Multiple Model Ensemble for Cross-Lingual Question Answering

Neal Rakesh Vaidya
Stanford University
nealv@stanford.edu

Satita Vittayaareekul
Stanford University
satita97@stanford.edu

Seyed Shahabeddin Mousavi
Stanford University
ssmousav@stanford.edu

Abstract

As the amount of information available to the average person has increased dramatically in recent years, much progress has been made in the field of automated question answering systems. Much of this effort has gone into monolingual question answering, especially in English, due to the overwhelming presence of English-language information present on the web. It can be challenging to apply these results to languages with less readily accessible material, imposing barriers on global communities trying to find relevant information. In this project, we propose a novel ensemble method for utilizing large English corpora for answering questions in a variety of languages. We find that, when dealing with multilingual queries and an English only context corpus, query translation and late-interaction retrieval outperforms dense passage retrieval with a multilingual encoder.

1 Introduction

As the arrival of the internet to every corner of the world has broken down barriers to publishing content and diminished the role of gatekeepers, the amount of information available at any person's fingertips has grown at an astonishing rate. In the past, the venues available for getting an answer to a question you had were visiting the local library or asking some people you knew. Now, any given person has the knowledge of billions of people available to them. As noted by Thompson (2015), value in the pre-internet age was generated by being able to publish content to a wide audience, while value today is generated by being able to narrow down the vast amount of available information to those pieces which are relevant to an individual.

One increasingly prominent way of finding relevant information is the process of automated question answering. In these systems, people input a natural language question, and a program will either find or generate an answer. Several prominent datasets, like SQuAD (Rajpurkar et al., 2018) and the Natural Questions dataset (Kwiatkowski et al., 2019), have arisen as standards for this problem, and numerous systems have been applied to these datasets with high degrees of success. Given the nature of data available on the internet, most of these datasets have been composed of English questions and answers, and consequently the bulk of publicly available automated QA systems deal solely with English. This has meant that users of non-English languages face significant barriers in leveraging these systems. We try to address this issue by building a question answering tool where users can pose questions and receive answers in any language, with the answers leveraging English-language data.

2 Prior Literature

While the bulk of question answering work has been applied to English-only questions and corpora, multilingual QA systems have seen increased attention in recent years. In particular, we note two benchmark datasets which have arisen and some of the various approaches that have been used in addressing them.

XOR QA (Asai et al., 2021a) introduces a dataset and benchmark for training and evaluating models for cross-lingual question answering. The dataset is adapted from the previous TyDi-QA dataset (Clark et al., 2020), where question were gathered from native speakers in 11 different languages. Asai et al. propose 3 tasks based on this dataset. These are termed: XOR-RETRIEVE, for retrieving English passages from target language questions; XOR-ENGLISHSPAN, for extracting or generating English answers to target language questions; and XOR-FULL, for extracting or generating answers in the original language of the question. The authors also propose a set of baseline systems for completing these tasks, which we discuss below.

Similarly, MKQA (Longpre et al., 2021) introduce a dataset for multilingual open-domain question answering based on human translations of question-answer pairs from the well known Natural Questions dataset (Kwiatkowski et al., 2019). For a full discussion on the similarities and differences between these two datasets, please see the Data section.

2.1 XOR QA Baseline Systems

The baseline systems presented by Asai et al. (2021a) fall into two categories – translation and multilingual. For the translation systems, questions are translated into English, and passages retrieved using either term-based (BM25) (Robertson and Zaragoza, 2009) or neural (DPR) (Karpukhin et al., 2020) IR. Specific answers are predicted using a fine-tuned BERT (Devlin et al., 2019) model, which are then translated back into the original language. In the multilingual systems, passages in multiple languages are retrieved using DPR with a multilingual BERT encoder, with answers in the original language predicted using XLM-RoBERTa (Conneau et al., 2020). In both cases, passage recall is evaluated by calculating the percentage of of questions which have passages containing the exact answer within the first n (2000 or 5000) tokens, and answer prediction is evaluated using token level F1 scores for both the English and target language answers. In each task, the translation pipeline achieved better results than the multilingual pipeline, with DPR substantially outperforming BM25.

The authors also experimented with using Google Search and Google Translate in place of their own IR and translation methods; the Google methods tended to outperform other methods for most languages.

2.2 Cross-lingual Open-Retrieval Answer Generation (CORA)

Asai et al. (2021b) propose CORA as a model to address one of the tasks presented in XOR-QA, XOR-Full, where a question is asked in a target language and the answer must be generated or retrieved in that same language. CORA utilizes a pipeline of a multi-lingual retriever to retrieve passages from any language, and a multilingual answer generator which uses a sequence-to-sequence model to generate an answer phrase in the target language. Here, the retrieval system is an extended version of DPR using a multilingual BERT model, which the authors term mDPR, and the answer generator (called mGEN) is derived from mT5 (Xue et al., 2021). For the answer generator, a "language token" is added to the prompt to specify what language the generated output should be in.

Parameters for these two systems are also iteratively updated in a fine-tuning process where new training data for mGEN is gathered using the "This page in other languages" feature of Wikipedia, which is then labeled as positive or negative based on whether the answer generator can successfully generate an answer from the new passages. CORA is also evaluated on the MKQA dataset (Longpre et al., 2021), where it achieves state of the art performance by improves on the original baseline on a substantial margin, outperforming the baseline DPR + Machine Translation system by 34 F1 points, and the DPR + Google Translate system by 24 on the XOR-FULL task.

2.3 Single Encoder Retriever (Sentri)

The Sentri+MFiD model is similar to CORA in that it combines a retriever and generator model into a single system. In contrast to CORA, the authors use a modified version of the FiD sequence-to-sequence model (Izacard and Grave, 2020) that has been retrained on a multilingual dataset, hence MFiD. The authors also fix the parameters of the generator model after pre-training and only fine-tune the retriever, unlike CORA which updates both models. With the additional MFiD base, Sentri bumps up 26.1 F1 points, which outperforms CORA over 2.7 F1 scores.

As for data choices, both of these models mine data to overcome data scarcity issues, especially in some low-resource languages. Sentri and CORA choose different methods, Sentri uses M2M100 machine translation models to translate NQ and Trivia QA datasets from English to other languages, while CORA uses Wikipedia language links and create new synthetic answers to mine new training data cross-lingually.

It is worth mentioning that the authors of Sentri think that all languages have complex morphology and other linguistic features will make information retrieval less effective if just make use of token comparison. Thus they choose to normalize the morphology of each language using different approaches.

When it comes to the actual system, both of them adopt the idea of self-training, and create a closed

circle to iterate their system.

2.4 DR.DECR: Dense Retrieval with Distillation-Enhanced Cross-Lingual Representation

With DR.DECR the authors (Li et al., 2021) propose a single-model retriever for the XOR-RETRIEVE task from XOR QA. They find the best performing pipeline where questions are first translated to English with a proprietary machine translation model, and then passages are retrieved using ColBERT (Khattab and Zaharia, 2020). In order to create a single model that can achieve similar performance, Li et al. propose using knowledge distillation to distill information from the translation pipeline into a single retriever model that does no explicit translation at inference time. First, a ColBERT model fine tunes a pretrained multilingual language model on an English only QA dataset. This model will act as the teacher. Then, the student ColBERT model is trained, with distillation happening at two points.

1. The query-passage similarity scores for queries in all languages are trained to match the similarity scores between the corresponding English question and the passages from the teacher model

2. The query encodings themselves of the student model are trained to match the corresponding English encodings of the teacher model. Because ColBERT operates on token-level encoding, this involves a procedure for aligning tokens between the original and translated English queries.

The authors find that while this distillation process does improve the performance of the multilingual IR model, it is still outperformed by the translation + English IR pipeline.

3 Data

For the information retrieval, we utilized the English Wikipedia dump, split into passages of 100 tokens at a time, as our corpus. As for our evaluation set, we adopted the evaluation set made available via the MIA 2022 Shared Task on Cross-lingual Open-Retrieval Question Answering. We used data from both XOR QA & MKQA, a brief description of both of which is made available below.

3.1 XOR QA: Cross-lingual Open-Retrieval Question Answering

We selected the dataset introduced in XOR QA (Asai et al., 2021a) as the main dataset we worked with as it not only presented a solid dataset but also came with comprehensive benchmarking for training and evaluating models for cross-lingual question answering in the related paper. The XOR QA dataset is adapted from the previous TyDi-QA dataset, where questions were gathered from native speakers in eleven different languages. XOR QA selects those questions deemed unanswerable in the original dataset due to lack of same-language information and uses human translation to translate them into English. Human annotators then collected passages from Wikipedia that contained answers for the English questions, in addition to extracting specific answers from those passages. Finally, these answers were verified and professionally translated back into the original question language. Due to this series of steps, the XOR QA dataset contains:

1. Questions in 7 languages
2. English translations of those questions
3. English Wikipedia passages that contain answers to those questions
4. English answers to the translated questions
5. Answers in the original language of the question.

Asai et al. propose 3 tasks based on this dataset corresponding to points 3, 4, and 5 in the above list. These are termed: XOR-RETRIEVE, for retrieving English passages from target language questions; XOR-ENGLISHSPAN, for extracting or generating English answers to target language questions; and XOR-FULL, for extracting or generating answers in the original language of the question.

The authors also propose a set of baseline systems for completing these tasks. The baseline systems fall into two categories – translation and multilingual. For the translation systems, questions are translated into English, and passages retrieved using either term-based (BM25) or neural (DPR) IR. Specific answers are predicted using a fine-tuned BERT model, which are then translated back into the original language. In the multilingual systems, passages in multiple languages are retrieved using DPR with a multilingual BERT encoder, with

answers in the original language predicted using XLM-RoBERTa. In both cases, passage recall is evaluated by calculating the percentage of of questions which have passages containing the exact answer within the first n (2000 or 5000) tokens, and answer prediction is evaluated using token level F1 scores for both the English and target language answers. In each task, the translation pipeline achieved better results than the multilingual pipeline, with DPR substantially outperforming BM25.

The authors also experimented with using Google Search and Google Translate in place of their own IR and translation methods; the Google methods tended to outperform other methods for most languages.

3.2 MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering

LONGPRE, LU, AND DAIBER, 2021 (Longpre et al., 2021) introduce Multilingual Knowledge Questions and Answers (MKQA), an open-domain question answering evaluation set comprising 10k question-answer pairs aligned across 26 typologically diverse languages (260k question-answer pairs in total). Of importance in our decision that MKQA makes a valuable contribution to the field is how well it reflects realistic, real-world settings and how reliable its annotations are: MKQA explicitly makes the decision to create "retrieval-independent answer annotations" that are linked to Wikidata entities and a few other value types. MKQA also utilizes an answer collection procedure that offers a highly reliable & independent evaluation set that is unbiased towards the QA technique used, allowing it to compare the performance of vastly distinct techniques such as knowledge graph-based, dense and sparse retrieval and generative QA techniques on a large number of languages which comes in handy in our research project. Another aspect of MKQA that allows it to offer a comparable evaluation across so many languages is that it consists of one of the largest sets of fully aligned examples across such a diverse set of languages.

4 Model

Our ensemble approach relies on the combination of a number of pre-existing language models, most of which were trained and fine-tuned on different corpora and applied zero-shot to our question answering dataset. In this section we will discuss briefly the models we utilized.

4.1 Query Translation

For language translation model, we choose to use Opus-MT, a set of neural machine translation models built by the Helsinki-NLP group and released on Hugging Face (Tiedemann and Thottingal, 2020). Their models are based on Marian-NMT and trained on Open Parallel Corpus (OPUS) data using a data augmentation technique called "Back Translation" where they only keep the back-translated text to datasets if it's different from the original text after translating it to a different language and translating it back. Moreover, this model could translate from English to French, Portuguese, and Spanish by just a single API, opus-mt-en-ROMANCE.

We choose Opus-MT because the authors have released a large number pretrained models for various translation pairs. Especially well covered are the translation pairs with a target language of English, allowing us to make use of our monolingual information retrieval system with our English corpus. Thus, for every query q^L where the Opus-MT project has an available L to English translation model, we generate a corresponding English query q^{eng}. For languages where no Opus-MT model is available, we remove the queries from our evaluation. We discuss some of the potential impacts of this in the Project Limitations section below.

4.2 Information Retrieval

For our information retrieval module, we utilize a ColBERTv2 late-interaction neural IR model (Santhanam et al., 2021). Unlike the mDPR model utilized by CORA, ColBERTv2 encodes each query and document into a multi-vector representation, allowing for more fine-grained similarity calculations than the single-vector encodings of DPR. Another benefit of ColBERT as discussed by the authors is the ability to ColBERT retrieval models outside of their initial training domains. Because of this quality, we felt that it would be feasible to use publicly available ColBERT model weights trained for passage ranking on the MS MARCO dataset (Bajaj et al., 2016), and apply the model with no fine-tuning to rank our English Wikipedia corpus. Thus, we use ColBERT to retrieve a set of passages $\mathcal{P}^{\text{eng}}$ from our English Wikipedia corpus for each of our translated queries q^{eng}.

4.3 Answer Generation

For answer generation, we use the mGEN model proposed and trained by Asai et al. (2021b) for CORA. This allows us to isolate our model changes to the information retrieval part of the pipeline and make the most direct comparison in performance. In particular, we use the publicly released weights from the $\text{mDPR}_1 + \text{mGEN}_1$ model where the iterative training process of the full CORA model is not performed. As in Asai et al. (2021b), mGEN is based on the mT5 (Xue et al., 2021) sequence-to-sequence model that generates a set of answer tokens for each query, conditioned on the original query q^L (i.e. not the translated query q^{eng}), a language tag L denoting the language of the query, and the English context passages $\mathcal{P}^{\text{eng}}$. Note that while unlike CORA we perform translation on the initial query, like CORA we do not perform any translation on the generated answer, and rely on the pretrained mGEN model to produce answers in the correct language given q^L and L.

5 Methods

For our experiment, we follow closely the examples set by XOR-QA TyDi benchmark and the 2022 Workshop on Multilingual Information Access. In particular, we adopt their primary metric of using macro-avg F1 score, separately reported on the XOR QA and MKQA evaluation sets, while reporting Exact Match (EM) and BLEU scores as secondary metrics. Although EM is the more commonly used metric in evaluating QA systems, Asai et al. (2021a) prefer F1 due to the lower risk of surface-level mismatches in the cross-lingual setting. Similarly, we found implementations of BLEU to be somewhat unreliable due to the difficulty in some languages of accurately identifying word boundaries. We also adopt CORA's use of 15 reference passages, along with the question and language tag, as the input of our sequence-to-sequence answer generation model.

6 Results

Tables 1 and 2 report our primary and secondary metrics for each individual language, as well as the overall macro average, on the XOR-QA TYDI and MKQA datasets respectively. Table 2 compares our XOR-QA TYDI results against a set of baseline models, while Table 4 compares our results on the MKQA dataset against those same models. Our comparison models are as follows:

	F1	Em	BLEU
Avg.	21.2	16.0	10.7
Fi (974 exs)	28.0	23.1	21.6
Ja (693 exs)	24.4	19.3	2.6
Ru (1018 exs)	23.5	16.9	11.2
Ar (1387 exs)	22.8	15.5	13.7
Bn (490 exs)	12.4	8.2	8.3
Ko (473 exs)	16.1	12.9	6.9

Table 1: Our results on XOR-TYDI QA

Note: exs = examples

	F1	Em	BLEU
Avg.	16.7	12.6	14.6
Fi (1758 exs)	22.5	19.3	21.7
Ja (1758 exs)	14.7	6.7	5.2
Ru (1758 exs)	15.1	6.5	13.4
Ar (1758 exs)	11.5	7.4	11.4
Ko (1758 exs)	8.6	5.9	5.6
Es (1758 exs)	26.6	23.2	26.0
Sv (1758 exs)	24.8	21.7	24.5
Tr (1758 exs)	21.1	18.0	20.2
Zh_cn (1758 exs)	5.6	5.0	3.2

Table 2: Our results on MKQA

- CORA - The full CORA model from Asai et al. (2021a), with a multilingual text corpus for context retrieval and after the iterative training updates.
- $\text{mDPR}_1 + \text{mGEN}_1$ - CORA with multilingual corpus, but without the iterative training process completed. Our answer generation model is identical to the model used here.
- DPR (trained NQ) + mGEN - CORA but instead of training the retriever on a multilingual retrieval dataset, a multilingual retriever is trained on the english only Natural Questions dataset.
- CORA, C^{multi}={En} - CORA but the model only retrieves English language documents.

We find that our model outperforms all other models that use an English only corpus for information retrieval on the XOR-QA TYDI dataset, but that the

Models	XOR-TYDI QA			
	Avg. F1	Ar	Ja	Te
CORA	31.4	42.6	33.4	26.1
CORA(ii) DPR (trained NQ)+mGEN	24.3	30.7	29.2	19.0
CORA(iii) CORA, C^{multi}={En}	**19.1**	**20.5**	**23.2**	**11.5**
Our model (ColBERT_QA?)	**21.2**	**22.8**	**24.4**	**–**

Table 3: Our results comparing with CORA's model in same setting on XOR-TYDI QA

Models	MKQA					
	Avg. F1	Fi	Ru	Es	Th	Vi
CORA	22.3	25.9	20.6	33.2	6.3	22.6
CORA(ii) DPR (trained NQ)+mGEN	17.9	20.1	16.9	29.4	5.5	18.2
CORA(iii) CORA, C^{multi}={En}	**20.5**	**24.7**	**15.4**	**28.3**	**8.3**	**21.9**
Our model	**16.7**	**22.5**	**15.1**	**26.6**	**–**	**–**

Table 4: Our results comparing with CORA's model in same setting on MKQA

7 Analysis & Conclusion

As previously discussed, there appear to be two dominant strategies in the field of Cross Lingual Information Retrieval (CLIR) – translation methods and multilingual methods. In the translation models, the query is first translated into various languages and then monolingual IR is performed, using either neural or lexical IR methods. The multilingual models by contrast are strictly neural, and involve projecting both the untranslated queries and the passages into a single embedding space. Previous projects have chosen different approaches, and have come to differing conclusions on which is more effective.

In particular, Asai et al. (2021a) and Li et al. (2021) each found that pre-translating the query achieved better results, while Asai et al. (2021b) and Anonymous (2022) achieved higher recall scores by using a single multilingual encoder model. Li et al. (2021) attempt to reconcile the two approaches in DR.DECR by using knowledge distillation to make the results of the multilingual "student" model more similar to the pre-translated "teacher" model, though this process still does not achieve scores higher than the original teacher model.

In general, multilingual question answering systems that utilize translation based IR also tend to perform translation on the answer generation end as well (Asai et al., 2021a). Our approach is unique in that it utilizes translation and a strong monolingual retriever along with a multilingual answer generation model. We find that this approach outperforms CORA on the XOR-QA dataset when retrieving from an English-only corpus, but that CORA achieves much better results when opened up to retrieve from a multilingual corpus. We believe this shows that the translation approach is more effective when dealing with a monolingual corpus and multilingual queries, though it remains to be seen whether our mixed translation-generation approach can be effectively extended to multilingual corpora.

Known Project Limitations

A major limitation of our approach is that it relies on the availability of existing models for translating queries into English. While we feel that there are some benefits to using existing translation models, mainly due to the time and cost saved by not needing to train dataset-specific models, a major downside is that translation models may not exist for all of the necessary language pairs. For example, we were unable to locate reliable publicly avaialble translation models for Telugu-English and Khmer-English, despite having queries for those languages in our datasets. This may also have a biasing effect on our results, as it is likely that any models we did train on these language pairs would offer worse performance than our reported average.

Authorship Statement

Neal Rakesh Vaidya reimplemented the CORA system with ColBERTv2 as retrieval model instead of DPR and performed the experiments for generating our results. Satita Vittayaareekul implemented the use of Helsinki-NLP/Opus-MT for translating all queries target languages to English as the input of ColBERTv2 and refactor the output of ColBERTv2 for the use of feeding them into mT5. All members of the team conducted the literature review experimental protocol together in the earlier stages of the project. Shahab Mousavi looked into the datasets for the project and contributed to writing the final paper. Shahab unfortunately suffered from a harsh case of COVID starting May 24 and couldn't contribute as much as he would have liked to the implementation of the models due to his condition and modeling portion was done by the time he was back to normal in early June, so he could only contribute to writing the final paper.

References

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021a. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hannaneh Hajishirzi. 2021b. One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. In *NeurIPS*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. *ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT*, page 39–48. Association for Computing Machinery, New York, NY, USA.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Yulong Li, Martin Franz, Md Arafat Sultan, Bhavani Iyer, Young-Suk Lee, and Avirup Sil. 2021. Learning Cross-Lingual IR from an English Retriever.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. *ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction*, page arXiv:2112.01488. arXiv.

Ben Thompson. 2015. Aggregation theory.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. 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. Association for Computational Linguistics.