Training data reduction for multilingual Spoken Language Understanding systems

Anmol Bansal [*†§], **Anjali Shenoy**[*‡], **Chaitanya P. K.**[‡], **Kay Rottmann**[‡], **Anurag Dwarakanath**[‡]
[‡]Alexa AI, Amazon
[§]IIT Kharagpur
{anshen, kppappu, krrottm, adwaraka}@amazon.com
anmolbansal632@gmail.com

Abstract

Fine-tuning self-supervised pre-trained language models such as BERT has significantly improved state-of-the-art performance on natural language processing tasks. Similar fine-tuning setups can also be used in commercial large scale Spoken Language Understanding (SLU) systems to perform intent classification and slot tagging on user queries. Fine-tuning such powerful models for use in commercial systems requires large amounts of training data and compute resources to achieve high performance. This paper is a study on the different empirical methods of identifying training data redundancies for the fine tuning paradigm. Particularly, we explore rule based and semantic techniques to reduce data in a multilingual fine tuning setting and report our results on key SLU metrics. Through our experiments, we show that we can achieve on par/better performance on fine-tuning using a reduced data set as compared to a model fine-tuned on the entire data set.

1 Introduction

In recent years, a variety of smart voice assistants such as Apple's Siri, Samsung's Bixby, Google Home, Amazon Echo, Tmall Genie, have been deployed and achieved great success. These voice assistants facilitate goal-oriented dialogues and help users to accomplish their tasks through voice interactions. One component of such spoken language understanding (SLU) systems is Natural Language Understanding (NLU) which aims to extract the intent of the query (intent classification) and semantically parse the user utterance (slot tagging). As an example, if a user requests "play madonna" to the voice assistant, SLU would classify the intent as *PlayMusic* with slot filling of tokens "play" as *Action* and "madonna" as *Artist*.

As in many other language processing fields, pre-trained language models have seen major success for natural language understanding. Pre-trained language models (Radford and Narasimhan, 2018; Howard and Ruder, 2018; Baevski et al., 2019; Dong et al., 2019) are generic language models learned in a semi-supervised fashion whose underlying large scale knowledge is then leveraged for fine-tuning towards down-stream tasks (Ruder et al., 2019). BERT (Devlin et al., 2019) is one such example of a language model based on the Transformer Network architecture (Vaswani et al., 2017), pre-trained on a corpora of 3300M words extracted from publicly available unannotated data and then fine-tuned on smaller amounts of supervised data for specific tasks, relying on the induced language model structure to facilitate generalization beyond the annotations. It provides powerful and general-purpose linguistic representations, triggering strong improvements and significant advances on a wide range of natural language processing tasks. Chen et al. (2019) also observed the success of BERT to jointly learn intent classification (IC) and slot filling (SF) tasks by leveraging pre-trained representations. Leveraging this joint IC and SF set up to interpret user utterances in commercial SLU systems requires large volume of annotated training data (Ezen-Can, 2020), compute resources (such as GPUs) and model build time to cover the variability of customer utterances. Such resources (human and compute) are not only expensive but also time consuming, unscalable and may not be fully optimized to achieve the same performance.

In this paper, we perform an empirical analysis on identifying subsets of multilingual training data which can achieve on par or better performance as compared to the same model trained

[*]equal contribution
[†]work done while interning with Alexa AI

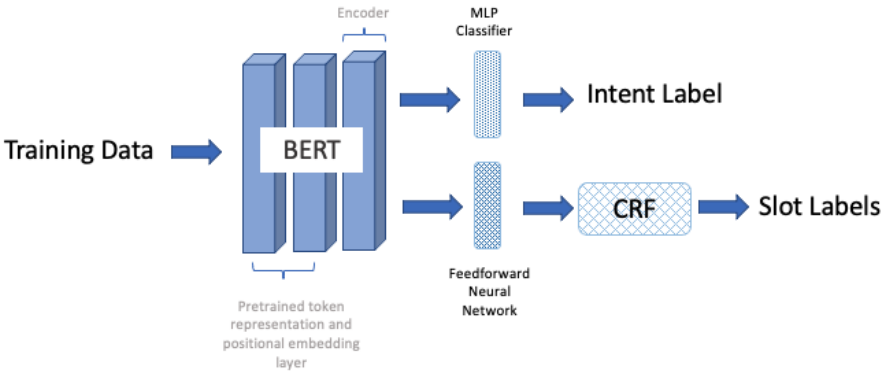


Figure 1: IC/SF Bert architecture

on the full data-set for the same task. In particular, we fine-tune a Lean-BERT sized model (Conneau et al., 2020), on the IC-SF task for Hindi and English SLU data. Such mutlilingual large scale pre-training is known to effectively promote cross-lingual generalization (Choi et al., 2021; Pires et al., 2019) giving rise to an opportunity to exploit latent space similarities of such languages to identify redundancies in training data for fine-tuning. We experiment with various semantic and rule based data reduction approaches and report fine-tuning performance on key SLU metrics.

2 Related work

Finding the right data reduction technique for BERT fine tuning while maintaining evaluation performance can be considered as a part of two major classes of problems - fine tuning regime in the low resource setting to leverage insights from incorporated best practices, and the few shot classification class of problems where the model is trained using only a few samples from each class. Note that the above two classes of problems are not disjoint and is concurrently explored in this work.

2.1 Low resource fine tuning

A newly discovered approach to fine-tune a transformer based model using low resource data is pre-finetuning, introduced in Aghajanyan et al. (2021). In pre-finetuning, the BERT based model undergoes large-scale multi-task learning between language model pre-training and fine-tuning to encourage learning of representations that generalize better to many different downstream tasks. Pre-finetuning gains are particularly strong in the low resource regime, where there is relatively little labeled data for fine-tuning. Our proposed approaches can be used as an extension on top of pre-finetuning to use the gains of the pre-finetuning and benefit from smaller data-sets during the real finetuning.

Active learning is also a widely popular space involving few shot learning where the number of examples to learn a concept are much lower than that required in a normal supervised learning setting. Grießhaber et al. (2020) explore active learning in conjunction with BERT finetuning in the low resource setting with less than 1000 data points. The method involves using Bayesian approximations of model uncertainty (Gal and Ghahramani, 2016) to efficiently select unannotated data for manual labeling. The method utilizes pool-based active learning to speed up training while keeping the cost of labeling new data constant. They also demonstrate the benefits of freezing layers of the pre-trained language model during fine-tuning to reduce the number of trainable parameters, making it more suitable for low-resource setting. Drawing inspiration from this, we conduct our experiments by initially freezing the input embedding layer and gradually unfreezing it by applying an increasing fraction of the learning rate over the training steps.

Shnarch et al. (2021) introduce a new unsupervised learning layer between pre-training and fine-tuning called the Clustering Layer which helps train BERT on predicting cluster labels and can significantly reduce the demand for labeled examples for topical classification tasks. This technique however affects the overall latency of the model in real time systems and we only wish to consider those techniques which modify the input training data rather than the model itself.

Zhang et al. (2021) explore commonly observed instabilities in few-sample scenarios for fine-tuning BERT. Several factors which were identified as causes of instability were the limited applicability

of significant parts of the BERT network for downstream tasks and the prevalent practice of using a pre-determined small number of training iterations. We have leveraged insights from this work and accordingly tuned the various hyper parameters of our model.

2.2 Few shot classification & entity recognition

While few shot and one shot learning techniques are very popular in computer vision for tasks such as image recognition (Koch, 2015), in NLP Lampinen and McClelland (2018) was the first to introduce one-shot and few-shot learning for word embeddings. Geng et al. (2019) explore leveraging the dynamic routing algorithm in meta-learning (Yin, 2020) to simulate the few-shot task and introduce a novel Induction Network to learn generalized class-wise representations. Huang et al. (2020) explore few shot learning for the entity recognition task with meta learning, supervised pre-training (similar to BERT) and self-training to leverage unlabeled in-domain data. Yang and Katiyar (2020) explore entity recognition in the nearest-neighbour paradigm.

The first works in data reduction techniques in Machine Learning (Wilson and Martinez, 2004; Arnaiz-González et al., 2016) are based on instance selection methods broadly classified into two categories. The first is the incremental method which begins with an empty set and the algorithm keeps adding instances to the this subset by analyzing instances in the training set. The decremental method, on the contrary, starts with the original training data set removes those instances that are considered superfluous or unnecessary. We would consider our approach of selecting the subset of data as a decremental method since we start from the original set and proceed to extract a smaller set from it.

Koh and Liang (2020) introduced the concept of influential data instances - those training points which are most responsible for a given prediction - and how to identify them. However, this approach can only be applied to machine learning models trained on convex losses and is also not scalable due to the computationally heavy Hessian matrix multiplication involved. Pruthi et al. (2020) extended this concept to estimate training data influence by tracing its gradient descent. Using first-order approximation for Hessian computation and extending the algorithm to mini-batches, they made this approach scalable and showed results on an image classification task. This is however unexplored in the language processing setting for joint intent classification and slot filling task which is more complex than binary classification.

3 Method

In this paper, we first describe the SLU architecture used for the IC-SF task and the four methods for data reduction.

3.1 SLU model Architecture

We use a common SLU architecture (Chen et al., 2019) for joint intent classification and slot filling, which is depicted in figure 1. It consists of a BERT based 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 one standard task specific layer and a softmax layer on top. Meanwhile, the slot decoder uses a CRF layer on top of one task specific layer 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 as in Chen et al. (2019)

3.2 Data reduction approaches

We define terminologies that we use throughout the paper as the following - Let an utterance $u_i \in S$, where S is the set of all utterances in the training data, have an intent $intent_i \in I$ and annotation $a_i \in A$, where *I* and *A* is the set of all intents and annotations and $a_i \in A$ is a string of tokens with each token annotated with a slot label. From our previous example, u_i = "play madonna", $intent_i$ = PlayMusic, a_i = "play<Action> madonna<Artist>"

3.2.1 Baseline

In the baseline, we fine-tune model on the entire data set (S, I, A) without any modifications.

3.2.2 Unique

In this method, for an utterance $u_i \in S$, we extract unique utterance annotations filtered using annotation $a_i \in A$ as a key. This helps in uniformly representing variations in SLU data by removing any bias due to frequency of occurrence.

3.2.3 Log N

In Log N reduction, if an utterance $u_i \in S$ has a frequency of occurrence n_i, we downsample the

Domain	Unique	Log	Clustering 70%	Singular Score
Music	-55%	-38%	-30%	-25%
Shopping	-51%	-40%	-30%	-35%
Video	-44%	-24%	-27%	-18%
Notifications	-52%	-36%	-33%	-11%
Weather	-53%	-38%	-20%	-3%

Table 1: Reduction achieved by different techniques.

utterance to have a frequency $\log_2(n_i)$. This maintains the utterance distribution as in the original dataset but reduces the absolute magnitude of the frequency. This way the model learns the original input distribution of the SLU data but the reduced representation helps avoid overfitting the model to the more prevalent classes. We experimented with other variants of Log N subsampling such as k-Log N where $k \in \mathbb{N}$ and would involve scaling the frequency to k times $\log_2(n_i)\ \forall u_i \in S$ but we did not see any significant gains in this approach.

3.2.4 Clustering

The first two methods described above only account for the frequency statistics of the utterance in the training data and is language agnostic. In the clustering approach, we try to reduce the data distribution *semantically*.

The steps in the clustering approach are as follows:

- Identify a subset of intents $I' \subseteq I$ by filter for those where the number of utterances $u_i \in S$ labeled with $intent_i \in I'$ is greater than 1000. This is done so that we do not reduce the data from underrepresented intents.
- Extract unique utterances for all utterances having $intent_i \in I'$ using a_i as the key. Repetitions of utterances in the data will have the same word embedding representation and creates redundancy in the input to the clustering algorithm and a compute resource bottle neck.
- Extract embedding representation $e_i \in \mathbb{R}^m$ having dimensions m for these unique utterances from the max pooling representation of the model's [CLS] token. (Devlin et al., 2019).

$$M_{n\times m} = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{bmatrix}$$

- Condense these unique utterance's embedding representations into a smaller number of dimensions $d < m$ by computing SVD on the input matrix

$$M_{n\times m} = U_{n\times m}\Sigma_{m\times m}V_{m\times n}$$

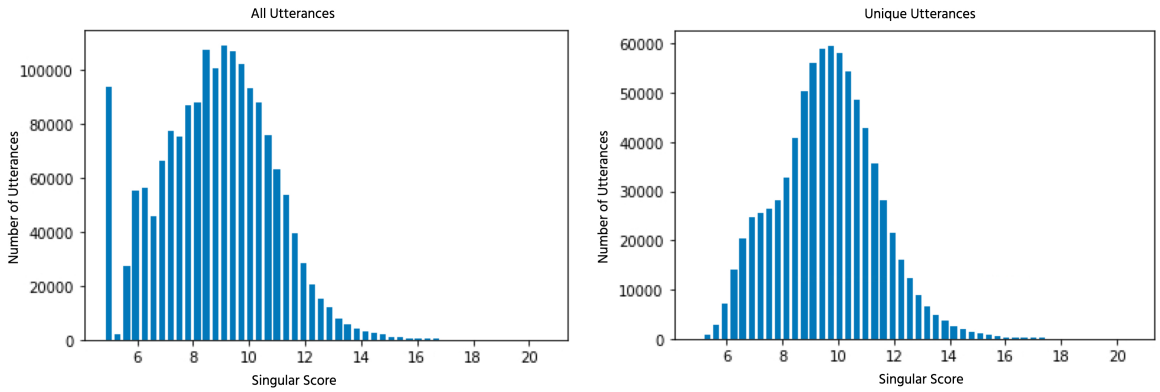
$$M_{n\times m}\bar{V}^T_{m\times d} = U_{n\times m}\bar{\Sigma}_{m\times d}$$

  where $\bar{V}^T_{m\times d}$ and $\bar{\Sigma}_{m\times d}$ are simply the first $d < m$ columns of V and Σ.
- Obtain the condensed representation for each unique utterance's data point in the rows of $U_{n\times m}\bar{\Sigma}_{m\times d}$. Note that unlike PCA we do not normalize the input here since it is computationally expensive.
- For each $intent_i \in I'$, perform k-means clustering from the extracted and condensed utterance embedding representations and find the optimal number of clusters K using the Elbow Method.
- Restore the frequency of the clustered utterances to the original frequency as observed in the full dataset S.
- Per cluster $k_i \in K$ where the clustered utterances have their original frequency of utterance, randomly sub-sample 30% of the utterances and use the remaining 70% as the training set.

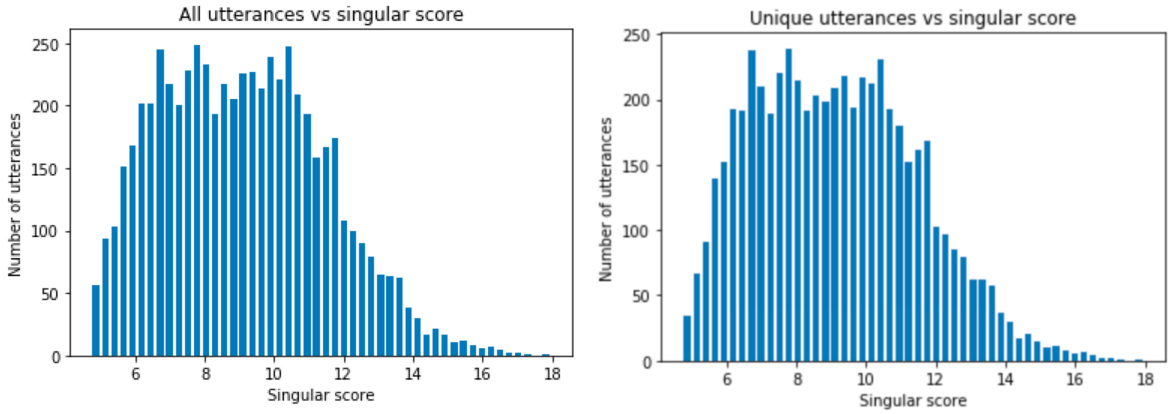
Note that we experimented with choosing a subset by randomly subsampling [10%, 20%, 30%, 50%] of the data and observed that subsampling 30% of the data had the best balance with respect to amount of reduction versus performance drop.

3.2.5 Singular Score

Golub and Reinsch (1971) introduced Singular Value Decomposition (SVD), a technique to factorize the matrix into two unitary matrices and a diagonal matrix. The diagonal values of this matrix are the singular values. This approach has been used extensively in multiple fields since its onset in 1970s such as the work described in (Kabsch,



(a) Our dataset



(b) Multi ATIS++ dataset

Figure 2: Singular Scores Distribution

1978), which uses SVD to compute an ideal rotation matrix for 3-D molecular comparisons. (Walton et al., 2013) explores the decomposition offered by SVD to reduce the degrees of freedom and interpolate the flow problem to lower complexity, with minimal loss in accuracy of representation. In the field of Statistics and Machine Learning, it has been primarily used to achieve dimensionality reduction with minimal loss in information content. One such application in field on Information Retrieval is Latent Semantic Analysis (LSA) (Furnas et al., 1988) where sparse representations of documents were reduced significantly to a few dimension that hold most information and these were used as representations for the original documents.

For this work, we explore using SVD on subsampling data-points instead of subsampling dimensions as in regular applications. As seen in equation 1, M is the embedding matrix with n datapoints and m dimensions per datapoint. We performed experiments with treating data-points analogous to dimensions and subsampling them. However, this wasn't favorable as the datapoints being treated as dimensions for reduction were very large in number and did not have the correlation factor as seen with regular dimensions of embeddings.

$$M_{n\times m} = U_{n\times m}\Sigma_{m\times m}V^T_{m\times m} \tag{1}$$

$$B = MV = U\Sigma \tag{2}$$

We instead analyse the projection of each utterance on principal axes and formulate a score, which we will refer to as the Singular Score going forward. We use this score value to quantize the dataset into buckets and apply appropriate downsampling methods per bucket, giving up to 25% reduction in the data while also showing improvements in the SemER and Intent Classification metric consistently.

From equation 2 we can see that MV, which is the projection of embedding matrix along principal components is the same as $U\Sigma$. This is because U, V are unitary matrices and $VV^T = V^TV = I$. Each row in this matrix $B = MV$ represents the projection of corresponding input embedding along

principal axes. We use absolute sum (i.e L_1 norm, or Manhattan distance from origin) of each row in B as the Singular score corresponding to the data point $u_i \in S$.

$$score_i = \sum_{j=1}^{m} |B_{ij}| \tag{3}$$

We conducted experiments on the representation power of this score and find interesting observations.

Correlation with frequency: As shown in figure 2a we find that there is a correlation between the frequency of an utterance and the score it generates where the lower ranges of scores represent more than 60% of the data. We also performed this experiment on the English and Hindi subset of the Multi-Atis++ data (Xu et al., 2020) to verify our observations. The Multi-Atis++ dataset is an extension of the ATIS (Air Travel Information Services) dataset (Upadhyay et al., 2018) developed to support the research and development of speech understanding systems. This data comprises of 5928 user spoken utterances (4488 English, 1440 Hindi) of which 5621 (94%) utterances are unique. These utterances are based on various hypothetical travel planning scenarios and are obtained by users interacting with a partially or completely automated ATIS system which is then recorded and transcribed. As shown in figure 2b we see that due to the unique utterance composition of Multi-Atis++ Hindi and English subset, the singular score distribution of the graphs remain majorly the same.

Correlation with Sparsity: We also observe that singular scores are a loose indicator of sparsity in the principal axis space as shown in figure 3. A datapoint with higher singular score is observed to have dense representation in its projection along principal axes while an utterance with low singular score has a representation on one or two of the first few axes only. Since the first few principal axes in SVD indicate the spread of the data on those axes, lower singular scores which primarily contain scores in the first few axes belong to those utterances which are common in the data.

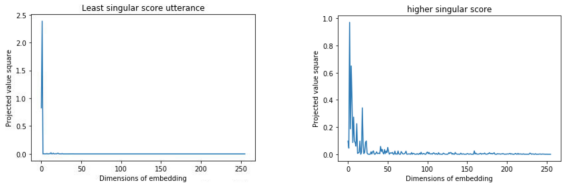


Figure 3: Singular Score Comparison

Compounding these two observations, and based on the pattern we observe in figure 2a we use the Singular Scores to quantize the utterance bins into three and apply different degrees of subsampling to each bucket. For an utterance $u_i \in S$ with frequency n_i and singular score value $score_i$:

- *Head:* Utterances with low singular scores, ($score_i \leq 7.7$) which have a higher degree of repetition. The frequency n_i is reduced to $10 * \log_2(n_i)$
- *Mid:* Utterances with medium singular scores ($7.7 < score_i \leq 10$) has its frequency n_i reduced to $\log_2(n_i)$
- *Tail:* Utterances with high singular scores ($score_i > 10$) which have almost negligible repetitions have their frequency retained.

The amount of reduction achieved by the Unique, Log N, Clustering 70% and Singular Score approaches is summarized in table 1.

4 Experimental Setup

4.1 Data

Since we present approaches with practical applications to real-world SLU modelling systems, we present results on real world data. In particular, use 3 months of data from an internal de-identified data authority and include a random sample from the remainder of the year to account for seasonality in the utterance requests. We use English and Hindi data from five domains, i.e. *Music, Video, Weather, Notifications,* and *Shopping*.

Data statistics are shown in table 3; for each domain, we have atleast 100k training samples of English and Hindi data in equal distribution and use 90% for training and 10% for validation.

	Domain	Unique		Log N		Clustering 70%		Singular Score	
		SemER	IC	SemER	IC	SemER	IC	SemER	IC
English	Music	-6.18	-18.24	-3.77	-14.46	-2.48	-8.83	-3.08	-14.45
	Shopping	-3.52	-3.29	-2.88	-3.84	-4.19	-4.53	-3.79	-4.97
	Video	+9.90	+10.04	+4.40	+4.45	+6.31	+5.70	+3.80	+4.05
	Notifications	-1.10	-0.19	-1.90	-0.62	-1.00	-0.08	-1.15	-2.56
	Weather	-4.40	-9.25	-2.20	-4.29	-5.28	-7.86	-8.05	-17.14
Hindi	Music	-6.60	-24.48	-4.66	-20.74	-2.32	-10.77	-4.51	-20.11
	Shopping	-7.20	-9.71	-4.45	-3.83	-2.23	-3.87	-5.13	-6.63
	Video	+9.54	+10.96	+5.29	+6.62	+1.76	+6.16	-0.00	+6.16
	Notifications	-2.72	-6.21	-2.16	-7.45	-3.47	-13.69	-2.60	-11.20
	Weather	-1.56	-28.21	-0.00	-15.38	-1.12	-25.64	-1.12	-23.08

Table 2: Relative change results

Domain	Intents	Slots
Music	27	103
Shopping	25	45
Video	36	73
Notifications	24	47
Weather	4	18

Table 3: Dataset distribution

4.2 Model parameters

We use an in-house distilled multilingual Lean BERT (Conneau et al., 2020) sized model (50Mparameters) pre-trained on multiple languages including English and Hindi on a large variety of tasks. We use max-pooling for sentence representation. Each of our decoders, i.e. for IC and slot filling components, have one dense layer of size 128 and 256 with relu and gelu activation each respectively. The dropout values used in IC and SF decoders are 0.1. For optimization, we use Adam optimizer with learning rate $10e^{-4}$ with a step scheduler. We trained our model for 15 epochs with batch size of 64 and gradually unfreeze the initial embedding layer (Howard and Ruder, 2018) over 5000 steps.

4.3 Metrics

We evaluate our models on two standard SLU metrics - Intent Classifcation accuracy (IC) and Semantic Error Rate (SemER) following Gaspers et al. (2018), which jointly measures IC and SlotF1 and is defined as

$$SemER = \frac{\#(slot + intent\_errors)}{\#slots\_in\_reference + 1} \quad (4)$$

5 Results & Discussion

For each domain, we build four SLU models trained on the combined English and Hindi data, each named after the data reduction approach applied to the training data fed to it: Baseline, Unique, Log N, Clustering 70%, and Singular Score. We report the performance for each model on SemER and IC accuracy metrics in table 2

We break down our results into three categories: discussion on Video domain degradation, performance analysis of various data reduction technique and performance comparison across metrics.

5.1 Video domain degradation

Video domain consistently sees degradation in SemER metric as compared to the baseline model trained on the complete dataset. This is an indicator that subsampling data is not always beneficial and should be leveraged to make decisions on whether the data slice should be subsampled or not. However, degradation was also observed to be the least in the Singular Score approach, with 0% delta for Hindi SemER and the least IC degradation score. The Video domain training data singular scores captures the essence of frequency and semantic variety in training data which the Unique, Log N and Clustering 70% methods individually could not, furthering concreting our belief in the intuition behind these scores.

5.2 Data reduction techniques

The method of uniquing the input data performs well across languages and metrics as compared to the other approaches. However, this is not practical for commercial SLU systems where the natural distribution of utterance weights is determined by its

frequency of occurrence. Similarly, for the Log N approach, we see consistent improved performance, yet this approach affects tail frequency utterances which are already under represented. We can perform Log N reduction on only the top few most frequent utterances and generate a uniform representation from the long tail distribution of data, but that would be a scaled version of the unique experiment and we expect results to be pretty similar.

An interesting observation we extract from the results table is that for the Singular Score approach, across most domains, most metric values have the behaviour

$$\text{Sing. Score} \leq min(\text{Log N}, \text{Clust. } 70\%)$$

Singular Score method shows combined improvements from Log N and Clustering 70% indicating that the Singular Score approach factors in frequency response as well as semantic similarity in its reduction step. Singular Score can be computationally heavy as it calculates the SVD of the input embedding and will scale exponentially as the input dimension size increases.

5.3 Metrics

Intent Classification benefits from all data reduction techniques across different languages. This indicates that the model has abundance of training data for intent classification given the simpler nature of the task as compared to Slot Filling. In the joint IC/SF BERT model context, we see that intent classification accuracy improves while also showing improvements in SemER indicating no compromise on the Slot Filling task.

6 Conclusion

In this paper, we investigated various inexpensive approaches for identifying data redundancy in training data used to fine-tune BERT based models in SLU systems for the IC and SF task. To the best of our knowledge, this work is the first step in the direction of identifying inexpensive techniques to fine-tune BERT model without affecting offline metrics. We presented empirical results on a real-world SLU dataset, showing that data reduction techniques benefit SemER and Intent Classification metrics. In particular, we proposed a novel data redundancy identification and reduction technique which we call the Singular Score approach. This method helps jointly filter utterances based on their frequency and semantic representation and also helps achieve one of the best results among the techniques experimented with. Future work may target more complex forms of identifying training data redundancy such as influential instances (Pruthi et al., 2020; Koh and Liang, 2020) or active learning Grießhaber et al. (2020)

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning.

Álvar Arnaiz-González, J. Díez-Pastor, Juan José Rodríguez Diez, and C. García-Osorio. 2016. Instance selection of linear complexity for big data. *Knowl. Based Syst.*, 107:83–95.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven pretraining of self-attention networks.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling.

Hyunjin Choi, Judong Kim, Seongho Joe, Seungjai Min, and Youngjune Gwon. 2021. Analyzing zero-shot cross-lingual transfer in supervised nlp tasks.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation.

Aysu Ezen-Can. 2020. A comparison of lstm and bert for small corpus.

George W. Furnas, Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, Richard A. Harshman, Lynn A. Streeter, and Karen E. Lochbaum. 1988. Information retrieval using a singular value decomposition model of latent semantic structure. In *SIGIR*, pages 465–480. ACM.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning.

Judith Gaspers, Penny Karanasou, and Rajen Chatterjee. 2018. Selecting machine-translated data for quick bootstrapping of a natural language understanding system. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 137–144, New Orleans - Louisiana. Association for Computational Linguistics.

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. 2019. Induction networks for few-shot text classification.

Gene H. Golub and Christian Reinsch. 1971. Singular value decomposition and least squares solutions. *Linear Algebra*, pages 134–151.

Daniel Grießhaber, Johannes Maucher, and Ngoc Thang Vu. 2020. Fine-tuning bert for low-resource natural language understanding via active learning.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification.

Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. 2020. Few-shot named entity recognition: A comprehensive study.

W. Kabsch. 1978. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 34(5):827–828.

Gregory R. Koch. 2015. Siamese neural networks for one-shot image recognition.

Pang Wei Koh and Percy Liang. 2020. Understanding black-box predictions via influence functions.

Andrew K. Lampinen and James L. McClelland. 2018. One-shot and few-shot learning of word embeddings.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert?

Garima Pruthi, Frederick Liu, Mukund Sundararajan, and Satyen Kale. 2020. Estimating training data influence by tracing gradient descent.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *NAACL-HLT*.

Eyal Shnarch, Ariel Gera, Alon Halfon, Lena Dankin, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2021. Cluster & tune: Enhance {bert} performance in low resource text classification.

Shyam Upadhyay, Manaal Faruqui, Gökhan Tür, Dilek Z. Hakkani-Tür, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

S. Walton, O. Hassan, and K. Morgan. 2013. Reduced order modelling for unsteady fluid flow using proper orthogonal decomposition and radial basis functions. *Applied Mathematical Modelling*, 37(20-21):8930–8945.

D. Wilson and T. Martinez. 2004. Reduction techniques for instance-based learning algorithms. *Machine Learning*, 38:257–286.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual nlu.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning.

Wenpeng Yin. 2020. Meta-learning for few-shot natural language processing: A survey.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021. Revisiting few-sample bert fine-tuning.