Multi-task pre-finetuning for zero-shot cross lingual transfer

Moukthika Yerramilli[*]
Amazon / Bangalore
mky@amazon.com

Pritam Varma[*]
Amazon / Bangalore
spv@amazon.com

Anurag Dwarakanath[*]
Amazon / Bangalore
adwaraka@amazon.com

Abstract

Building machine learning models for low resource languages is extremely challenging due to the lack of available training data (either un-annotated or annotated). To support such scenarios, zero-shot cross lingual transfer is used where the machine learning model is trained on a resource rich language and is directly tested on the resource poor language. In this paper, we present a technique which improves the performance of zero-shot cross lingual transfer. Our method performs multi-task pre-finetuning on a resource rich language using a multilingual pre-trained model. The pre-finetuned model is then tested in a zero-shot manner on the resource poor languages. We test the performance of our method on 8 languages and for two tasks, namely, Intent Classification (IC) & Named Entity Recognition (NER) using the MultiAtis++ dataset. The results show that our method improves IC performance in 7 out of 8 languages and NER performance in 4 languages. Our method also leads to faster convergence during finetuning. The usage of pre-finetuning demonstrates a data efficient way for supporting new languages and geographies across the world.

1 Introduction

Recent advances in Natural Language Processing include the development of language models trained in an unsupervised fashion on large amounts of data (Devlin et al., 2019), (Radford and Sutskever, 2018). These models were extended to a multilingual context by training on data from a number of languages (Devlin et al., 2019), (Conneau et al., 2019). These multilingual models were found to perform well in a zero-shot cross lingual transfer tasks - i.e. the multilingual model is fine-tuned for a task in a resource rich language (such as English) and is directly tested on a resource poor language (such as Swahili) (Schlinger, 2019a)

In this paper, we investigate the usage of multi-task pre-finetuning in zero-shot cross lingual tasks. Pre-finetuning is a step between pre-training and fine-tuning, where a pre-trained model is trained on additional supervised learning tasks with the aim to improve pre-trained representations (Aghajanyan et al., 2021)(Liu et al., 2019). Typically, these additional tasks are unrelated to the tasks that the model will finally be fine-tuned on. Past work on multi-task learning showed its usefulness in a monolingual setting. In our work, we extend the pre-finetuning concept to a multilingual setting with zero-shot cross lingual transfer.

Our method is demonstrated on the XLM-Roberta (Conneau et al., 2019) pre-trained model and uses 8 additional auxiliary tasks from the GLUE benchmark (Wang et al., 2018) for the multi-task pre-finetuning on English. The resulting model is then applied for the joint Intent-Classification & Named Entity Recognition tasks (IC-NER). We show cross lingual transfer by fine-tuning for IC-NER on English and directly test on 8 different languages in a zero-shot manner. We use the MultiAtis++ dataset (Xu et al., 2020) for the IC-NER data in English and 8 other languages for fine-tuning and testing. Our results bring out multiple insights. We find that the multi-task pre-finetuned model is better by 5.12% relative (or 391 absolute basis points (bps)) on average across all 8 languages for the IC task at early stages of fine-tuning. This shows the ability of multitask pre-finetuning to improve the learnt representations of the pre-trained model. The results however indicate that such out of the box improvement in pre-trained models is not seen in the NER task, where the performance at early stages of fine-tuning degrades by 10% relative (or 525 bps absolute). We also find that pre-finetuning has improved the performance

[*]Equal contribution

Proceedings of the 18th International Conference on Natural Language Processing, pages 474–480
Silchar, India. December 16 - 19, 2021. ©2021 NLP Association of India (NLPAI)

in monolingual setting where the results for both IC and NER for English have improved by 6.53% relative (or 549 bps absolute) and 11.86% relative (or 700 bps absolute) respectively.

Improving the ability of cross lingual transfer in a zero shot setting has large implications. Through our method, products using language models can improve machine learning support for new languages where sufficient training data is not available. For example, the Indian sub-continent has 179 languages and 544 dialects (Wikipedia contributors, 2021b) and building machine learning services to cater to all multilingual users is a daunting task since enough data, both in the form of annotated and un-annotated, is not available.

This paper is structured as follows. In Section 2, we present the related work in zero-shot cross lingual transfer and progress in multi-task pre-finetuning. Section 3 presents our method of improving zero-shot cross lingual transfer through the use of multi-task pre-finetuning. We present the results in Section 4 and conclude with directions for future work in Section 5.

2 Related work

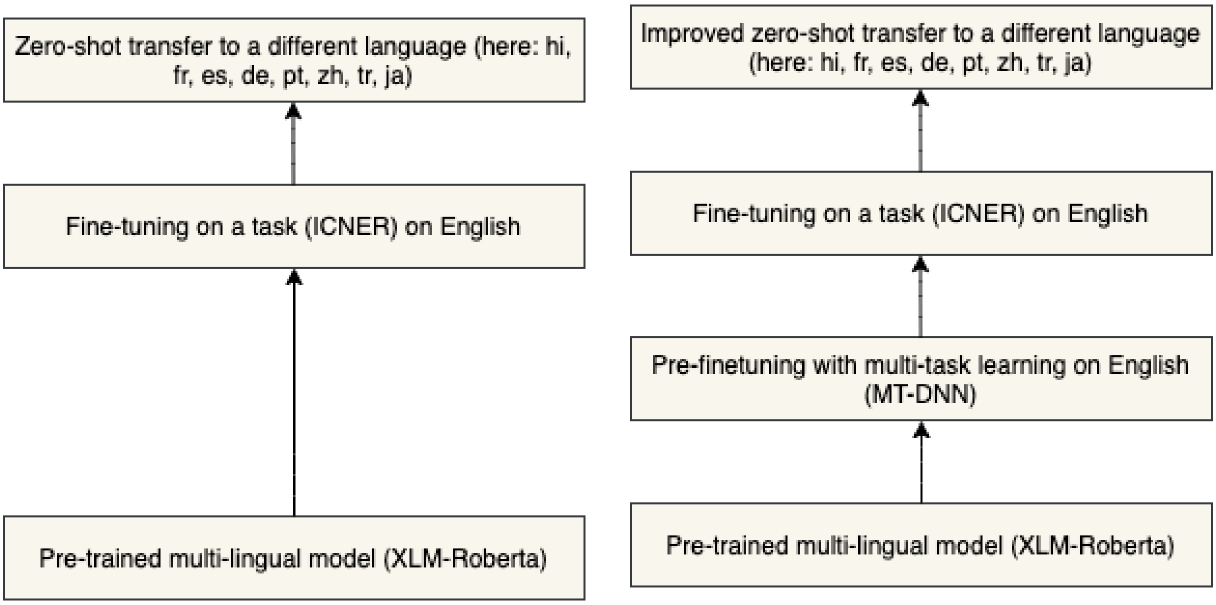
Multilingual language models such as mBERT (Devlin et al., 2019)(Devlin et al., 2021), XLM (Conneau et al., 2019) and MuRIL (Khanuja et al., 2021) have advanced the state-of-the-art on cross-lingual natural language understanding tasks by training large Transformer models (Vaswani et al., 2017) on data from many languages.

These multilingual pre-trained models can be used as generic task agnostic neural network architectures and can be applied in different natural language processing tasks by attaching a task specific decoder (such as a linear classifier) and fine-tuning on the task specific training data. The usage of multilingual pre-trained models also enabled zero-shot cross lingual transfer. In zero-shot transfer, the pre-trained model is fine tuned on annotated data from a resource rich language (such as English) and directly tested in a resource poor language (such as Swahili). Studies (Schlinger, 2019b) (Libovický et al., 2020a) (Wu and Dredze, 2019a) have shown such multilingual pre-trained models have the ability to learn common representations across languages even though the training methodology was not explicitly designed to build common representations.

Independently, approaches have been developed to improve the performance of pre-trained models in the monolingual setting. Prominent work includes the usage of multi-task pre-finetuning (Liu et al., 2020) (Aghajanyan et al., 2021). In pre-finetuning, the pre-trained model such as BERT is trained in a supervised learning setting on auxiliary tasks. The work in (Liu et al., 2020) uses 8 different tasks from the GLUE dataset (Wang et al., 2018) and trains for different tasks in random batches. The work in (Aghajanyan et al., 2021) scales the pre-finetuning concept by training on 50 different tasks. Their results show that multi-task pre-finetuning can significantly improve the performance of the pre-trained models. Both these works test the improvement in a monolingual setting. In contrast, our work in this paper explores the applicability of pre-finetuning in a multilingual setting.

3 Multi-task pre-finetuning for zero shot cross lingual transfer

In this section, we present our method to improve the performance of zero-shot transfer of knowledge across languages. Current state-of-the-art methods have demonstrated the great ability of a neural network to transfer knowledge of a given task across languages using a pre-trained model that is trained generically (Conneau et al., 2018), (Schlinger, 2019a), (Wu and Dredze, 2019b). In our work, we demonstrate a further intriguing property of deep neural networks. We take a multi-lingual pre-trained neural network and train it over multiple tasks in English using task specific supervised data (we choose English since there is significant amount of supervised training data available for the language). We then fine tune the network for the specific down-stream task in English. We denote the resulting neural network model through our method as **MT-DNN-MultiLingual-Finetuned**. Now, when the **MT-DNN-MultiLingual-Finetuned** model is tested directly (i.e. zero-shot) on the downstream task in a different language, we observe better performance. This demonstrates the ability of the neural network to not only transfer knowledge of the same task across languages (as demonstrated by existing literature), but also shows its ability to transfer cumulative knowledge of multiple different tasks across languages. Figure 1 depicts the novelty of our work in comparison with extant literature.


Zero-shot transfer to a different language (here: hi, fr, es, de, pt, zh, tr, ja)
Fine-tuning on a task (ICNER) on English
Pre-trained multi-lingual model (XLM-Roberta)
Improved zero-shot transfer to a different language (here: hi, fr, es, de, pt, zh, tr, ja)
Fine-tuning on a task (ICNER) on English
Pre-finetuning with multi-task learning on English (MT-DNN)
Pre-trained multi-lingual model (XLM-Roberta)

(a) Baseline model: XLM-R with fine-tuning on English

(b) Experiment model: XLM-R with multi-task pre-finetuning and fine-tuning – both done on Enlgish

Figure 1: Baseline vs Experimental Model with multi-task pre-finetuning setup for zero-shot transfer.

3.1 Multi-Lingual Pre-trained Model

We base our method on the XLM-RoBERTa base (XLM-R) (Conneau et al., 2019) as the pre-trained model for all the experiments in this paper. XLM-R is trained on 100 languages (including the 8 languages we use for our tests) using the CommonCrawl Corpus. The XLM-R was trained using the Masked Language Model training objective and does not use any supervised data (or parallel corpus) for its training. Recent work (Schlinger, 2019a) (Libovický et al., 2020b) has shown that the XLM-R model is able to build similar representations for semantically similar sentences across languages.

3.2 The Multi-Task Pre-finetuning step

In existing work, the XLM-R model would be fine-tuned on a downstream task. In contrast, our method introduces a pre-finetuning step before fine-tuning is done. We chose the GLUE benchmark dataset (Wang et al., 2018) which contains supervised training data in English for 8 tasks spanning - 1) Single-Sentence Classification; 2) Pairwise Text Similarity; 3) Pairwise Text Classification; and 4) Pairwise Ranking. For each task, a task specific decoder was attached to the XLM-R pre-trained model and trained for a fixed number of epochs. We made use of the publicly available framework (Liu et al., 2020) for pre-finetuning on the GLUE benchmark. No layer freezing was done for the pre-finetuning steps. The different tasks and the size of the training data is shown in Table 1.

3.3 The finetuning step

Post the pre-finetuning on English using the GLUE benchmark dataset, we attach a decoder for the joint Intent Classification & Named Entity Recognition (IC-NER) task. The formulation of the IC-NER task is shown below.

$$y^i = softmax(W^i h_1 + b) \qquad (1)$$

where y^i represents the IC hypothesis and h_1 represents the hidden state of the classification head (also denoted using the special token [CLS]).

For NER, we feed the final hidden states of other tokens $h_2, ..., h_n$ into a softmax layer to classify over the slot filling labels. To make this procedure compatible with WordPiece or SentencePiece tokenization, we feed each tokenized input word into a tokenizer and use the hidden state corresponding to the first sub-token as input to the softmax classifier.

$$y_n^s = softmax(W^s h_n + b), n \in 1...N \qquad (2)$$

where h_n is the hidden state corresponding to the first sub-token of word x_n and s is the slot label. To jointly model intent classification and slot filling, the objective is formulated as:

$$p(y^i, y^s|x) = p(y^i|x) \prod_{n=1}^{N} p(y_n^s|x), \qquad (3)$$

Table 1: GLUE dev set results. ST-DNN has the same architecture as MT-DNN-Original but without the Multi-Task pre-finetuning. MT-DNN-MultiLingual is the Multi-Task pre-finetuned model with XLM-R pre-trained model as the base. The results in columns 3, 4 and 5 (BERT-Large, ST-DNN and MT-DNN-Original) are obtained from (Liu et al., 2019) The results in bold(not indicative of best performance) represent the performance of MT-DNN-MultiLingual, which is the pre-finetuned model used in experiments later.

.

Dataset	Train Data Size	BERT-Large	ST-DNN	MT-DNN-Original	MT-DNN-MultiLingual
QQP(F1score/Acc)	364k	86.3/86.2	91.3/88.4	91.9/89.2	**88.532/91.373**
MNLI-m/mm(Acc)	393k	86.3/86.2	86.6/86.3	87.1/86.7	**84.035/84.123**
RTE(Acc)	2.5k	71.1	72	83.4	**77.978**
QNLI(Acc)	108k	92.4	-	92.9	**90.427**
MRPC(F1/Acc)	3.7k	89.5/85.8	89.7/86.4	91.0/87.5	**92.199/89.216**
SST-2(Acc)	67k	93.5	-	94.3	**92.775**
CoLA(Mcc)	8.5k	61.8	-	63.5	**49.217**
STS-B(Pc/Sc)	7k	89.6/89.3	-	90.7/90.6	**89.086/88.868**

The learning objective is to maximize the conditional probability $p(y^i, y^s|x)$. The model is fine-tuned end-to-end via minimizing the cross-entropy loss.

3.4 Zero-shot transfer evaluation task

We evaluate the performance of pre-finetuned and finetuned XLM-R model (on English) directly on different languages for the joint IC-NER task. We use 8 languages in the MultiATIS++ corpus - Hindi (hi), French (fr), Spanish (es), German (de), Portuguese (pt), Chinese (zh), Japanese (ja) and Turkish (tr). These languages also belong to a diverse set of language families - Indo-European, Sino-Tibetan, Japonic and Altaic.

4 Experiments and Results

4.1 Pre-Finetuning

Using the XLM-R pre-trained model, pre-finetuning is conducted over eight English language GLUE datasets: CoLA(Single-Sentence Classification), SST-2(Single-Sentence Classification), STS-B (Pairwise Text Similarity), RTE(Pairwise Text Classification), MNLI(Pairwise Text Classification), QQP(Pairwise Text Classification), MRPC(Pairwise Text Classification), QNLI(Pairwise Ranking). Pre-finetuning is carried out using MT-DNN (Multi-task DNN) setup where the training is done on only English. The training is conducted for four epochs with standard (as per the original MT-DNN implementation(Liu et al., 2020)) hyper-parameter values such as learning rate of $5 \times e^{-5}$, batch-size of 32 and with adamax optimizer.

Table 1 shows the GLUE dev set results, where MT-DNN-MultiLingual is the proposed model with pre-finetuning over XLM-R. The MT-DNN-Original is the original MT-DNN model that is pre-finetuned on BERT-Large model. The results indicate that MT-DNN-MultiLingual, which uses a multi-lingual pre-trained model as its base, is able to beat the mono-lingual non-prefinetuned models such as ST-DNN (stands for single task DNN which implements task-wise finetuning) and BERT-Large in most validation datasets (except for CoLA, MNLI-m/mm). However, its unable to beat MT-DNN-Original which is pre-trained and pre-finetuned exclusively on English. These results re-emphasize the effectiveness of multilingual models even for mono-lingual tasks.

4.2 IC-NER fine-tuning and Zero-Shot Transfer

The baseline **XLM-R-Finetuned** model for zero-shot transfer experiments consists of XLM-R model that is fine-tuned on English and evaluated on other languages in a zero-shot manner. These experiments are conducted on IC-NER Joint Task (Chen et al., 2019) over MutiATIS++ dataset. Our method **MT-DNN-MultiLingual-Finetuned** consists of MT-DNN-MultiLingual that is fine-tuned on English. Tables 2, 3 and 4 show the zero-shot performance of baseline model vs **MT-DNN-MultiLingual-Finetuned** across all the available languages in MultiAtis++ dataset. The results are averaged across three different fine-tuning runs and hyper-parameters such as batch-size (256), learning rate $(5 \times e^{-6})$ remain the same across both baseline and experimental model.

As seen in the results, **MT-DNN-MultiLingual-Finetuned** beats the baseline on 7 out of 8 languages at the 10th epoch for the IC task. The average improvement in accuracy is 5.12% con-

Table 2: XLM-R-finetuned vs MT-DNN-MultiLingual-finetuned for English, Hindi and French

Language		en		hi		fr	
Task		IC (Accuracy)	NER (F1 score)	IC (Accuracy)	NER (F1 score)	IC (Accuracy)	NER (F1 score)
XLM-R-finetuned	10th Epoch	78.52	59	72.57	55	72.14	46
	20th Epoch	86.52	81	78.4	58	80.82	55
	30th Epoch	90.06	83	80.68	54	82.3	59
	40th Epoch	90.52	83	81.37	55	84.58	59
MT-DNN-MultiLingual-finetuned	10th Epoch	84.01	66	77.71	53	80.36	49
	20th Epoch	90.41	82	83.77	55	86.64	57
	30th Epoch	90.75	83	84.68	54	86.64	58
	40th Epoch	90.75	84	84.57	53	86.3	58

Table 3: XLM-R-finetuned vs MT-DNN-MultiLingual-finetuned for Spanish, German and Portuguese

Language		es		de		pt	
Task		IC (Accuracy)	NER (F1 score)	IC (Accuracy)	NER (F1 score)	IC (Accuracy)	NER (F1 score)
XLM-R-finetuned	10th Epoch	72.34	47	72.34	47	73.6	54
	20th Epoch	81.05	69	80.22	53	80.57	62
	30th Epoch	84.7	70	84.68	57	86.17	63
	40th Epoch	84.81	70	88.68	58	87.88	64
MT-DNN-MultiLingual-finetuned	10th Epoch	75.68	51	83.08	48	74.62	54
	20th Epoch	87.1	69	88.57	58	82.62	62
	30th Epoch	87.67	72	88.91	59	83.42	63
	40th Epoch	87.44	72	89.48	59	83.2	63

Table 4: XLM-R-finetuned vs MT-DNN-MultiLingual-finetuned for Chinese, Japanese and Turkish

Language		zh		ja		tr	
Task		IC (Accuracy)	NER (F1 score)	IC (Accuracy)	NER (F1 score)	IC (Accuracy)	NER (F1 score)
XLM-R-finetuned	10th Epoch	72.14	61	72.03	52	73.6	54
	20th Epoch	75.57	65	73.07	56	68.42	51
	30th Epoch	79.68	67	77.33	55	72.68	52
	40th Epoch	79.79	67	78.82	55	74.82	51
MT-DNN-MultiLingual-finetuned	10th Epoch	75.57	64	73.53	33	71.55	22
	20th Epoch	85.5	66	84.11	43	71.4	27
	30th Epoch	85.38	67	85.5	45	68.7	27
	40th Epoch	85.38	0.67	85.5	45	66.7	27

sidering all languages. The results indicate Turkish (tr) is showing regressions for both IC and NER and appears to be an outlier. Discounting the Turkish language, we see an average improvement of 6.11% for IC. A similar improvement of 6.54% in IC is seen on English. These results, at the 10^{th} epoch, indicate that multi-task pre-finetuning has improved the performance of the pre-trained model for the IC task in a zero-shot setting. At the 40^{th} epoch, we see that the performance improvement tapers and achieves an average improvement of 1.17% across all languages. This reduction in gain is expected as the continual training on English starts to improve baseline performance but reduces some of the gains across other languages - i.e. the usage of pre-finetuning allows for early convergence (convergence in terms of performance on non-English languages). Discounting the results from Turkish, we see an average improvement of 2.64% across 7 languages for the IC task at the 40^{th} epoch.

The early convergence of our method can be seen in figure 2. We observe that MT-DNN-MultiLingual-Finetuned converges faster (at epoch 20) than the baseline method. We also see that in Chinese, the baseline model appears to have a constant test accuracy value till 15 epochs and the accuracy starts to increase post that. In contrast, MT-DNN-MultiLingual-Finetuned accuracy starts to improve after 8 epochs. Since MT-DNN-MultiLingual-Finetuned gains such early momentum on IC task, its able to progressively beat the baseline performance. Similar early gains in accuracy are observed for Hindi and Japanese as well. This can be attributed to the improved generalisation via the pre-finetuning step.

We see a degradation of 5.32% and 12.17% for pt (Portuguese) and tr (Turkish) respectively on IC task (at 40th epoch). For Portuguese, we see that MT-DNN-MultiLingual-Finetuned beats the baseline until 25 epochs. However, fine-tuning for further epochs shows better gains in baseline model compared to MT-DNN-MultiLingual-Finetuned. Although the MT-DNN-MultiLingual-Finetuned beats the baseline for zero-shot performance in Turkish at 20th Epoch, the performance does not

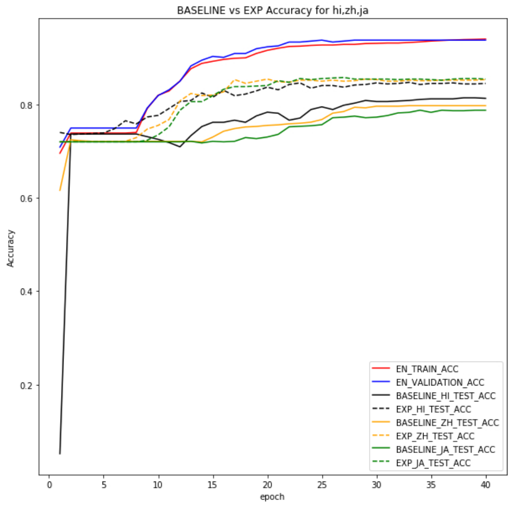


Figure 2: Learning profiles of Train (en), Validation (en) and Test (hi,zh,ja) datasets for Baseline and Experimental Models

improve with further finetuning as compared to baseline.

The results from NER indicate that pre-finetuning hasn't improved the performance. On average across all 8 languages, NER performance at the 10^{th} epoch has degraded by 10.09% relative to baseline. For NER, we see significant degradation of Turkish (- 47%) (- indicates degradation). Discounting Turkish, we see NER performance averaging across 7 languages to be -2.72% relative to baseline.

The reasons for degradation in Turkish (for both IC & NER) and for the lack of improvement in NER is not exactly clear. We hypothesise that the degradation in Turkish could perhaps be attributed to the fact that the language is highly agglutinative in nature. Agglutination is a linguistic process of derivational morphology in which complex words are formed by stringing together morphemes without changing them in spelling or phonetics (Wikipedia contributors, 2021a). While Japanese and Hindi do show partial agglutination, the morphemes/words are much more complex in Turkish. This hypothesis needs to be further investigated by checking for zero-shot performance on other agglutinative languages such as Hungarian, languages of the Dravidian family etc and the investigation forms the part of our future work.

5 Conclusion and Future Directions

In this work, we have investigated the effectiveness of multi-task pre-finetuning for cross lingual zero-shot transfer. Our method takes a multilingual pre-trained model and further trains it on auxiliary supervised tasks. The pre-finetuned model is then finetuned on a task specific language and tested directly on other languages in a zero-shot setting. We test our method for the tasks of Intent Classification (IC) and Named Entity Recognition (NER). The results indicate that the method indeed improves the performance for the IC task. This improvement is seen the most in early steps of finetuning and our method allows the training to converge faster. However, we see that the pre-finetuning does not improve results for NER. Further, we see that both IC and NER results degrade in Turkish.

Our furture directions include scaling our method to cover large number of auxiliary tasks for pre-finetuning. While our current method used 8 auxuliary tasks, we aim to scale this to beyond 50. Large scale multi-task learning has been shown to be effective in a monolingual setting (Aghajanyan et al., 2021) and we would like to explore this phenomenon in a multilingual setting. We will also explore the role of language families and its interaction with multi-task learning to test the hypothesis of poor performance in agglutinative languages (such as Turkish).

The pratical application of zero-shot learning provides a data-efficient method to expand the language capability of machine learning based techniques. The results from our technique show that such zero-shot performance can be further improved and also provide impetus for further research.

References

A. Aghajanyan, A. Gupta, A. Shrivastava, X. Chen, L. Zettlemoyer, and S. Gupta. Muppet: Massive multi-task representations with pre-finetuning. *CoRR*, abs/2101.11038, 2021. URL https://arxiv.org/abs/2101.11038.

Q. Chen, Z. Zhuo, and W. Wang. BERT for joint intent classification and slot filling. *CoRR*, abs/1902.10909, 2019. URL http://arxiv.org/abs/1902.10909.

A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://www.aclweb.org/anthology/D18-1269.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116.

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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Multilingual BERT. https://github.com/google-research/bert/blob/master/multilingual.md, 2021. [Online; accessed 27-May-2021].

S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, S. C. B. Gali, V. Subramanian, and P. Talukdar. Muril: Multilingual representations for indian languages. *CoRR*, abs/2103.10730, 2021. URL https://arxiv.org/abs/2103.10730.

J. Libovický, R. Rosa, and A. Fraser. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online, Nov. 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.150. URL https://www.aclweb.org/anthology/2020.findings-emnlp.150.

J. Libovický, R. Rosa, and A. Fraser. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online, Nov. 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.150. URL https://www.aclweb.org/anthology/2020.findings-emnlp.150.

X. Liu, P. He, W. Chen, and J. Gao. Multi-task deep neural networks for natural language understanding. *CoRR*, abs/1901.11504, 2019. URL http://arxiv.org/abs/1901.11504.

X. Liu, Y. Wang, J. Ji, H. Cheng, X. Zhu, E. Awa, P. He, W. Chen, H. Poon, G. Cao, and J. Gao. The microsoft toolkit of multi-task deep neural networks for natural language understanding. *CoRR*, abs/2002.07972, 2020. URL https://arxiv.org/abs/2002.07972.

A. Radford and I. Sutskever. Improving language understanding by generative pre-training. 2018. URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

E. Schlinger. How multilingual is Multilingual BERT ? pages 4996–5001, 2019a.

E. Schlinger. How multilingual is Multilingual BERT ? pages 4996–5001, 2019b.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461, 2018. URL http://arxiv.org/abs/1804.07461.

Wikipedia contributors. Agglutination — Wikipedia, the free encyclopedia, 2021a. URL https://en.wikipedia.org/w/index.php?title=Agglutination&oldid=1029218007. [Online; accessed 28-June-2021].

Wikipedia contributors. Languages of india — Wikipedia, the free encyclopedia, 2021b. URL https://en.wikipedia.org/w/index.php?title=Languages_of_India&oldid=1030136775. [Online; accessed 26-June-2021].

S. Wu and M. Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China, Nov. 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1077. URL https://www.aclweb.org/anthology/D19-1077.

S. Wu and M. Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China, Nov. 2019b. Association for Computational Linguistics. doi: 10.18653/v1/D19-1077. URL https://www.aclweb.org/anthology/D19-1077.

W. Xu, B. Haider, and S. Mansour. End-to-end slot alignment and recognition for cross-lingual nlu. *arXiv preprint arXiv:2004.14353*, 2020.