Cross-Lingual Question Answering

A Aparajitha(2022814001), Darshana S(2022701012)

November 16, 2022

1 Introduction

The objective of extractive Question Answering (QA) is to find the answer to a question as a span of words from a given context paragraph. The span is set by the start and end positions in the context. The data for Extractive QA is of the form $\{context(c), question(q), answer(a)\}$. Datasets for QA in languages other than English are sparse. Creation of such datasets is expensive and requires either a lot of manual work or good quality machine translations.

Cross lingual learning is an approach that tries to solve these issues by transferring the knowledge acquired from one language to another. More specifically, the goal is to transfer the knowledge from a high resource language to a low resource language. In the context of QA, for example, this can be from English to Hindi. Formally, Cross Lingual Transfer (XLT) task for QA requires a model trained with $\{c_{l_x}, q_{l_x}, a_{l_x}\}$ to be able to predict the answer span in $\{c_{l_y}, q_{l_y}, a_{l_y}\}$ where l_x is typically a high resource language and l_y is low resource. Generalized Cross Lingual Transfer (G-XLT) task extends XLT for extracting answers from $\{c_{l_y}, q_{l_z}, a_{l_y}\}$ where l_y and l_z can be any two different languages. Fair evaluation of such systems requires high quality parallel multilingual benchmarks. MLQA (Lewis et al. (2020)) is a benchmark covering 7 languages and diverse domains.

We attempt to build a cross-lingual QA model (CLQA) by training it with English data and evaluating with English, Spanish and Hindi in different zero-shot and few-shot settings. We also experiment various methods to improve and analyze the results.

2 Dataset

SQuAD (Rajpurkar et al. (2016)) is a monolingual dataset $\{c_{en}, q_{en}, a_{en}\}$ covering over 100,000 samples from several articles. We use SQuAD v1.1, which does not contain any unanswerable questions to fine-tune over pre-trained large language models.

MLQA contains QA samples for 7 different languages including English. For the purpose of the project we are choosing English, Spanish and Hindi. The languages are chosen based on their similarity and distance from English and the results in the paper. Spanish due to its similarity with English performed significantly better compared to Hindi which is syntactically and typologically different from English. The data is evaluated with the standard metrics F1 and Exact Match(EM).

EN	ES	HI
11590	5253	4918

Table 1: Number of Instances

3 Approaches

The details of the architecture for all the approaches maybe be seen in Appendix A.

3.1 Zero Shot Transfer

We produced the results of MLQA using mBERT (Devlin et al. (2018)) and XLM-R (Lample and Conneau (2019)) as the pre-trained models and then fine-tuned[1] them on SQuAD v1.1 dataset. We chose the most popular and best performing SOTA language models. Results[2] for the same are presented below.

[1]All models are fine-tuned for 1 epoch.

[2]All tables have the question language as columns and context language as rows.

Q/C	EN	ES	HI
EN	78.85	58.77	49.67
ES	66.16	67.17	36.25
HI	57.82	39.05	59.84

Table 2: F1 on BERT

Q/C	EN	ES	HI
EN	65.80	44.92	37.02
ES	48.98	49.07	22.52
HI	41.68	25.36	42.94

Table 3: EM on BERT

Q/C	EN	ES	HI
EN	79.27	66.29	42.45
ES	67.29	63.65	36.54
HI	54.67	46.11	49.11

Table 4: F1 on XLM-R

Q/C	EN	ES	HI
EN	66.16	52.31	29.80
ES	49.41	46.24	21.47
HI	39.85	32.09	34.01

Table 5: EM on XLM-R

We have also evaluated the benchmark using MuRIL (Khanuja et al. (2021)) pre-trained model instead of BERT. MuRIL is a multilingual LM specifically built for Indian Languages. It is trained on 16 Indian languages and English. The F1 scores and Exact match for XLT in Hindi and G-XLT in En-Hi and Hi-En did not improve, contrary to what we expected. We discuss more about the reasons for this in Section 5.

Q/C	EN	ES	HI
EN	58.08	-	8.61
ES	-	6.24	0.36
HI	47.09	1.39	43.33

Table 6: F1 on MuRIL

Q/C	EN	ES	HI
EN	48.74	-	5.73
ES	-	2.93	0.23
HI	35.94	1.10	32.20

Table 7: EM on MuRIL

3.2 Few Shot Transfer

As a way to improve the performance we explored to see if few-shot transfer would perform better than zero shot transfer. From the research conducted by Lauscher et al. (2020) it was concluded that for higher level language tasks the gains are less pronounced with few-shot even after seeing 1,000 target language instances. To test in a few shot setting, we fine-tuned the model on 500 instances of Hindi separately and Spanish and Hindi together. Considering the resources, we used 500 samples from the MLQA dev data.

As can be seen in the results, there is a performance improvement for Hindi and Spanish.

Q/C	EN	ES	HI
EN	-	-	47.97
ES	-	-	43.38
HI	55.17	46.19	54.45

Table 8: F1 on BERT Few HI

Q/C	EN	ES	HI
EN	-	-	35.17
ES	-	-	27.91
HI	39.04	31.39	38.14

Table 9: EM on BERT Few HI

Q/C	EN	ES	HI
EN	78.92	69.69	48.33
ES	67.29	66.92	46.19
HI	56.26	48.55	53.74

Table 10: F1 on BERT Few HI ES

Q/C	EN	ES	HI
EN	65.66	55.39	36.21
ES	49.30	48.63	31.39
HI	40.17	32.15	37.02

Table 11: EM on BERT Few HI ES

3.3 Two-Stage Training with MML

We implemented a two-stage training method with Maximum Marginal Likelihood (MML) loss.

3.3.1 k-Best Answers

Evaluation on MLQA is done with just one top answer. We tried evaluating using the top 3 best answers instead and found the models to perform better which was in line with what we expected. The models are able to find the spans correctly but it is not always the best answer. In fact, taking the top 20 answers for English question and English context using XLM-R resulted in a $F1$ of 96.49 and EM of 93.67.

Q/C	EN	ES	HI
EN	87.84	75.20	69.88
ES	83.13	83.88	59.79
HI	76.26	62.08	77.46

Table 12: F1 on BERT Top 3

Q/C	EN	ES	HI
EN	80.37	63.75	56.95
ES	67.73	68.37	39.23
HI	63.76	47.35	65.10

Table 13: EM on BERT Top 3

Q/C	EN	ES	HI
EN	89.08	71.71	62.80
ES	79.00	80.40	49.58
HI	72.42	52.49	73.70

Table 14: F1 on XLM-R Top 3

Q/C	EN	ES	HI
EN	81.96	61.08	50.75
ES	65.39	66.49	32.44
HI	58.29	39.11	59.90

Table 15: EM on XLM-R Top 3

3.3.2 Maximum Marginal Likelihood Loss

We established that the model is able to find the answer in the top 20 results in English-English XLT task. The models just was not able to rank it as the top answer. We

Since, optimization with the standard Cross Entropy Loss considers only the top one prediction, the model is sub-optimized (Chen et al. (2022)). Hence, we utilize the tope 20 best predictions pre-obtained from the model trained on original SQuAD. If the the top predictions did not contain ground truth, it was substituted with the last answer. In stage 2, the model is optimized on all the top 20 answers using max marginal likelihood loss.

$$L_{mml} = -log \sum_{z_l \in Z} P(z_l | q_i, c_i) \tag{1}$$

Q/C	EN	ES	HI
EN	78.34	59.77	52.41
ES	65.75	66.33	38.35
HI	58.33	39.17	59.62

Table 16: F1 MML

Q/C	EN	ES	HI
EN	64.97	46.12	39.69
ES	47.59	47.97	24.72
HI	41.64	25.07	42.43

Table 17: EM MML

4 Analysis

We can see from all the results that zero-shot transfer with a multilingual pre-trained model performs well, reaching an exact match score of 66 for the task in English. For a model that has not seen any samples in the target language, the performance is commendable. In general, with all approaches other than MuRIL the models perform better on the XLT task than G-XLT which is as expected. They perform best for English, then Spanish and finally Hindi.

In G-XLT, the performance for English-Spanish pair is the best, followed by English-Hindi and finally Spanish-Hindi.

Few-Shot transfer with some samples in Spanish and Hindi show improvements in both the languages but show a slight dip in the performance of English.

To perform a two stage training, we first evaluated against the top 3 best answers instead of 1. As can be seen in the results, there is a high increase in the performance of both tasks across all the languages. Hindi-Hindi shows the maximum performance improvement with an increase of 28.08 in the F1 and 23.72 in EM scores using BERT. The least is seen in Spanish-Hindi with an increase of 7.83 in F1 and 13.60 in EM. The trend continues in XLM-R with Hindi-Hindi increasing by 24.58 in F1 and 25.88 in EM. In least improvement is seen in English-Spanish with 5.42 in F1 and Hindi-Spanish with 7.02 in EM scores.

We then evaluated in Top 20 answers setting for BERT in English-English and saw the results go up to 96.49 in F1 and 93.67 in Exact Match. Having seen that the results improved, there was evidence that the model was able to identify the correct spans but they just were not the top 1 answer. Using this intuition, we trained with MML loss. The results show minor improvement in some of the G-XLT language pairs but a dip in the scores for XLT. We believe that better hyperparameters and training may improve the results.

5 Performance of Hindi

5.1 Transliteration

As can be seen in the Appendix B, some of the questions in Hindi are transliterations and not translations whereas the context has the translation for the words or vice versa. This inconsistency might be possible due to the approach used for data collection. While the context is directly taken from parallel Wikipedia, professional translators, translated the question from English to Hindi and other target languages.

Prior experiments (Pires et al. (2019)) on the effective transfer to transliterated languages suggest that mBERT might not be effective in such scenarios. As we can also see from the results, the model performs better with XLM-R (Conneau et al. (2019)) which is better equipped for code-switching due to its pre-training objective but there is no evidence to point for improved performance with transliteration. Though MuRIL uses a transliterated dataset (Roark et al. (2020)) and transliterated (Bhat et al. (2014)) Wikipedia, the transliteration goes from Indian languages to Latin script and not from English to Indian scripts. We believe that this might be one of the reasons for the models to not perform very well in Hindi.

5.2 Low Resource

While mBERT uses Wikipedia, XLM-R trains on the Common Crawl corpus. (Wu and Dredze (2020)), we can see that though English accounts for a large amount of data, cross-lingual transfer learning should work even on low-resource languages.

6 Other Approaches - A Survey

6.1 Language Models

Large Language Models (LLMs) with the vast amount of data learn the syntax and semantics of all the languages they are trained on. As can be seen in (Petroni et al. (2019); Brown et al. (2020)) LLMs learn not just language but also facts and can be considered Knowledge Bases.

6.2 Language Models for QA

Current SOTA models like XLM-R (Conneau et al. (2019)), XLM-E (Chi et al. (2021b)), InfoXLM (Chi et al. (2021a)) and mT5 (Xue et al. (2021)) are all trained with a Language Modeling objective. Replicating a massive language modeling objective is beyond our current scope. Observing the XTREME (Hu et al. (2020)) leaderboard, it is clear that Question Answering task is not the one pulling up the averages.

6.3 Alternate Datasets

We explored the idea of using a different dataset to fine-tune with instead of SQuAD but were limited by the availability of the same. We were looking for datasets that might in be better languages (Lin et al. (2019)) to transfer from in a zero-shot setting. It is evident that the lack of data is what led to cross-lingual transfer learning when we tried to consider languages other than English.

6.4 Knowledge Graphs

We have also looked into using knowledge graphs through models like XLM-K (Jiang et al. (2021)) and through approaches like (Xu et al. (2022)) and could not see any improvement as expected.

6.5 Data Augmentation

We looked into data augmentation (Riabi et al. (2021)) and once again faced with underwhelming performance on QA task. We also considered data augmentation through translation. (Debnath et al. (2021)) do several experiments on few-shot transfer and data augmentation through translation on the TyDi QA dataset (Clark et al. (2020)).

6.6 Multitask Learning

We also considered the approach of Multitask Learning with an auxiliary task. Extractive summarization seemed to be the closest task to extractive QA. (Ahuja et al. (2022)) explore the idea of joint training to improve zero-shot performance in multilingual models. The relationship between languages is also explored.

7 Future Work

Traditionally, extractive QA has been worked on with an independence assumption for modeling the span probabilities i.e. $P(span) = P(span_{start}) * P(span_{end})$. Work (Fajcik et al. (2021)) on a joint probability of the span start and end model $P(span_{start}, span_{end})$ may be attempted in a multilingual setting.

We did not attempt to use any of the generator family models (Xue et al. (2021); Dabre et al. (2022)) for performing the task of extractive QA. With minor modifications, it is possible to test the performance of these models on a non-generation task.

A span-based pre-training (Ram et al. (2021); Glass et al. (2020)) objective that was evaluated with SQuAD could also be evaluated with multilingual data.

8 Conclusion

We conducted several experiments to achieve Cross Lingual Transfer and Generalized Cross Lingual Transfer. We analyzed both the results, and the raw predictions from the model. We put forth a hypothesis on why the cross lingual transfer does not perform up to the expectations in Hindi.

References

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. Multi task learning for zero shot performance prediction of multilingual models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5454–5467, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.374. URL https://aclanthology.org/2022.acl-long.374.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, page 48–53, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450337557. doi: 10.1145/2824864.2824872. URL https://doi.org/10.1145/2824864.2824872.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sas-

try, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL `https://arxiv.org/abs/2005.14165`.

Nuo Chen, Linjun Shou, Ming Gong, and Jian Pei. From good to best: Two-stage training for cross-lingual machine reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10501–10508, Jun. 2022. doi: 10.1609/aaai.v36i10.21293. URL `https://ojs.aaai.org/index.php/AAAI/article/view/21293`.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online, June 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.280. URL `https://aclanthology.org/2021.naacl-main.280`.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. Xlm-e: Cross-lingual language model pre-training via electra. *arXiv preprint arXiv:2106.16138*, 2021b.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020. doi: 10.1162/tacl_a_00317. URL `https://aclanthology.org/2020.tacl-1.30`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.145. URL https://aclanthology.org/2022.findings-acl.145.

Arnab Debnath, Navid Rajabi, Fardina Fathmiul Alam, and Antonios Anastasopoulos. Towards more equitable question answering systems: How much more data do you need? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 621–629, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.79. URL https://aclanthology.org/2021.acl-short.79.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Martin Fajcik, Josef Jon, and Pavel Smrz. Rethinking the objectives of extractive question answering. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 14–27, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrqa-1.2. URL https://aclanthology.org/2021.mrqa-1.2.

Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, G P Shrivatsa Bhargav, Dinesh Garg, and Avi Sil. Span selection pre-training for question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2782, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.247. URL https://aclanthology.org/2020.acl-main.247.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization, 2020. URL https://arxiv.org/abs/2003.11080.

Xiaoze Jiang, Yaobo Liang, Weizhu Chen, and Nan Duan. Xlm-k: Improving cross-lingual language model pre-training with multilingual knowledge, 2021. URL https://arxiv.org/abs/2109.12573.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. Muril: Multilingual representations for indian languages, 2021. URL https://arxiv.org/abs/2103.10730.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.363. URL https://aclanthology.org/2020.emnlp-main.363.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Mlqa: Evaluating cross-lingual extractive question answering, 2020.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1301. URL https://aclanthology.org/P19-1301.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? 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 2463–2473, Hong Kong, China, November 2019. Association for Computational

Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL https://aclanthology.org/P19-1493.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. Few-shot question answering by pretraining span selection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3066–3079, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.239. URL https://aclanthology.org/2021.acl-long.239.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. Synthetic data augmentation for zero-shot cross-lingual question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7016–7030, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.562. URL https://aclanthology.org/2021.emnlp-main.562.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. Processing South Asian languages written in the Latin script: the Dakshina dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2413–2423, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.294.

Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online, July 2020. Association for

Computational Linguistics. doi: 10.18653/v1/2020.repl4nlp-1.16. URL https://aclanthology.org/2020.repl4nlp-1.16.

Liyan Xu, Xuchao Zhang, Bo Zong, Yanchi Liu, Wei Cheng, Jingchao Ni, Haifeng Chen, Liang Zhao, and Jinho D. Choi. Zero-shot cross-lingual machine reading comprehension via inter-sentence dependency graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11538–11546, Jun. 2022. doi: 10.1609/aaai.v36i10.21407. URL https://ojs.aaai.org/index.php/AAAI/article/view/21407.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

A Architecture

All the models were trained using AdamW optimizer and QA Cross Entropy Loss for a single epoch.

$$L_{qa} = -logP(start = a_{i,s}|c_i, q_i) - logP(end = a_{i,e}|c_i, q_i) \quad (2)$$

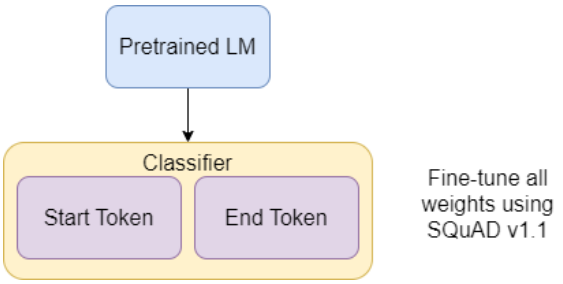

Figure 1: Zero-Shot Transfer Architecture

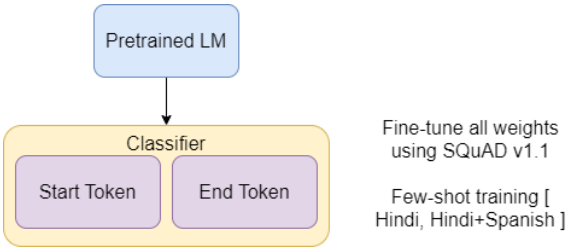

Figure 2: Few-Shot Transfer Architecture

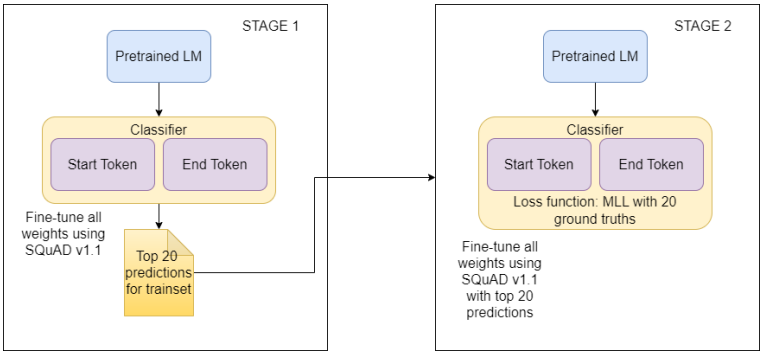

Figure 3: MML Architecture

B Performance of Hindi

Q: किस प्रकार की सड़कें बड़े खेतों और पशु-फार्मों तक जाती हैं?

A: डर्ट-रोड

ग्रूम झील (XLM-R) छोटे-छोटे (mBERT) No Answer (MuRIL)

C: मैं खानों की ओर जाती थीं, लेकिन उनके बंद होने के बाद इन्हें सुधारा गया है। इसके घुमावदार दिशाकोण एक सुरक्षा चौकी से होकर गुजरते हैं, लेकिन अड्डे के चारों ओर का प्रतिबंधित क्षेत्र आगे बढ़कर पूर्व तक फैला हुआ है। प्रतिबंधित क्षेत्र छोड़ने के बाद, ग्रूम झील सड़क पूर्व की ओर तिकाबू घाटी के फर्श की ओर उतरते हुए डर्ट-रोड के प्रवेश द्वार से गुजरती हुई राचेल के दक्षिण, "एक्स्ट्राटेरेस्ट्रिअल हाईवे", स्टेट रूट 375 की ओर अभिमुख होने से पहले कई छोटे-छोटे खेतों से गुजरती है।

Q: What type of roads lead to the ranches?

C: … After leaving the restricted area, Groom Lake Road descends eastward to the floor of the Tikaboo Valley, passing the *dirt-road* entrances to several small ranches, before converging with State Route 375, the "Extraterrestrial Highway", south of Rachel.

Figure 4: Translation in Question, Transliteration in Answer

Q: झील के सापेक्ष ग्रूम लेक रोड कहाँ जाती थी?

A: उत्तर पूर्व

पेचीदा पहाड़ियों (XLM-R) ग्रूम बॉक्स (mBERT) No Answer (MuRIL)

उसी "एरिया XX " नामकरण प्रणाली का प्रयोग नेवादा परीक्षण स्थल के अन्य भागों के लिए किया गया है। मूल रूप में 6 बटे 10 मील का यह आयताकार अड्डा अब तथाकथित 'ग्रूम बॉक्स " का एक भाग है, जो कि 23 बटे 25.3 मील का एक प्रतिबंधित हवाई क्षेत्र है। यह क्षेत्र NTS के आंतरिक सड़क प्रबंधन से जुड़ा है, जिसकी पक्की सड़कें दक्षिण में मरकरी की ओर और पश्चिम में युक्का फ्लैट की ओर जाती हैं। झील से उत्तर पूर्व की ओर बढ़ते हुए व्यापक और और सुव्यवस्थित ग्रूम झील की सड़कें एक दर्रे के जरिये पेचीदा पहाड़ियों से होकर गुजरती हैं। पहले सड़कें ग्रूम घाटी

Q: Where does the Groom Lake Road head relative to the lake?
C: ... Leading northeast from the lake, the wide and well-maintained Groom Lake Road runs through a pass in the Jumbled Hills...

Figure 5: Transliteration in Question, Translation in Answer

Q: शहर में कितने क्वार्टर हैं?\n

A: आठ

5,000 (XLM-R) ईसाई अल-नजाज़्रेह (mBERT) No Answer (MuRIL)

C: बेथलहम के केंद्र में इसका प्राचीन शहर है। यह प्राचीन शहर आठ भागों से मिलकर बना है, जो मैंगर चौक (Manger Square) के आस-पास के क्षेत्र का निर्माण करते हैं। इन भागों में ईसाई अल-नजाज़्रेह (al-Najajreh), अल-फ़राहियेह (al-Farahiyeh), अल-अनात्रेह (al-Anatreh), अल-तराजमेह (al-Tarajmeh), अल-क़वाव्सा (al-Qawawsa) और ह्रीज़त (Hreizat) भाग तथा अल-फ़वाघरेह (al-Fawaghreh)—एकमात्र मुस्लिम भाग— शामिल हैं। अधिकांश ईसाई भागों के नाम उन अरब घैसनिद (Ghassanid) संप्रदायों के नाम पर रखे गए हैं, जो वहां बस गए। अल-क़वाव्सा (Al-Qawawsa) भाग का निर्माण अठारहवीं सदी में समीपस्थ नगर तुक़ु' (Tuqu') से आए अरब ईसाई उत्प्रवासियों द्वारा किया गया था। पुराने शहर के भीतर एक सीरियाई भाग भी है, जिसके नागरिक तुर्की में मिदयात (Midyat) और मा'असार्ते (Ma'asarte) से आए हैं। इस प्राचीन शहर की कुल जनसंख्या लगभग 5,000 है।

How many quarters are there?
... The old city consists of *eight* quarters, laid out in a mosaic style, forming the area around the Manger Square ...

Figure 6: Transliteration in Question, Translation in Answer

इस अवधि के दौरान उनके रिंग नाम को छोटा करके बस ट्रिपल एच कर दिया गया।

During this period, his ring name was shortened to simply Triple H, even though he would still be referred for a while as Helmsley from time to time and Hunter for the rest of his career.

Figure 7: Incomplete translation