise, the experiments on internal data help us to assess different mitigation methods under a more realistic scenario. In particular, the internal datasets contain noise that occurs in practice due to the nature of manual annotations and inherent ambiguities in data. Our results show that all methods are indeed able to recover from label noise in practice as well. We also find that while no single method shows the best performance across different datasets, Label smoothing [3] with early stopping [4] and Limiting label Information Memorization In Training (LIMIT) [5] show more consistent improvements.

The rest of the paper is structured as follows. We briefly review related work in Section 2. We present the problem formulation in Section 3 and the different mitigation approaches are detailed in Section 4. The datasets and the experimental setup are provided in Section 5 with the results in Section 6. We conclude with directions for future research in Section 7.

2. Related Work

Classification under label noise has been a long standing problem. Several methods have attempted to mitigate the effect of noise by designing a robust loss function [6] or by estimating label-noise [7, 8, 9]. Some works use meta-learning to treat the problem of noisy/incomplete labels as a decision problem in which one determines the reliability of a sample [10, 11], while others seek to detect incorrect examples and relabel them [12, 13, 14].

Other noise mitigation approaches, such as LIMIT [5], work by adding an information theoretic regularization to the objective function and have illustrated the effectiveness of the approach on versions of MNIST, CIFAR-10, and CIFAR-100 datasets.

Additionally, we study and compare the mitigation effectiveness of less complex approaches, such as Label Smoothing [3] and Early Stopping [4]. Label Smoothing introduces noise in label by converting hard 0 or 1 classification targets to softer values between 0 an 1. Early stopping prevents overfitting by stopping the training when validation error starts to increase.

While most of the previous work have focused on vision tasks, recently, [15] studied the impact of label noise for text

* equal contribution

classification problems and proposed a noise mitigation method by adding a layer to the classifier. To the best of our knowledge, our work is one of the first to compare the leading noise mitigation methods on both public and industrial data for SLU text classification tasks.

3. Problem Statement

We consider a typical K-class classification problem with n i.i.d. samples $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ corresponds to the input vector for the i-th data point and $\mathbf{y}_i = \{0,1\}^K$ with $\mathbf{y}_i^T \mathbf{1} = 1$ corresponds to the label. For the classification tasks considered $\mathbf{x}_i$ correspond to (pre-trained) encoding of the text to be classified. The classifier that maps input to labels is a non-linear function $f(\cdot)$ parameterized by a DNN. The accuracy of the trained classifier is evaluated on the *clean* dataset (i.e. the test set that does not contain any label noise).

We conduct experiments on both the public and internal industry SLU data. While noise in the labels is naturally occurring in the industrial data, we inject artificial noise into the publicly available datasets.

Noise Model To inject noise in the labels, we use a uniform noise model where the label of each data point is randomly flipped according to the following transition

$$p(\bar{y}|y, \mathbf{x}) = (1-p)\mathbf{I} + p\mathbf{\Pi} \tag{1}$$

where $\mathbf{I}$ is the identity matrix, p characterizes the noise intensity, and the $K \times K$ matrix $\mathbf{\Pi}$ is matrix with zeros on the diagonal, and $\mathbf{\Pi}_{k \neq l} = \frac{1}{K-1}$, $k, l = 1, ..K$. In other words, the label of each data point is randomly flipped with uniform probability over the other $K-1$ classes. Note that in this model, the noise does not depend on the input.

4. Mitigation Methods

In this section, we describe the different noise mitigation methods used in our experiments.

Noise Layer The method of Noise Layer[15] learns the noise distribution adding an additional final layer to a DNN. As reported in the work, our implementation uses a noise layer with softmax non linearity with layer weights initialised to the identity matrix. The model is trained using early stopping. During inference, the noise layer is removed and the base model is used for the final predictions.

Robust Loss The method of Robust Loss [16] proposes a loss function called Active Passive Loss (APL) which is a combination of an "active" and a "passive" loss function. The former loss explicitly maximizes the probability of being in the labeled class, while the latter minimizes the probabilities of being in other classes. Among the different APL loss functions, combination of normalized cross entropy (NCE) and reverse cross entropy (RCE) were shown to achieve state-of-the art performance. In our work, we perform experiments with αNCE + βRCE as robust loss function.

LIMIT The method of LIMIT [5] proposed a noise mitigation method that works by adding an information theoretic regularization to the objective function and tries to minimize the mutual information between model weights and the labels conditioned on data instances: $I(w : y|x)$. Since this objective is hard to minimize directly, LIMIT leverages an auxiliary network that predicts gradients in the final layer of a classifier without accessing label information.

Label Smoothing Label Smoothing [3] is a regularization technique that introduces noise for the label to account for the fact that datasets could have mistakes. Assume for a small constant ϵ, the training set label y is correct with probability $1-\epsilon$ and incorrect otherwise. Label Smoothing regularizes a model based on a softmax with k output values by replacing the hard 0 and 1 classification targets with targets of $\frac{\epsilon}{K-1}$ and $1-\epsilon$ respectively.

Early Stopping The method of Early Stopping [4] aims to prevent overfitting by stopping the training when the generalization error starts to increase. Error on the validation set is used as a proxy for generalization error and training is stopped when validation error starts to increase for a certain number of epochs.

5. Datasets and Experimental Setup

5.1. Publicly available datasets

We use three publicly available and widely used SLU datasets - Air Travel Information System (*ATIS*) corpus [17], *SNIPS* [18], which is collected from Snips personal voice assistant, and task oriented parsing (TOP) datasets from Facebook [19].

The public datasets consists of utterances along with intent and slot labels. The TOP dataset is available for multiple languages and we work with the English portion of the data. We focus on intent classification tasks which involves predicting one of the labels for a query utterance. Table 1 provides summary of datasets including the number of utterances in training (N) and test (T) sets; the number of classes (K); the average word length of utterances (L); and the balance factor - B. The balance factor is computed using equation 2 where n_i is the number of instances in class i. A balance factor of 0 indicates complete imbalance and 1 perfect balance.

$$\frac{-\sum_{i=1}^{K} \frac{c_i}{N} \log \frac{c_i}{N}}{\log K} \tag{2}$$

Dastaset	N	T	L	K	B
ATIS	4478	893	11	21	0.42
SNIPS	13084	700	9	7	1
TOP	28414	8241	16	9	0.63

Table 1: *Summary of the text classification datasets.*

The industrial dataset consists of a random sample of data used to train a domain classifier for a large-scale SLU implementation. The training data is annotated by human data associates (DA) and is known to contain annotation errors. Additionally, we have access to gold data that goes through rigorous quality checks with multiple experienced DAs and has fewer or no annotation errors. Similar to intent classification task on public datasets, the domain classification (DC) task hypothesizes a target domain such as Music, Books, etc. given an utterance text [20]. We evaluate all our models on gold data test set and report the improvement in accuracy of each method on overall data and top three domains. Due to the nature of industrial data,

we only provide the duration in hours for training and test data of the three studied domains in Table 2.

Domain	Train (in hours)	Test (in hours)
Domain 1	49	6
Domain 2	47	6
Domain 3	57	7
Overall	153	19

Table 2: *Dataset distribution of the random samples from the industrial dataset used for experiments*

5.2. Multi-Class Classification Approach

For the multi-class single sentence classification task as described in Section 3, we leverage BERT model architecture [21]. The input to the BERT model is WordPiece, positional and sequence embeddings. A special classification embedding ([CLS]) is inserted as the first token and a special token ([SEP]) is added as the final token. Given an input token sequence $t = (t_1, t_2, t_3, ..., t_S)$, the BERT output is $H = (h_1, h_2, ...h_S)$.

The BERT model is pre-trained in self-supervised fashion on two tasks - next sentence prediction and mask language models and provides powerful sentence representation for multiple NLP tasks [21]. Similar to the approach used for intent classification in [22], we add a softmax layer on top of hidden state of the first special token ([CLS]) denoted by h_1. The probability for class i, y^i is given by equation 3, where W^i and b^i are weights and biases respectively.

$$y^i = softmax(W^i h_1 + b^i) \quad (3)$$

5.3. Training and Evaluation Details

We use the DistilBERT model from Hugging Face library. DistilBERT is trained by distilling BERT base model, which has 40% fewer parameters than bert-base-uncased and runs 60% faster while preserving over 95% of BERT's performance as measured on the GLUE language understanding benchmark [23]. DistillBERT has 6 layers, 12 heads, and 768 hidden states.

All hyper-parameters of mitigation methods are tuned using the Tune library [11]. The batch size is 32. Adam is used for optimization with an initial learning rate of 5e-6. The dropout probability is set to 0.1. All the experiments are run on p3.16xlarge AWS instances that contain 8 GPUs, 64 vCPUs, 128GB GPU memory and 488 GB CPU memory.

We report the accuracy, which is the ratio of correctly predicted to all instances. We run the training for 100 epochs when Early Stopping is not used. We run experiments three times and report mean accuracy using seeds 1, 2 and 3, since the noise model randomly flips the utterance label to incorporate noise.

6. Results and Discussion

6.1. Evidence of Overfitting

Figure 1 shows the training, validation (dev) and test accuracy over epochs. Our first observation is that the training error converges to 100% accuracy, signalling extreme overfitting. At the same time, both validation and test accuracies increases at the beginning of the training and then rapidly deteriorate. We observe that the peak accuracies on both validation and test sets are well aligned, which justifies early stopping as an effecting baseline strategy for mitigating the noise in such cases.

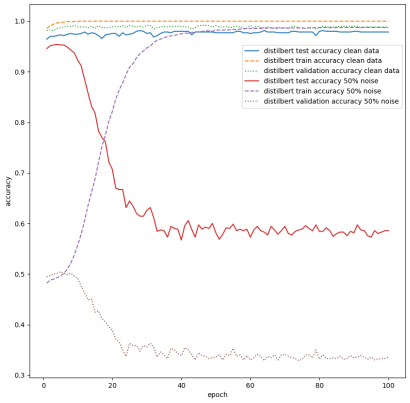

Figure 1: *Training, dev, and test accuracy over learning epochs for the SNIPS dataset with 50% label noise*

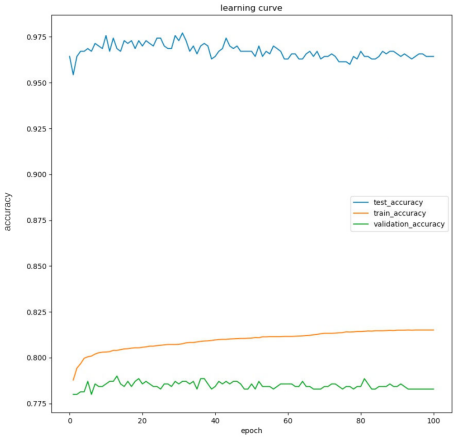

Figure 2: *Learning Curve for Robust Loss mitigation method for SNIPS dataset with 20% noise*

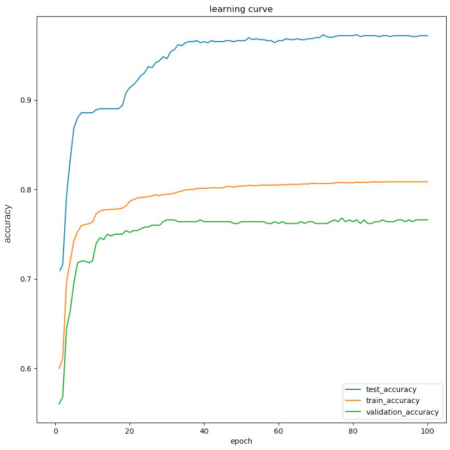

Figure 3: *Learning Curve for LIMIT mitigation method for ATIS dataset with 20% noise*

The noise mitigation approaches fall into three categories. In first category, training is stopped when overfitting starts, such as Early Stopping. Second category delays or reduce the impact of overfitting and can further benefit from early stopping, such as Noise Layer and Label Smoothing. The third kind updates the loss function in such a way that model doesn't overfit, which occurs in Robust Loss and LIMIT. Learning curve for Robust Loss is shown in Figure 2 and for LIMIT in Figure 3, we don't observe any drop in the accuracy of validation set. These methods do not seem to benefit from early stopping. There may not be an increase or drop in validation accuracy for a few steps, as observed in Figure 3. In such scenarios, Early Stopping is not the most effective solution and the number of steps have to be increased steadily.

6.2. Public Datasets

Table 3 presents the intent classification accuracy on public datasets. As expected, we observe drop in accuracy for all 3 datasets as the noise is added to the training data. The accuracy on ATIS, SNIPS, and TOP datasets drop from 97.31 to 64.69, 97.86 to 59.43 and 98.75 to 59.61 percent respectively. Early stopping is able to recover most of the dropped accuracy in all 3 datasets. Robust Loss demonstrates the best performance on SNIPS and Label Smoothing with Early Stopping has the best performance on TOP dataset. Early Stopping has only 2.4% and 1.04% absolute drop in accuracy at 50% noise and other methods show marginal improvements over Early Stopping. Unlike in the case of ATIS where Early Stopping has a drop of 7.88% absolute drop in accuracy and LIMIT produces 1.87% absolute improvements on Early Stopping. We also observe improvements in clean ATIS and SNIPS dataset, which indicates the publicly available data may contain noise.

Next, we measure how much accuracy mitigation methods are able to recover on 20% noise to compare it with industrial dataset in Table 4. LIMIT performs best and able to recover 98.04%, 96.18% and 97.07% of the dropped accuracy, followed by Label Smoothing which recovered 78.67%, 85.62% and 95.21% accuracy on ATIS, SNIPS and TOP datasets respectively. Early Stopping matched Label Smoothing on SNIPS(85.62%) and performed third on ATIS and TOP.

Dataset	Model	Clean	10	20	30	40	50
ATIS	DistilBERT	97.31	94.51	89.62	85.74	75.74	64.69
	ES	97.09	96.98	95.52	94.70	92.42	89.44
	Noise Layer	97.31	96.04	94.62	90.52	80.74	80.74
	Robust Loss	**97.76**	96.01	95.20	94.18	**95.16**	88.84
	LIMIT	**97.76**	**97.50**	**97.16**	**96.19**	94.74	**91.30**
	LS	97.26	96.23	95.67	94.70	92.91	89.77
SNIPS	DistilBERT	97.86	93.14	87.91	79.81	70.95	59.43
	ES	97.43	97.05	96.43	96.43	96.14	95.48
	Noise Layer	97.23	96.71	96.38	94.24	84.52	81.76
	Robust Loss	**98.00**	97.38	97.38	96.86	**96.57**	**96.10**
	LIMIT	**98.00**	**97.52**	**97.48**	**97.05**	96.14	94.68
	LS	97.57	97.05	96.43	96.71	96.14	95.48
TOP	DistilBERT	**98.75**	94.17	88.51	79.54	70.19	59.61
	ES	98.45	**98.37**	98.24	98.05	97.78	97.72
	Noise Layer	97.31	96.04	94.62	90.52	80.74	80.74
	Robust Loss	96.72	96.62	96.65	96.58	97.04	97.39
	LIMIT	98.58	97.62	**98.45**	**98.30**	**98.04**	97.63
	LS	98.45	**98.37**	98.27	98.08	97.79	**97.72**

Table 3: *Accuracy of various mitigation methods on public datasets. Best accuracy for clean and error rates is highlighted. ES - Early Stopping, LS - Label Smoothing.*

6.3. Industrial Dataset

We now report our results on applying noise mitigation techniques on a large-scale industrial SLU system. In our experiments, we have noisy and clean (gold) versions of the dataset. We train our models on the noisy version and test them on the gold counterpart. Table 4 presents the results for the three domains. For reference, we also train models with gold data of the same duration and report the accuracy gain for the model. We believe the noise mitigation models should vie to achieve the accuracy of the gold model.

First, we note that training on the noisy data results in a measurable accuracy drop. Indeed, the last row of Table 4 is the relative improvement in performance when trained on gold data over the one trained on noisy data. On an average, training on gold data yields 2.69% relative improvement in accuracy.

Then we look at the effect of different noise mitigation strategies. Early Stopping (1.54%), Robust Loss (1.07%) and LIMIT (0.85%) seem to recover 57%, 40% and 31% of the underfitting caused by the noise in the data. Robust Loss (3.74%) comes closest to Gold (3.86%) on Domain 1. Noise Layer and Label Smoothing don't generate best accuracy gain overall on either of the domains. Overall, there isn't one method that works best across the board and multiple mitigation approaches should be explored to pick appropriate one for the dataset. However, we see that Early Stopping, LIMIT and Robust loss outperform the others in both public datasets and industrial dataset.

Model	Domain 1	Domain 2	Domain 3	Overall (%)
Early Stopping	2.92	0.11	**1.69**	**1.54** (57%)
Noise Layer	0.94	-0.11	-0.66	-0.01 (-0.2%)
Robust Loss	**3.74**	**0.31**	-0.38	1.07 (40%)
LIMIT	1.99	-0.85	1.41	0.85 (31%)
Label Smoothing	2.46	-1.06	0.38	0.52 (19%)
Gold	3.86	1.49	2.82	2.69(100%)

Table 4: *Accuracy improvement of various mitigation methods on industrial datasets.*

7. Conclusion

In conclusion, we have conducted a systematic study of Spoken Language Understanding text classification task in the presence of label noise in public dataset and the industrial system. Our results indicate that DNNs are indeed prone to extreme overfitting that leads to poor generalization abilities of learned models for spoken language classification task. We also observed that various methods are able to mitigate the impact of synthetically injected noise in public datasets, and there is not one method that works well across the board. Our experiments with synthetic data also indicate that even simple early stopping heuristics can be effective in mitigating the impact of label noise.

For the experiments with the real-world data, we found that early stopping is still useful and as competitive as more elaborate mitigation methods. However, we see there is scope for further improvements in real-world data. This might be due to the fact that the label noise distribution in real-world datasets is more complex than the simple label-flipping model considered here. It is an interesting open problem to characterize the noise distribution in real systems and devise better mitigation strategies tailored to such noise models.

8. References

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[2] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, "A closer look at memorization in deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 233–242.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[4] M. Li, M. Soltanolkotabi, and S. Oymak, "Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 4313–4324.

[5] H. Harutyunyan, K. Reing, G. V. Steeg, and A. Galstyan, "Improving generalization by controlling label-noise information in neural network weights," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 4071–4081.

[6] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Advances in neural information processing systems*, 2013, pp. 1196–1204.

[7] S. Sukhbaatar, J. Bruna, M. Paluri, L. D. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," in *ICLR 2015*, 2014.

[8] J. Goldberger and E. Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," in *ICLR*, 2017.

[9] D. Hendrycks, M. Mazeika, D. Wilson, and K. Gimpel, "Using trusted data to train deep networks on labels corrupted by severe noise," in *Advances in neural information processing systems*, 2018, pp. 10 456–10 465.

[10] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," *arXiv preprint arXiv:1803.09050*, 2018.

[11] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, and D. Meng, "Meta-weight-net: Learning an explicit mapping for sample weighting," in *Advances in Neural Information Processing Systems*, 2019, pp. 1917–1928.

[12] S. E. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," *CoRR*, vol. abs/1412.6596, 2014.

[13] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, "Joint optimization framework for learning with noisy labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5552–5560.

[14] J. Han, P. Luo, and X. Wang, "Deep self-learning from noisy labels," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5138–5147.

[15] I. Jindal, D. Pressel, B. Lester, and M. Nokleby, "An effective label noise model for DNN text classification," 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. 3246–3256. [Online]. Available: https://www.aclweb.org/anthology/N19-1328

[16] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, and J. Bailey, "Normalized loss functions for deep learning with noisy labels," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6543–6553.

[17] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "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*, 1990. [Online]. Available: https://www.aclweb.org/anthology/H90-1021

[18] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril *et al.*, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *arXiv preprint arXiv:1805.10190*, 2018.

[19] S. Gupta, R. Shah, M. Mohit, A. Kumar, and M. Lewis, "Semantic parsing for task oriented dialog using hierarchical representations," *arXiv preprint arXiv:1810.07942*, 2018.

[20] C. Su, R. Gupta, S. Ananthakrishnan, and S. Matsoukas, "A reranker scheme for integrating large scale nlu models," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 670–676.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

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

[23] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6