*Zijian Feng  
Abhash Sapkota  
Michael Rimboim*

1050 Kemper Hall

Davis, CA 95616

June 2nd, 2022

Dear Professor Oliver,

Our recommendation report is entitled "Improvement of Natural Language Processing on Low Resource Languages Using Combined Methods". It is a result of our research combining various solutions to overcome the obstacles of natural language process performance in low-resource languages.

We appreciate your mentorship and guidance throughout the quarter. And we are looking forward to hearing feedback on this recommendation report.

Sincerely,

Zijian Feng

Abhash Sapkota

Michael Rimboim

**Enclosure: Recommendation Report**

Improvement of Natural Language Processing on Low Resource Languages Using Combined Methods

A Recommendation Report

Zijian Feng

Michael Rimboim

Abhash Sapkota

UWP 102E

University of California, Davis

June 2, 2022

# **Table of Contents**

Executive Summary

Abstract

Introduction

Technical Background

Criteria

Comparison

Recommendation

References

# **Executive Summary**

Natural language processing (NLP) has advanced significantly in the past few decades. Innovations and advancements in both NLP algorithms and hardware had allowed wide commercial usage of NLP services such as translation and speech recognition services. However, the performance of NLP on low resource language (LRL) is still insufficient due to multiple reasons. LRLs generally lack sufficient datasets for training models. Variations in their dialects also introduce uncertainty for NLP tasks. Research groups rarely work dedicatedly on LRLs because of their low popularity and conventional methods’ requirement for additional human labor to create labeled datasets. Improvement of NLP performance on LRLs will need solutions that overcome the scarcity of datasets while maintaining the preexisting MT standards for high-resource languages (HRL).

Various groups have been researching and developing new methods to build reliable datasets without committing large amounts of human resources, or new training models for improving neural machine translation (NMT) performances, and models that can still perform with insufficient datasets. First, to solve the problem of dataset scarcity, a variation of bilingual word embeddings (BWE), called BiLex, is utilized to create precise datasets that can fulfill the need for NLP training models for LRLs. Then, the generated datasets are used by multilingual training models which already are proven to be reliable for maintaining high performance. Finally, when datasets are extremely insufficient, we utilize adversarial learning methods to maintain the performance at an acceptable standard.

# 

# **Abstract**

This recommendation’s objective is to suggest a feasible framework with combined methods that overcome the obstacles caused by the nature of low-resource languages to improve the performance of NLP in these languages. These methods include: a model that generates a precise dataset for LRL using a high resource language (HRL) while requiring minimum human intervention, a multilingual model which effectively transfers learned representations from clusters of similar languages and can increase performance by training with a more precise dataset such as the one generated in previous part, and finally a training schema that utilizes adversarial learning for application when no dataset is available.

# **Introduction**

Natural language processing technology (NLP) has been advancing in the past years. According to [1], we are now in the fourth era of NLP. In this era, the use of probabilistic and data-driven models became standard for most NLP tasks. Accordingly, algorithms of parsing, reference resolution, and other components of NLP are all incorporating probabilities and employ evaluation strategies. At the same time, technological innovations in hardware, such as an increase in usable memories, allowed commercial exploitation of NLP services.

However, performances of NLP tasks on low-resource languages (LRL) are still insufficient due to the challenging nature of LRLs. LRLs can be understood as languages that are less studied, resource-scarce, less computerized, less privileged, less commonly taught, and has lower density [3]. All of them introduce challenges to research groups that try to dedicate work on LRLs. Addressing these challenges becomes the key point to improving the performance of NLP on LRLs.

There are many motivations for us to care about LRLs and how to improve the performance of NLP on them. For example, according to [2], Africa and India host more than 2000 LRLs. They are home to over 2.5 billion inhabitants. Developing technologies for languages used by such a large population introduces significant economic perspectives. Moreover, supporting LRLs with NLP tools not only can prevent their extinction and encourage their expansion, but can also help discover knowledge hidden within the original works.

Therefore, in this recommendation, we will evaluate current research and compare different approaches for overcoming different challenges of NLP on LRLs. We will then explain the reasons for our choice of approaches. The technical background section will contain three major challenges and their according solutions: 1. Using BiLex method for dataset generation to overcome the scarcity of datasets. 2. Using a multilingual training model to maintain high NLP standards. 3. Using Adversarial learning to maintain acceptable performance when the access to datasets is extremely limited. The criteria section lists requirements that NLP needs to meet. The comparison section provides a comparison of methods introduced in the technical background section. Finally, the recommendation section presents the final recommendation based on our conducted research.

# **Technical Background**

### **Mitigating Data Scarcity with BiLex**

**Why supervised learning is necessary.**

LRLs lack both annotated, and unlabeled datasets. Annotated datasets are necessary for supervised learning. It requires frequent human interventions, collection of training examples, and manually labeling them. LRLs infrequent use makes this an expensive and infeasible task. Unlabeled datasets are the precursors to their annotated versions and usually consist of raw texts that are sometimes used for unsupervised training. Because unsupervised learning does not precisely sort data, the output data is less accurate and the spectral classes do not always correspond to informational classes. So far, no major application or model can serve as proof of effectiveness for unsupervised training for NPL. LRL’s lack of corpora exacerbates the situation for ML training.

**Bilingual Word Embedding and BiLex**

A common approach is looking for alternative resources of datasets for LRL processing. Bilingual word embeddings (BWE) are the most utilized methods of supervised training for LRL. Word embedding is a representation of words for text analysis where words are placed in vectors encoded with values that represent the meaning [8]. As an extension, bilingual word embeddings go further to learn the semantic relationship of word pairs across languages [9].

[4] proposes an alternative training method to decrease the BWE learning's dependency on human interventions. They use "Bilingual Word Embeddings Based on Lexical Definitions(BiLex) that leverages publicly available lexical definitions for bilingual word embedding learning."

Although various approaches to learning BWE already exist, most of them rely on aligned corpora. Such corpora can provide word-level mappings between two languages [10]. These methods commonly suffer from several deficiencies. First, lexicon seeds require extensive human resources to obtain which often hinders these seed-lexicon-based methods from performing well on LRLs. Second, parallel corpora provides coarse alignment that does not often accurately infer fine-grained semantics transfers of lexicons [11].

The Bilex model’s mechanism has two stages. In stage one, it first finds pairs of words with cross-lingual definitions and extracts bilingual strong word pairs from them. Then the model automatically utilizes induced word pairs for the induction process.

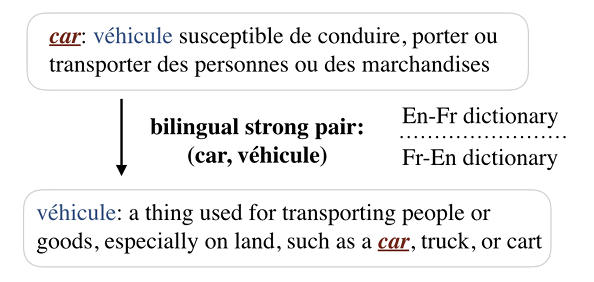
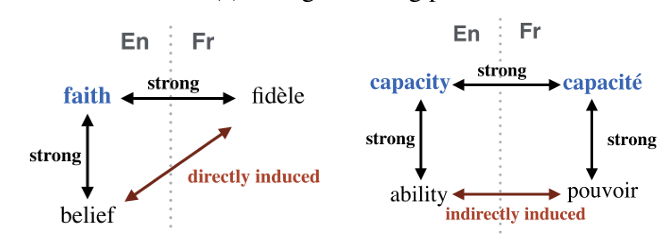


Figure 1: Bilingual strong pair [10]

 Figure 2: Directly and indirectly induced pair [10]

Figures 1, and 2 demonstrate the difference between bilingual strong pairs and induced word pairs.

Because bilingual lexical definitions cover only a limited number of words, [4]’s incorporation of bilingual strong pairs and induced word pairs, and this method resulted in a promising performance.

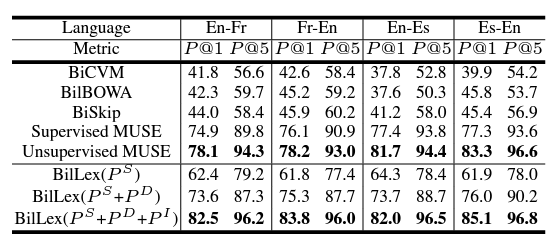
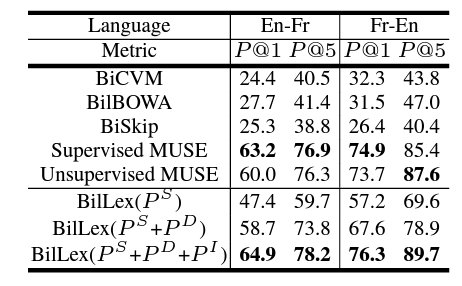
Table 1: Results of word translation task

Table 2: Results of sentence translation task



Tables 1 and 2 show a performance comparison between variants of BiLex and existing models. BiLex utilizes all three types of word pairs and consistently outperforms other methods in all settings.

### **Transfer Learning Using Multilingual Models**

Multilingual models have recently proven to be a successor to the currently used bilingual models. Multilingual models take a multi-task approach to get more use out of transfer learning [12]. Rather than one domain or task, the model trains on several domains. The multilingual model aims to solve the problem of Neural Machine Translation (NMT) quality being low for low-resource languages. Breakthroughs in architecture, optimization, and scalability have made multilingual models the top choice for HRLs and LRLs.

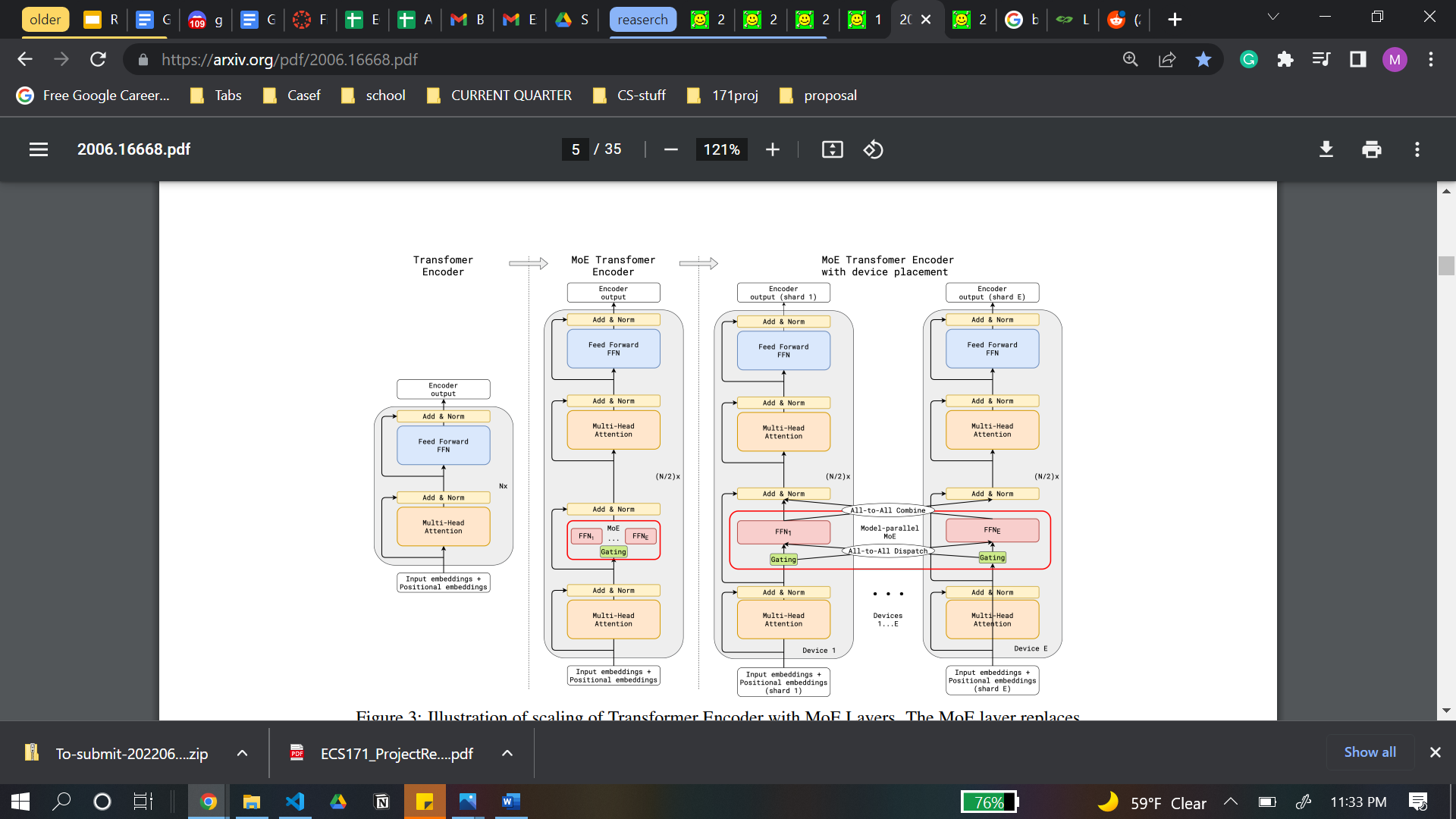
**Lack of transfer learning in bilingual models**

[13] creates both baseline and finetuned bilingual models using the transformer architecture [28]. Bilingual models are a gold standard for HRL with extensive corpora. Lack of corpora prevents bilingual models from achieving reliable quality for LRL [25]. Transfer learning is the most promising approach to achieving quality translations for LRL [26], yet bilingual models cannot take advantage of transfer learning..

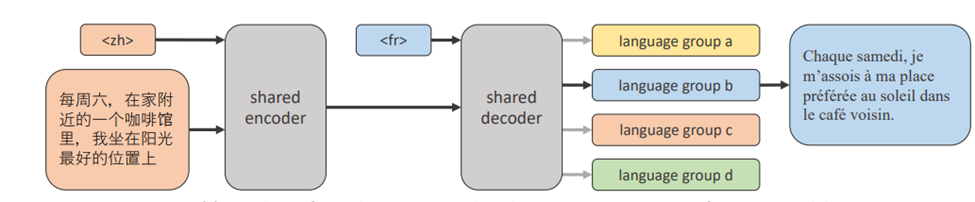
Transfer learning is the ability to train a model for a general task and retrain for a specific or different task reusing the pre-trained weights as starting points. The process transfers some general understanding of words or sentence structure to the next task. Many LRLs have some roots and similarities to HRLs, allowing us to train on large HRL corpora and then finetune for LRL. Using transfer learning, we achieve better understanding without needing extra corpora. Bilingual models only take two languages for each direction, meaning they have no way to use transfer learning to their advantage[24]. Multilingual models are composed of many languages, both HRL and LRL, allowing for transfer learning.

**Multilingual scalability**

As the number of languages grows, so does the positive transfer for LRL [29]. When a model increases the number of languages, it needs to increase model capacity to prevent a drop in quality for HRLs. The two ways to increase capacity in a model are densely and sparsely [13]. Dense multilingual models offer excellent performance but do not scale well as the number of languages increases. As n(the number of languages) increases, the computational cost increases exponentially for a dense multilingual model[29]. The second method of sparse matrices turns out to be a valuable method for increasing capacity.

Figure 7: Original transformer architecture vs. distributed expert based transformer[29] 

A transformer consists of self-attention layers and feed-forward network(FFN) layers. Usually, we scale a transformer by densely adding more layers, but this is not feasible beyond a certain point. Instead, as figure 7 shows, every other FFN layer can be replaced with a mixture of experts (MoE) layer[29]. The replacement allows subnetworks to be activated as needed. The size of a subnetwork is roughly consistent even as the number of experts per MoE layer increases, allowing for sub-linear computational scaling as the number of languages increases[29]. In addition, the rightmost part of figure 7 displays the MoE layer's ability to work on a distributed system, training across several devices simultaneously. A distributed approach is a massive leg-up for reducing training time[25]. If the model's size increases for LRL and HRL in the same manner, it may cause something known as negative transfer, in which HRL takes a hit in performance. By creating subcomponents of language groups that activate as needed, we can direct resources in a way that maximizes positive transfer and minimizes negative transfer. Varying the size of these sub-components to match the extensiveness of corpora is the mechanism by which this works. The combination of non-exponential scaling, architecture that targets positive transfer, and distributed system readiness means the multilingual models with MoE layers are exceptionally scalable.

Figure 8: Multilingual Model with sub-networks for language groups[27]

Translating several languages would require several bilingual models, each with a constant training time and storage requirements[12]. We would need to store n models for n languages which take up a lot of storage and are hard to manage after training. With a multilingual model, we only need one such model, and thanks to its intelligent architecture, it can scale efficiently in terms of space and training time[12]. Lastly, managing one model instead of n such models leads to decreased manpower, decresed resource usage, and less infrastructure to implement the model.

**Optimizations with multilingual models**

There are several optimization techniques available for both bilingual and multilingual models. These include in-domain finetuning and back-translation. In-domain finetuning is a technique in which a trained model is further trained on a specific domain such as news or medical text [23], or trained on a specific language direction such as Russian to English. In-domain finetuning has produced better results for the said task or domain [27]. Back-translation is the process of creating synthetic source sentences by translating in the reverse language direction. While it has shown improvements for all models, it may cause overfitting [13]. [13] discovers that multilingual systems respond better to optimization techniques than bilingual systems, meaning their change in BLEU score for the same optimization is greater than that of a bilingual system. Combining a sparse matrix of MoE layers with the sub-component architecture and the intelligent optimizations allowed facebook's team to finally beat out bilingual models in both LRL and HRL for the first time [13].

### **Turning Adversarial Losses to Datasets**

Adversarial learning has shown promising results in multiple experiments. In adversarial NMT, the training of the NMT model is assisted by an adversary, an elaborately designed 2D convolutional neural network (CNN) whose goal is to differentiate the translation result generated by the NMT model from that by a human [13]. The goal of this NMT model is to cheat the adversary by producing high-quality translations. The model forces output to be as good as human translations by detecting counterfeit from genuine.

**Why unsupervised learning fails with LRL**

[14] investigates a generative adversarial training (GAN) based approach in low resource situations for speech recognition. Figure 3 shows the architecture for the GAN model.

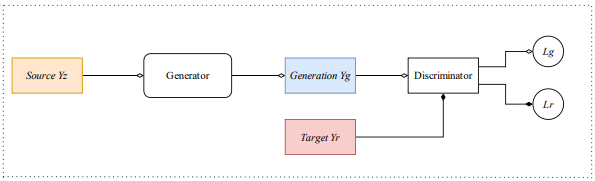


Figure 3: GAN model architecture [15]

Their procedure consists of two generators and two discriminators. The generators convert the source/target speaker’s voice into the target/source speaker’s voice, and discriminators judge whether the input voice is from a real dataset or a generator. The two components constantly play minmax games with each other based on the objective function shown in figure 4.

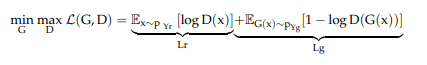


Figure 4: GAN Objective Function [15]

Here, acoustic features in the training data are converted to target speaker-like data via the generator of CycleGAN, and then the ASR model is trained with the original and the converted training data [16].

GAN methods are unsupervised and produce realistic outputs. They are heavily utilized to produce explicit images that can even fool the human eye [14]. However, they are dependent on large data distributions which can overlap between real and generated data. Although promising, they are not feasible with LRL which lacks large datasets.

**Improvements with LAC**

[14] propose a Low Resource, Adversarial, Cross-Lingual (LAC) model. It consists of a discriminator, a generator, a class of feed-forward neural networks called multilayer perceptron (MLP), and source and target embedding. Figure 5 shows the architecture for LAC.

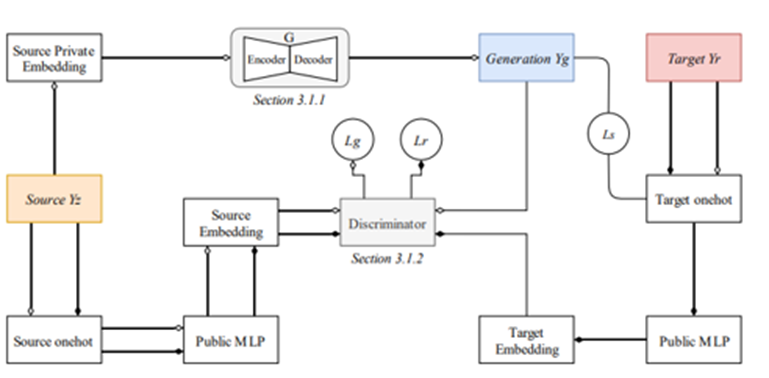


Figure 5: LAC model architecture [14]

Yz, Yr, and Yg are source language distributions, human translation, and generated translation. Lr and Lg are adversarial losses, whereas Ls is translation loss. The discriminator gets fed two different pairs of data, one with source and human translation and the other with source and generated translation.

The generator utilizes the rival loss from the discriminator to yield better results. The generator is based on RNNsearch which enhances performance. The discriminator is fed two different pairs of data, one with source and human translation, and other with source and generated translation. The objective function for this model is shown in figure 6.

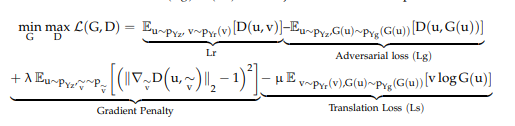


Figure 6: LAC model architecture [14]

Here, λ is the penalty coefficient, whereas µ controls the translation coefficient. v\*logG(u) is the cross-entropy of the real and generated translations. This objective function consists of the adversarial rival loss of GAN as well as cross-entropy loss between original and translated data.

Although multiple layers would have increased efficiency, only one layer was implemented for simplicity and proof of concept. The researchers then used Tatoeba dataset, which comprises short and clean parallel language pairs from 81 languages and widely used for rare language NMT research [17]. Preprocessing was done on the corpus, and parameters were set before feeding into the LAC. The results are described in detail in the comparison section.

# **Criteria**

1. **Maximize low-resource language performance**
   1. Low and Very Low corpara
      1. Must provide reliable translations for low-resource corpara
   2. Non-existent corpara
      1. Larger capacity allotments will allow for emergence of zero-shot approaches to unseen languages.
2. **Maintain high-resource language performance** 
   1. Must maintain comparable performance for standard High-resource language models
3. **Computational Complexity** 
   1. Training
      1. Training Time must be comparable to that of current bilingual models
   2. Excution
      1. Must be able to execute in a distributed system
      2. Must be able to execute on a singlular low-resource system
4. **Scalibility** 
   1. Must be able to train across multiple systems simultaneously i.e. distributed systems ready
   2. Complexity scales non-linear with the number of languages provided
   3. Integration of a new language or langue group must not require retraining of every sub-group.
   4. Must be modular to allow for configuration and swapping of sub-components.

# **Comparison**

### **Supervised training vs. unsupervised training**

Table 3: BiLex Comparisons [7]

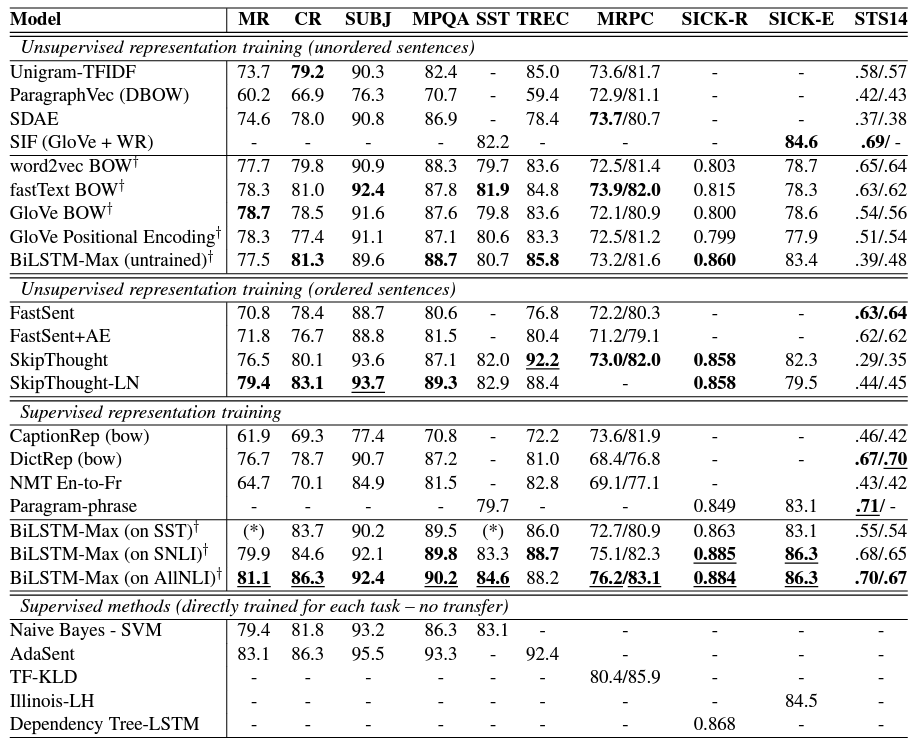


Table 3 shows the comparison results of supervised and unsupervised training for NLP in various training methods. The bold is the best result among the models trained similarly. From the experiment, supervised training consistently outperformed unsupervised training models. Therefore, to maintain promising performance, our group chose to work around supervised training.

### **BiLex model vs. existing model**

As shown in tables 1 and 2 on page 9, BiLex shows consistent and promising results when compared to both supervised and unsupervised models on both word translation tasks and sentence translation tasks. BiLex is able to maintain consistent performance when other supervised models, like BiCVM, BIlBOWA, and supervised MUSE are outperformed by unsupervised MUSE because of fewer available bilingual definitions.

### **Multilingual Models vs Bilingual Models**

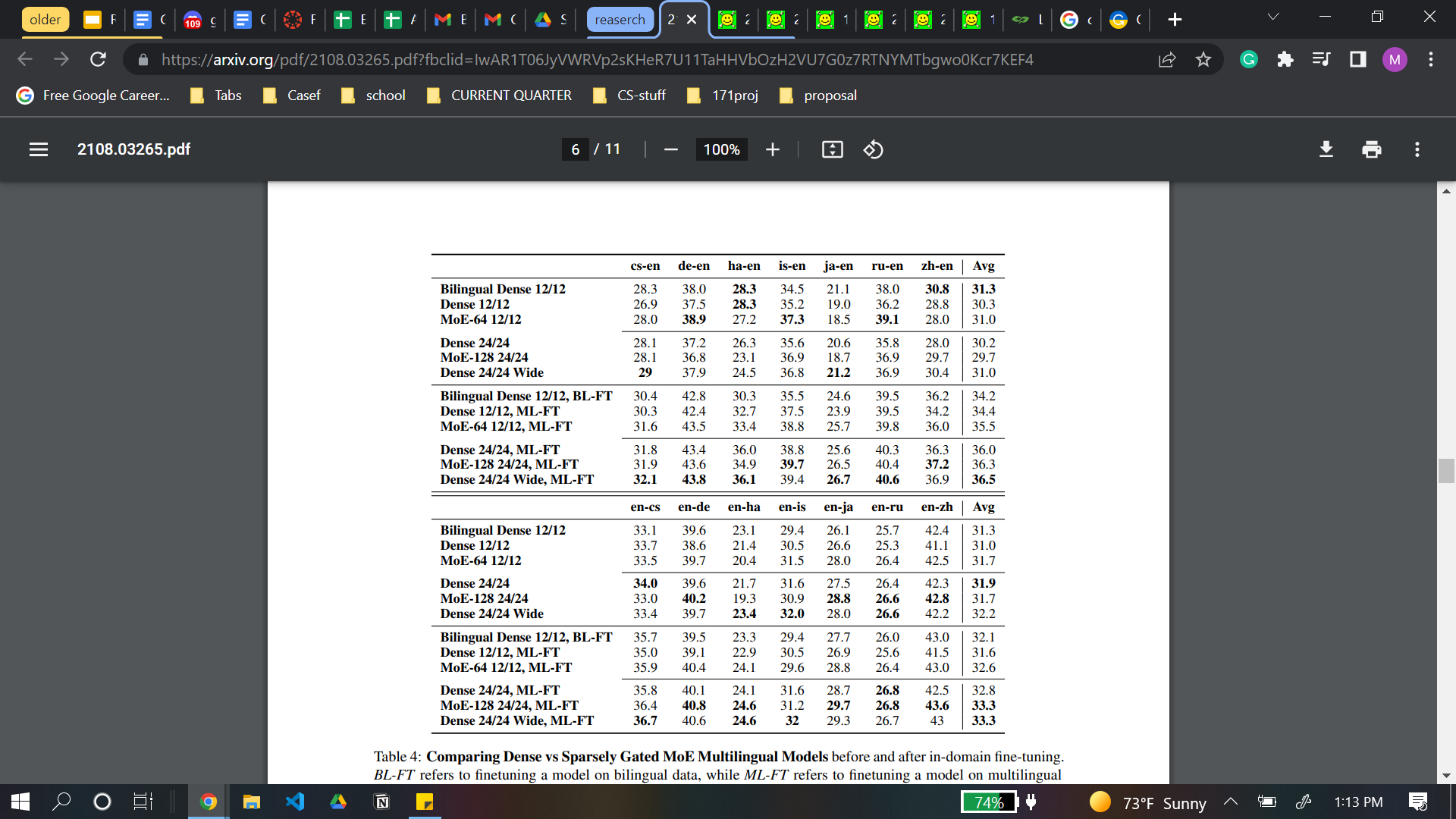
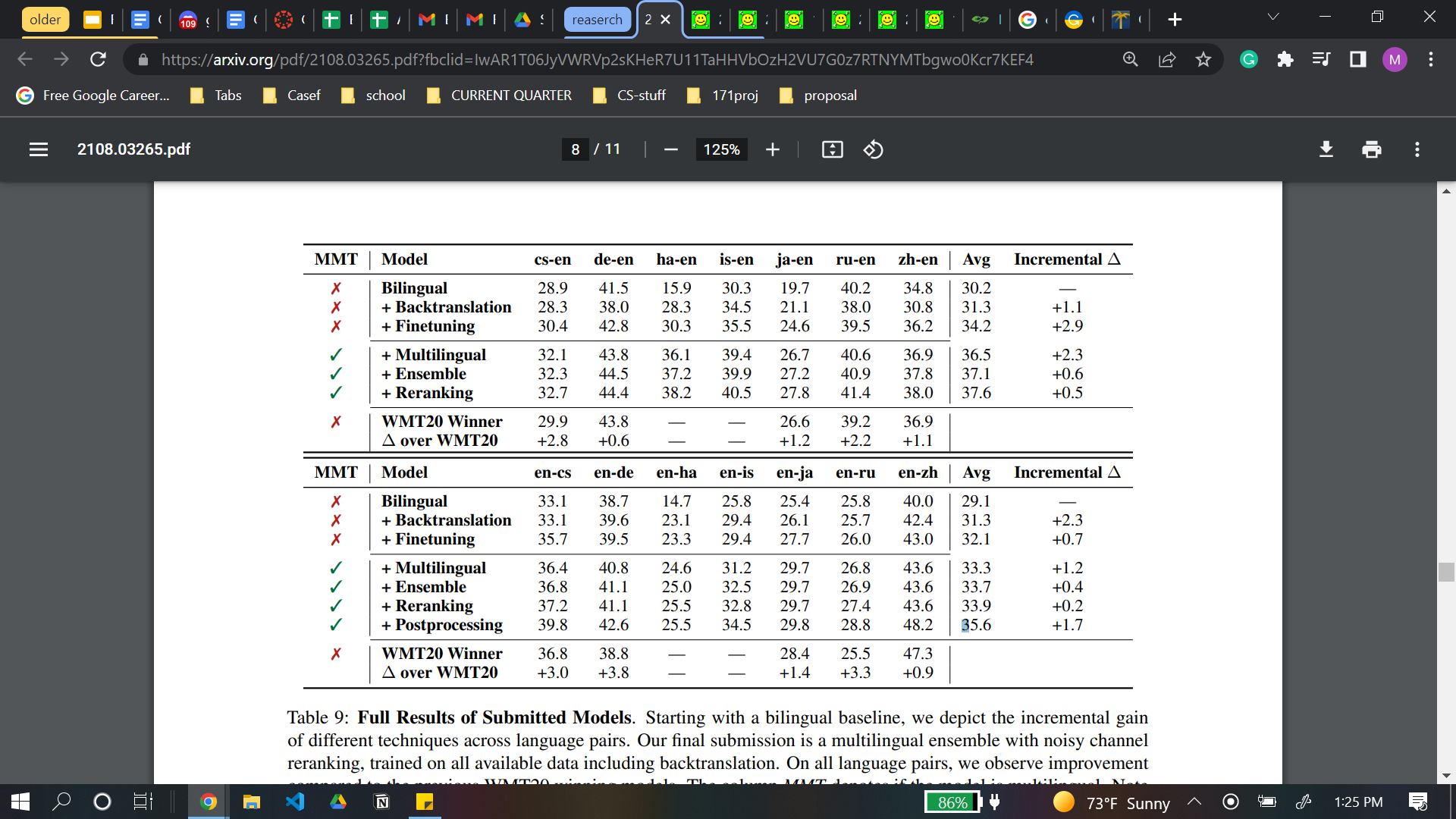
All comparisons were made using the Bilingual Evaluation Understudy(BLEU), a standard metric for measuring translation quality between languages, where the higher the BLEU score the better. Table 7 shows the BLEU scores between baseline multilingual and bilingual 

Table 7: Baseline and fintuned multilingual and bilingual models[13]

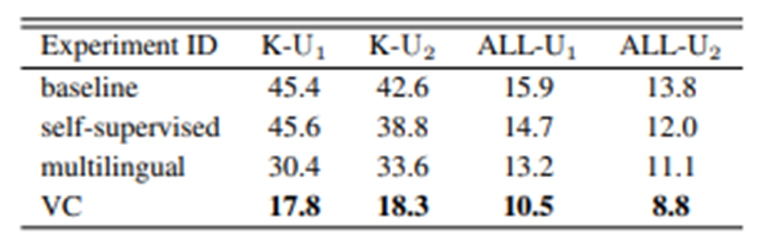
models and finetuned multilingual and bilingual models. It should be noted that several types of multilingual models are measured, which are all either dense or sparse. The biggest model is the MoE-128. This table shows the measurements for several language directions. Overall the best multilingual models beat out the best bilingual models, with the highest performance going to MoE-128 likely due to its increased capacity. The increased capacity that the MoE and sparse architecture enable. In most cases, the finetuned models outperform the baselines except for some language directions, which may suffer from back-translations adverse effects.

Table 8 measures both BLEU scores and changes in BLEU when specific optimization techniques were applied. Table 8 also shows the increase in BLEU score over the previous year's winner. In almost every case, the multilingual model beats out the bilingual model. In addition, the new multilingual model beats out the previous year's model from 0.5 to 4 BLEU, depending on the language direction. Lastly, the complexity and architecture of the multilingual model allow for new optimization techniques not available to a bilingual model, such as reranking and ensembling. Reranking shows an improvement of 0.3 - 0.5 BLEU, while ensembling seems to vary greatly depending on the language direction. This table shows how this new multilingual model beats the previous models in HRL and LRL, paving the way for a more efficient and possibly zero-shot-able approach to NMT.

Table 8: Change in BLEU based on optimizations[13]

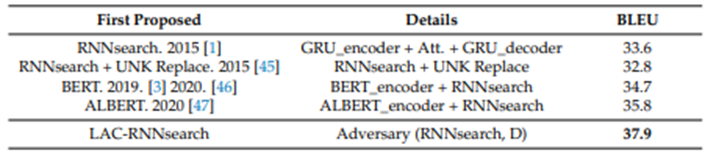
### **LAC vs Traditional Methods**

The results of preliminary experiments performed by extracting data every 10 ms over a 25-ms window can be seen in table 4. Here, VC refers to CycleGAN-VC2, the primary training model being used while multilingual refers to multilingual training, where corpora of another HRL is used for training and is another popular method in the field of LRL NLP. K-U1, K-U2 etc. simply refer to different known and unknown target speakers chosen for the experiment. There was a 33.7% relative performance increase (lower the better) by using the adversarial training method over traditional ones such as multilingual which can be seen in table 1.

Table 4: Comparison data among models [18]

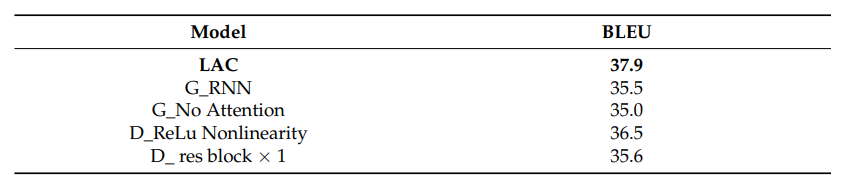
Bilingual Evaluation Understudy (BLEU) score is a standard metric which is used to compare translation between texts by measuring similarity between machine translated and human translated texts. Higher BLEU score is better. The results of preprocessing done with LAC was compared to other traditional networks. The LAC model presented here had a BLEU score of 37.9 in the Turkish-English dataset, which is better than all the previous models, and is shown in table 5.

Table 5: Baseline comparison of LAC to previous models [19]



Ablation study is the analysis of AI systems by removing certain key components [20]. Such study was also performed on LAC and the results are summarized in table 6. It shows that the encoder-decoder and discriminator used in LAC is superior compared to other models, as indicated by the higher score.

Table 6: Ablation study comparison [14]



Character n-gram F3-score (ChrF3), another metric, also had higher performance in all languages. The ChrF3 curves for four language pairs can be seen in figure 7. The solid line refers to the LAC model whereas the dotted line refers to the previous RNNsearch model. The LAC scored higher in all translations when compared to RNNsearch. The difference grew larger as time went by. The LAC was able to capture more information through the adversary since an extra rival loss from the discriminator was learned by the generator [14].

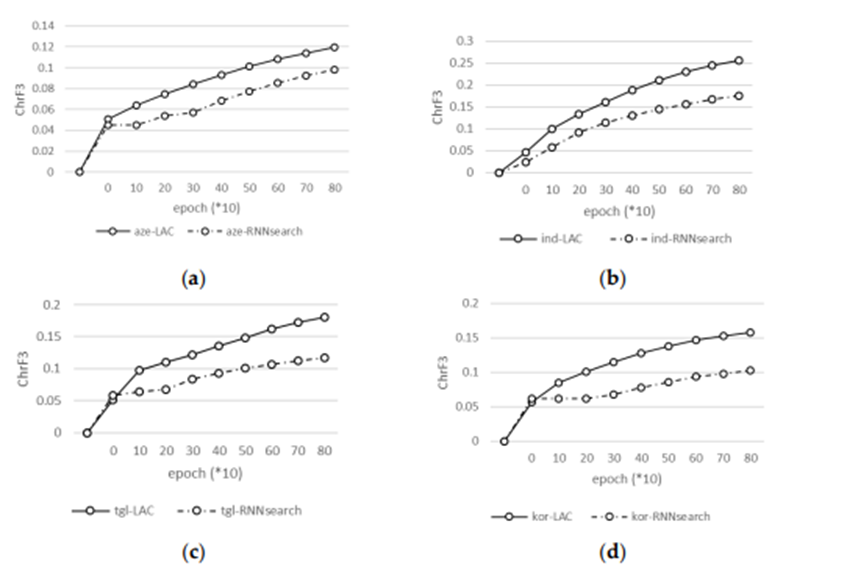


Figure 2: ChrF3 curves for four language pairs: (a) aze-eng (b) ind-eng (c) tgl-eng (d) kor-eng [20]

LAC combines the best part of its predecessors. The downside to this model is its higher use of computational power. It is also a supervised model and requires human intervention [21]. As it is much harder to gain datasets of several LRL quickly, especially when there are thousands of languages and multiple dialects, improving machine learning methods to generate such data accurately and reliably is one feasible way. Due to these reasons, we propose that LAC is optimal when no datasets are available.

# **Recommendation**

The overall recommendation for the improved natural language process model contains two major parts.

First, we generate datasets for supervised training using the BiLex method. The results are more accurate compared to unsupervised training and it does not require the involvement of a large amount of human resources. Therefore, it is an ideal and feasible solution specific to low resource languages.

Second, the generated dataset will be fed into a multilingual NMT model . Wrapping everything so far in an adversarial training schema will help to filter any noise in the system. The end result will be a single model which will be high quality for LRLs, and even useful for some HRLs. It will also be space efficient because it only takes as much space as one model.

Lastly, the system can take advantage of fine tuning and back translation. So, we recommend incorporating a hyperparameter search system into the adversarial training schema will likely yield better results. Since we are using a sparse matrix for the architecture, we can reasonably conclude with no drastic overhead. It should be noted that other preprocessing techniques may also be used.

# 

# **References**

[1]Kumar, Ela, Natural Language Processing, I. K. International Pvt Ltd, 2013

[2]Magueresse, Alexandre et al. “Low-resource Languages: A Review of Past Work and Future Challenges.” *ArXiv* abs/2006.07264 (2020): n. Pag.

[3]​​Anil Kumar Singh. 2008. Natural language processing for less privileged languages: Where do we come from? where are we going? In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages.

[4] Weijia Shi, Muhao Chen, Yingtao Tian, and Kai-Wei Chang. 2019. [Learning Bilingual Word Embeddings Using Lexical Definitions](https://aclanthology.org/W19-4316). In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 142–147, Florence, Italy. Association for Computational Linguistics.

[5] Z. Z. Hlaing, Y. K. Thu, T. M. Oo, T. Supnithi and P. Netisopakul, "Supervised Bilingual Word Embeddings for Low-Resource Language Pairs: Myanmar and Thai," 2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2021, pp. 1-6, DOI: 10.1109/iSAI-NLP54397.2021.9678157.

[6] M. Maimaiti, Y. Liu, H. Luan and M. Sun, "Enriching the transfer learning with pre-trained lexicon embedding for low-resource neural machine translation," in Tsinghua Science and Technology, vol. 27, no. 1, pp. 150-163, Feb. 2022, DOI: 10.26599/TST.2020.9010029.

[[7]Alexis Conneau](https://dblp.uni-trier.de/pid/182/2017.html), [Douwe Kiela](https://dblp.uni-trier.de/pid/136/9140.html), [Holger Schwenk](https://dblp.uni-trier.de/pid/92/6322.html), [Loïc Barrault](https://dblp.uni-trier.de/pid/86/7823.html), [Antoine Bordes](https://dblp.uni-trier.de/pid/49/4572.html): Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. [CoRR abs/1705.02364](https://dblp.uni-trier.de/db/journals/corr/corr1705.html#ConneauKSBB17) (2017)

[8]Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, "A neural probabilistic language model", *The Journal of Machine Learning Research*, vol. 3, pp. 1137-1155, Feb 2003.

[9] K. Wu, X. Wang and A. Aw, "Bilingual word embedding with sentence similarity constraint for machine translation," *2017 International Conference on Asian Language Processing (IALP)*, 2017, pp. 119-122, doi: 10.1109/IALP.2017.8300560.

[10] TomasMikolov, Quoc VLe, andIlyaSutskever. 2013a. Exploiting similarities among languages for machine translation. CoRR,abs/1309.4168.

​​[11] Sebastian Ruder, Ivan Vuli´c, and Anders Søgaard. 2017. A survey of cross-lingual word embedding models. Journal of Artificial Intelligence Research.

[12] Y. Tang et al., “Multilingual Translation with Extensible Multilingual Pretraining and Finetuning.” arXiv, 2020. doi: 10.48550/ARXIV.2008.00401.

[13] C. Tran, S. Bhosale, J. Cross, P. Koehn, S. Edunov, and A. Fan, “Facebook AI WMT21 News Translation Task Submission.” arXiv, 2021. doi: 10.48550/ARXIV.2108.03265.

[14] M. Sun, H. Wang, M. Pasquine, and I. A. Hameed, “Machine Translation in Low-Resource Languages by an Adversarial Neural Network,” *Applied Sciences*, vol. 11, no. 22, p. 10860, Nov. 2021, doi: 10.3390/app112210860.

[15] Daniel Grießhaber, Ngoc Thang Vu, Johannes Maucher, “Low-resource text classification using domain-adversarial learning,” *Computer Speech & Language*, Volume 62, 2020, 101056,<https://doi.org/10.1016/j.csl.2019.101056>.

[16] Lee, H.; Yu, T. ICASSP 2018 tutorial: Generative adversarial network; Its applications to signal processing; Natural language processing. In Proceedings of the ICASSP 2018, Calgary, AB, Canada, 15 April 2018.

[17] Wu, L., Xia, Y., Tian, F., Zhao, L., Qin, T., Lai, J. &amp; Liu, T., “Adversarial Neural Machine Translation”, *Proceedings of The 10th Asian Conference on Machine Learning*, 95:534-549, 2018,<https://proceedings.mlr.press/v95/wu18a.html>.

[18] Choi, Y.-S.; Park, Y.-H.; Yun, S.; Kim, S.-H.; Lee, K.-J. Factors Behind the Effectiveness of an Unsupervised Neural Machine Translation System between Korean and Japanese. Appl. Sci. 2021, 11, 7662

[19] K. Matsuura, M. Mimura, S. Sakai and T. Kawahara, "Generative Adversarial Training Data Adaptation for Very Low-resource Automatic Speech Recognition", *arXiv.org*, 2022. [Online]. <https://arxiv.org/abs/2005.09256>.

[20] Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), Doha, Qatar, 25–29 October 2014; pp. 1724–1734.

[21] Fu, Z.; Xian, Y.; Geng, S.; Ge, Y.; Wang, Y.; Dong, X.; Wang, G.; de Melo, G. ABSent: Cross-lingual sentence representation mapping with bidirectional GANs. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020), New York, NY, USA, 7 February 2017.

[22] K. He, Y. Yan and W. Xu, "Adversarial Cross-Lingual Transfer Learning for Slot Tagging of Low-Resource Languages," *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9207607.

[23] Randhawa G, Ferreyra M, Ahmed R, Ezzat O, Pottie K. Using machine translation in clinical practice. Can Fam Physician. 2013;59(4):382-383.

[24]Z. Wang, K. K, S. Mayhew, and D. Roth, “Extending Multilingual BERT to Low-Resource Languages.” arXiv, 2020. doi: 10.48550/ARXIV.2004.13640.

[25]N. Arivazhagan et al., “Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges.” arXiv, 2019. doi: 10.48550/ARXIV.1907.05019.

[26]Y. Wang, C. Zhai, and H. H. Awadalla, “Multi-task Learning for Multilingual Neural Machine Translation.” arXiv, 2020. doi: 10.48550/ARXIV.2010.02523.

[27]A. Fan et al., “Beyond English-Centric Multilingual Machine Translation.” arXiv, 2020. doi: 10.48550/ARXIV.2010.11125.

[28]A. Vaswani et al., “Attention Is All You Need.” arXiv, 2017. doi: 10.48550/ARXIV.1706.03762.

[29]D. Lepikhin et al., “GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding.” arXiv, 2020. doi: 10.48550/ARXIV.2006.16668