Global Readiness of Language Technology for Healthcare: What would it Take to Combat the Next Pandemic?

♡Ishani Mondal,*♡Kabir Ahuja,*♡Mohit Jain, ◇Jacki O'Neil, ♡Kalika Bali, ♡Monojit Choudhury
♡Microsoft Research Labs, Bangalore, India
◇Microsoft Africa Research Institute, Nairobi, Kenya
{t-imonda, t-kabirahuja, kalikab, monojitc}@microsoft.com

Abstract

The COVID-19 pandemic has brought out both the best and worst of language technology (LT). On one hand, conversational agents for information dissemination and basic diagnosis have seen widespread use, and arguably, had an important role in fighting against the pandemic. On the other hand, it has also become clear that such technologies are readily available for a handful of languages, and the vast majority of the global south is completely bereft of these benefits. What is the state of LT, especially conversational agents, for healthcare across the world's languages? And, what would it take to ensure global readiness of LT before the next pandemic? In this paper, we try to answer these questions through survey of existing literature and resources, as well as through a rapid chatbot building exercise for 15 Asian and African languages with varying amount of resource-availability. The study confirms the pitiful state of LT even for languages with large speaker bases, such as Sinhala and Hausa, and identifies the gaps that could help us prioritize research and investment strategies in LT for healthcare.

1 Introduction

The world witnessed one of the worst pandemics in early 2020, COVID-19, infecting over 250 million people globally. Scientists and technologists from various fields joined hands, lending support to deal with this global crisis. Language Technology (LT), particularly the conversational agents (aka chatbots), played a crucial role during the pandemic by facilitating correct information dissemination (Li et al., 2020; Maniou and Veglis, 2020) and early disease screening (Judson et al., 2020a; Martin et al., 2020b). Nevertheless, today practically useful chatbots and other benefits of LT are available only in a handful of languages (Joshi et al., 2020). Despite impressive gains made by the Massively Multilingual Transformer based Language Models (MMLM) (Devlin et al., 2019; Lample and Conneau, 2019; Aharoni et al., 2019; Conneau et al., 2020; Xue et al., 2021) on standard NLP benchmark tasks (Pan et al., 2017; Conneau et al., 2018; Yang et al., 2019; Ruder et al., 2021), the real-world implications of such advancements remain largely unexplored. Joshi et al. (2020) has highlighted such a disparity and proposed a language hierarchy that comprises of the languages of world classified into six classes based on their resource-availability. In this hierarchy, Class 5 represents the most resource-rich languages for whom benefits of LT are readily available; and class 0 denotes the most under-resourced languages.

In this paper, we ask the following two questions: (1) Today, in which languages can we build practically useful LT systems, especially chatbots, that could serve as beneficial assistants during the pandemic? (2) How should we prioritize research and resource building investments so that LT is globally ready before the next pandemic?

In order to answer these questions, we review the existing literature and resources on COVID-19 chatbots, and classify them based on the languages they support and the solutions they provide. Quite unsurprisingly, the survey reveals a strong disparity in LT solutions between resource-rich and resource-poor languages. In order to quantify this gap and measure the pandemic-readiness of various languages today, we select 15 Asian and African languages (except English) with various degrees of resource-availability, and attempt to build COVID-19 FAQ bots for them. Since building an end-to-end chatbot is a substantial engineering effort, we scope the problem down to building an intent classifier for these languages, which forms the core of the Natural Language Understanding (NLU) unit. We also experiment with entity recognition for a subset of these languages. Our code and datasets have been made publicly available to foster future research.[1]

*Equal contribution

[1] https://github.com/kabirahuja2431/

Proceedings of the 29th International Conference on Computational Linguistics, pages 4320–4335
October 12–17, 2022.

Our study shows that despite using the best available commercial multilingual chatbot frameworks (e.g., Google Dialogflow[2], Microsoft Bot Framework (MS Bot)[3]), advanced Machine Translation (MT) systems[4], and state-of-the-art massively multilingual language models (mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020)), there is a 20-30% drop in performance for class 0 - 2 languages as compared to English. The drop is large for all the African languages (e.g., Hausa and Somali) and some of the Asian languages (e.g., Marathi and Sinhala). Note that our experiments were limited to languages which are supported by at least one of the chatbot frameworks, MT systems or MMLMs. There are thousands of other languages which are supported by none. We have experimented with one such language (Kikuyu) and observed near random performance. Therefore, we hypothesize that such languages are not at all ready for the next pandemic.

We extrapolate our findings at global scale and construct a global LT readiness map for pandemic-response and healthcare. Based on this map as well as error analysis of the chatbot experiments, we identify a set of research problems as well as resource-prioritization strategies which we believe are key to ensure global LT readiness before the next pandemic. More specifically, the purpose of our work is about comparing the systematic inequalities that exist across different languages while deploying chatbots for emergency situations, as well as showing that certain geographical regions are in a disadvantaged position because even the languages that are spoken by a large portion of their population are ill-supported by current LT. Finally, we provide recommendations on how this gap can be bridged by suggesting investment strategies for building LT systems which is otherwise a tough ethical question.

The rest of the paper is organized as follows: Sec 2 presents the literature survey on LT response to COVID-19, specifically focusing on chatbots built for the pandemic. Sec 3 describes the chatbot building experiments, where in 3.1 we motivate the choice of languages for the experiment, in 3.2 and 3.5 we discuss the intent and entity detection experiments respectively. In Sec 4 we present the global LT readiness map and in Sec 5, we conclude with our recommendations.

2 Literature Survey

In the recent years, NLP for Healthcare has witnessed a major uptake and an impressive volume of work has significantly pushed the research forward by developing sophisticated domain-specific language models (Alsentzer et al., 2019; Lee et al., 2020; Ji et al., 2021). These models have been adopted to serve different axes of healthcare such as patient provider communication (Min et al., 2020; Si et al., 2020), information dissemination (Maniou and Veglis, 2020), and self-care management and therapy (Morris et al., 2018; Kadariya et al., 2019; Park et al., 2019; Kamita et al., 2019). The role of healthcare chatbots becomes crucial along all these axes because of the recent adoption of telehealth technology services (Bhat et al., 2021).

Chatbots have received a considerable interest during the recent COVID-19 pandemic. Due to the worldwide spread and severity of the virus and subsequent global response, we believe that the study of COVID-19 chatbots can provide us an accurate picture of the global-readiness of LT. We surveyed COVID-19 chatbots that are mentioned in the literature and/or deployed in the real-world[5].

2.1 Use-Cases and Technological Support

From the survey, two primary use-cases of COVID-19 chatbots emerge – (1) information dissemination: answering pandemic-related questions asked by the users (Li et al., 2020; Desai, 2021; Prasannan et al., 2020; Mehfooz et al., 2020; Trang and Shcherbakov, 2021), and (2) symptom-screening: assessing risk factors associated with the symptoms provided by the user for quick diagnosis (Ferreira et al., 2020; Martin et al., 2020a; Judson et al., 2020b; Quy Tran et al., 2021). Existing commercial frameworks such as DialogFlow, Watson Assistant and MS Bot have been used primarily for building a majority of these chatbots (Li et al., 2020; Sophia and Jacob, 2021). However, open-source bot frameworks like Rasa (Quy Tran et al., 2021; Nguyen and Shcherbakov, 2021) have also been gaining

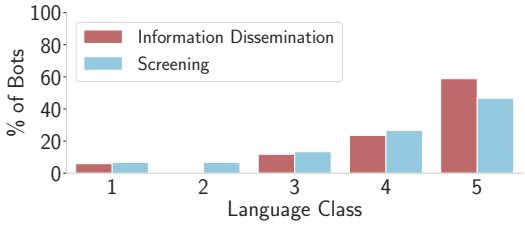
Covid19HealthBots

[2] https://cloud.google.com/dialogflow

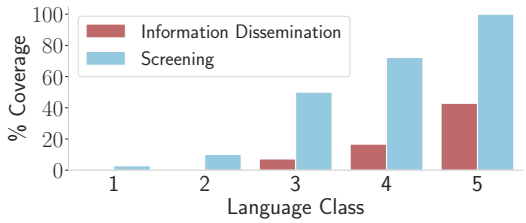
[3] https://dev.botframework.com/

[4] https://translate.google.co.in/ , https://www.bing.com/translator

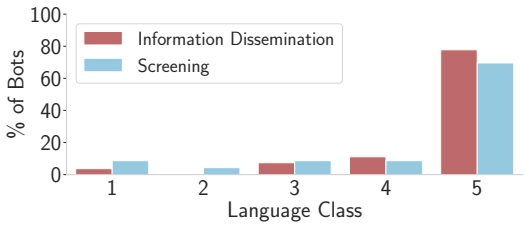
[5] Besides healthcare, NLP has also proved beneficial in providing aid during natural disasters like earthquakes and floods (Lewis, 2010; Rudra et al., 2015; Ghosh et al., 2019; Tsai et al., 2019; Basu et al., 2019). Strassel and Tracey (2016) leveraged existing LT for resource-poor languages to fight against the natural disasters. Though this study is limited to pandemic readiness, we believe the state of LT for disaster-readiness across the globe would be very similar.



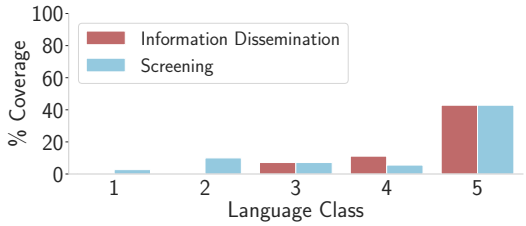
(a) Percentage of papers covering the development of COVID-19 chatbots per language class



(b) Percentage of languages of each class for which at-least one paper addressed the development of a chatbot.



(c) Percentage of bots deployed by governmental agencies , provider organizations etc. per language class, to serve people during the pandemic.



(d) Coverage of languages across different classes in the bots deployed by governmental agencies , provider organizations etc.

Figure 1: Bots developed (for both research and public deployment) for different language classes and their coverage.

traction in the community. The inbuilt NLU engines supported by these frameworks makes chatbot development easy, hence there is a significant uptake in utilizing these to develop new chatbots. Pre-trained LMs were also leveraged for COVID symptom identification (Oniani and Wang, 2020) and question answering (Park et al., 2020).

2.2 Language Diversity

Which languages are supported by these COVID bots? Of the 20 COVID-related bots mentioned in the existing literature and 34 others deployed by different countries to combat the pandemic, 26 ($\approx 50\%$) are exclusively for English, followed by German having 10 deployed bots. In Figure 1, we show the distribution of the chatbots by the language classes defined in (Joshi et al., 2020). As expected, for all the cases, we observe that chatbots were available primarily and almost exclusively for languages in class 5. We observe a slightly higher presence of class 4 and 3 languages in research papers on COVID chatbots (Fig. 1a). For instance, there are three research papers each for Hindi and Vietnamese, both class 4 languages. (Mabrouk et al., 2021) has been recently introduced to help in information dissemmination about covid in African languages. To the best of our knowledge, we could not find any publication or deployed bot for class 0 languages. This skew is more prominent when we consider the coverage of languages of different classes, i.e., the fraction of languages in each class for which at least one COVID-19 conversation system was developed (Figure 1b and 1d). This lack of attention to a large number of languages has also been highlighted by Anastasopoulos et al. (2020) who strongly advocated for the development of language resources for improving access to COVID-19 related information in 26 lesser-resourced languages, particularly from Africa and South and South-East Asia.

3 Rapid Chatbot Building Exercise

How quickly can one build a pandemic response chatbot in a language based on the best publicly available systems? In order to answer this question we have to understand the pandemic-readiness of various languages. To do this, we made an attempt to build chatbots for answering frequently asked questions about COVID-19 using Google Dialogflow, Microsoft Bot Framework (MS Bot), as well as two of the most popular Massively Multilingual Language Models (MMLM) – mBERT and XLM-R. Since building an end-to-end chatbot is complex, we chose to conduct rapid prototyping experiments for intent recognition in 16 languages, and entity recognition in 3 languages.

3.1 Language Selection Criteria

For our experiments, we chose a few languages from each language class (Joshi et al. (2020)) such that at least one language per class is supported by either of the two commercial chatbot frameworks, leading to the set: *English, Chinese* from Class 5, *Hindi, Korean* from Class 4, *Bengali, Malay* from Class 3, *Swahili, Hausa, Marathi, Amharic, Zulu* from Class 2, *Assamese, Gujarati, Kikuyu, Somali*

from Class 1, and *Sinhala* from Class 0.

3.2 Intent Recognition

Intent Recognition is an essential component of conversational systems. Given a user query, the task is to classify it into one of the pre-defined intent categories (Braun et al., 2017).

3.2.1 Dataset Creation and Characteristics

For training and evaluation, we curate a set of 147 queries categorized into one of the 14 intents: 1) `Airborne` (how COVID spreads by air), 2) `ClarifyCovid` (difference between COVID and other diseases), 3) `Country` (country-wise infection statistics), 4) `CovidTwice` (possibility of reinfection), 5) `ExplainSymptom` (COVID symptoms) 6) `Incubation` (how many days of incubation required), 7) `Length` (longevity of infection), 8) `Mask` (ways of wearing mask), 9) `Protection` (ways to protect against infection), 10) `Quarantine` (quarantine requirement of US), 11) `Spread` (how COVID spreads), 12) `Testing` (available COVID tests), 13) `Medication` (about drugs to protect from COVID), and 14) `Treatment` (about treatments or therapies related to COVID). Examples and definitions of each intent are present in Table 3.3.

We refer to the FAQs provided by the UN (Department of Operational Support, 2020) and user queries in the dataset released by Anastasopoulos et al. (2020), to identify the 14 types of questions that a user may ask. We manually paraphrase the questions to generate queries (Mean = 10.5, S.D. = 4.36 queries per intent) for each intent in English. Two annotators with native English proficiency independently classified these queries; the inter-annotator agreement (κ) was 0.89. We asked a few native speakers of each of the selected languages to translate these 147 queries manually. The dataset is split into train and test sets, using a stratified split over the intents, giving a total of 76 and 71 queries in train and test set respectively.

3.2.2 Strategies of Developing Chatbots

We consider three training and inference strategies, emulating the possible scenarios for developing such chatbots in practical settings (Table 3.4).

Train on English Data: In this strategy, we develop our bots by training them on the English queries, and evaluate the intent detection performance in different languages by automatically translating the test queries into English (e.g., similar to Gupta et al. (2021)).

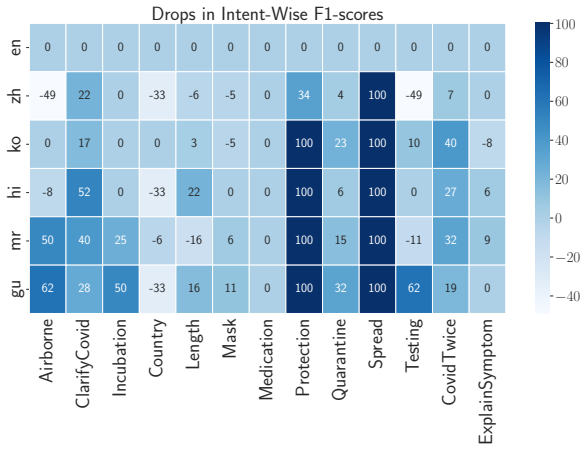


Figure 2: Relative drops (relative to English) in intent wise F1 scores for different languages in the *Train on Manual Translations* setup (in LUIS). Negative values indicate increase in the scores relative to English.

Train on MT Translations: Here we build target language intent classifier models from training data in different languages, which is obtained by automatically translating the English training data. The classifier is then tested on the manually translated test data in the corresponding target language. A similar method was adopted by Balahur and Turchi (2012) for sentiment analysis.

Train on Manual Translations: In this setup, we use the manual translations of the English training dataset to train our bots in different languages. Like the previous setups, here again we use the manually translated data to evaluate the intent detection performance of the developed chatbots. Jennifer Bot (Li et al., 2020) used a similar setup to extend their English bot to Spanish. Note that this is the most expensive setup in terms of data creation cost.

3.3 Intent Definitions and Descriptions

The different intents used for our experiments are described in table 5. We provide definitions and examples for each of the different intents used.

3.4 Bot Building Strategies

3.4.1 Experimental Setup

Commercial Frameworks: We use Google Dialogflow and MS Bot Framework to train and evaluate the FAQ bots in different languages. For Dialogflow we use the ES Console, and for MS Bot, we use Microsoft's Language Understanding

Intent Type	Example in English	Definition
Airborne	Can the virus that causes COVID-19 be transmitted through the air?	Queries related to how much COVID is carried by air
ClarifyCovid	How do I know if it is COVID-19 or just the flu?	Queries related to difference betwen COVID and other diseases
Incubation	Can someone in incubation infect other people?	Queries related to situations where a person is infected with COVID and is going through incubation phase
Length	How long does the illness make you poorly for?	Queries regarding longevity of COVID infection
Mask	Should I wear a mask while exercising?	Queries about wearing mask
Protection	Ways to keep safe from COVID-19	Queries about the ways of protection from COVID
Quarantine	Will I avoid coronavirus, if I self-isolate?	Queries about the effect of quarantining after getting infected with COVID
Spread	Aside from inhalation, are there other ways coronavirus can spread?	Queries about the spreading process of COVID
Testing	Where can I get my test done?	Queries about the testing process of COVID
CovidTwice	If you get COVID-19, can you get it again?	Queries about whether COVID can infect someone more than once
ExplainSymptom	I have a sharp pain here in the chest	User explaining COVID related symptoms
Country	How many people in Italy have COVID-19?	Querying about the statistics of infection in different countries
Medication	Do any of the drugs reduce mortality?	Querying about the medication to survive from COVID
Treatment	Which vaccines are good to protect against the virus?	Querying about the treatment strategies associated with COVID

Table 1: Different intents with definitions and examples present in our dataset.

Bot Building Setup	Training Strategy	Testing Strategy
Train on English Data	Train set comprises of the English queries	Test set comprises of English queries where the manually written queries in target language are translated to English using MT system
Train on MT Translations	Train set comprises of the English queries translated to target language using MT System	Test set comprises of manually written queries in target language
Train on Manual Translations	Train set comprises of manually written queries in target language	Test set comprises of manually written queries in target language

Table 2: Different strategies for building the chatbots.

Service (LUIS)[7] framework. Dialogflow and LUIS supports 7 and 6 out of our 16 selected languages, respectively. For the unsupported languages, we could only experiment with *Train on English Data*.

Pre-trained MMLMs: We evaluate two popular MMLMs, namely mBERT (*bert-base-multilingual-cased*) and XLMR (*xlm-roberta-base*), for our intent detection experiments. XLM-R supports all but Kikuyu, Somali and Zulu, while mBERT supports all but Amharic, Assamese, Hausa, Kikuyu and Zulu. For these models, we only evaluate the *Train on Manual Translations* setup. We experiment with two different approaches for building intent classifiers with these models: i) Using k-Nearest Neighbors on the sentence embeddings obtained through the MMLM to classify the intents as done in Caron et al. (2021), ii) Training an end-to-end classifier by fine-tuning the pre-trained MMLM. We report the best scores out of these two setups for both MMLMs (details in A.3).

Evaluation: We report the relative accuracy drop δ_l for each target language l from English (en), defined as $(A_{en} - A_l)/(A_{en}) \times 100$, where A_l is the accuracy of intent classification for l on the held-out test set[8]. Thus, lower the value of δ_l, better is the state-of-the-art of LT for the language l.

3.4.2 Results and Analysis

Table 3 presents the intent classification results which reports the relative drop of the model's accuracy with respect to English. While the relative drop δ_l is reported, we also mark the values with a † wherein the *absolute* accuracy, A_l falls below 67%. We use this as a minimum viable threshold of

[7] https://www.luis.ai/

[8] Absolute accuracies are not reported since we do not intend to compare the performances of commercial frameworks.

Class	Languages	Train on English Data		Train on MT Translations		Train on Manual Translations			
		DF	LUIS	DF	LUIS	DF	LUIS	mBERT	XLM-R
5	Chinese	0.60	5.00	17.50†	18.40†	0.04	(5.20)	(5.63)	(8.50)
4	Hindi	12.50	0.05	16.25†	25.01†	13.02	10.50	12.72†	19.15†
	Korean	6.50	13.71†	31.20†	11.55	23.75†	10.00	5.63†	10.88†
3	Bengali	20.50†	13.15†	26.25†	×	11.80	×	7.04†	2.13†
	Malay	21.24†	11.87	19.53†	×	12.50	×	4.77	14.89†
2	Swahili	28.08†	18.00†	×	×	×	×	32.39†	19.15†
	Hausa	40.97†	34.00†	×	×	×	×	×	29.79†
	Marathi	21.23†	14.00†	28.08†	28.7†	16.25†	17.76†	16.90†	29.79†
	Amharic	43.06†	34.82†	×	×	×	×	×	12.39†
	Zulu	30.56†	11.28	×	×	×	×	×	×
1	Assamese	19.52†	18.00†	×	×	×	×	×	29.79†
	Gujarati	15.55	10.03	×	22.88†	×	22.88†	4.77	19.15†
	Kikuyu*	97.60†	76.87†	×	×	×	×	×	×
	Somali	40.56†	27.58†	×	×	×	×	25.35†	×
0	Sinhala	35.00†	19.00†	34.93†	×	15.65	×	61.97†	19.15†

Table 3: δ_l for each language for the *Intent Recognition* task using the three different strategies. × indicates that the framework does not support end-to-end chatbot development for that language. Drops that lead to accuracy below 67% are marked by †, indicating the case where the bot mis-recognizes 1 out of every 3 queries. *Owing to non-availability of standard MT for Kikuyu, we used *Safarini*[6] app from Android playstore for translation. Note: The values mentioned in the parantheses indicate that we observe relative gain instead of drop.

Lang	Issue	Actual Example	Misclassified Translated Example
Si	**Terminology Mismatch**	I have hay fever though too.	I also have gonorrhea.
Bn	**Fluency**	Is SARS-CoV-2 airborne?	Does SARS-CoV-2 sit in the air?
Hu	**Relevance**	I got the virus. How long does it go on for?	I Nasami Cutar. How long will it take?
Hu	**Fluency, Terminology Mismatch**	How long should I wear a mask?	How long will I impose sanctions?
Hu	**Terminology Mismatch**	Is it healthy to wear a mask during swimming?	Is it safe, can I wear fascist sanctions when I swim?

Table 4: Excerpts of test instances showing bottlenecks of MT systems in the *Train on English Data* setup.

the performance, as below this the model will misclassify more than 1 out of every 3 queries which might not be useful for real world deployments. As expected, we observe high δ_l for languages belonging to class 3 or lower, with most of the accuracies below the acceptable limit.

Comparison across the three setups: We observe that for classes 4 and 5, *Translate on English Data* performs at par or even better than the most expensive *Train on Manual Translations* setup. This may be because the MT translations from these languages to English is highly accurate. On the other hand, for languages belonging to class 3 or lower, *Train on Manual Translations* led to better performance, arguably due to poorer performance of the MT system. Unfortunately, the *Train on Manual Translations* method is the most expensive in terms of data curation cost, hence may be the hardest to implement in the midst of a pandemic. The problem becomes worse because a majority of class 3 and lower languages are not supported by current chatbot frameworks. Even when supported, their performance is below the acceptable limit (e.g., Marathi, Gujarati). One of the reasons is the difficulty in correctly identifying technical intents like `Airborne` and `Incubation` in such low-resource languages (Figure 2). Since a few of these low-resource languages are present in the pre-training dataset of mBERT and XLM-R, we can evaluate them for *Train on Manual Translations*. There is a similar pattern in accuracy drop for MMLMs, however the accuracy begin to fall below the acceptable limit (67%) from class 4 languages onward. There is a remarkable drop in mBERT's accuracy for Sinhala (class 0). In general, we

find mBERT to outperform XLM-R, except for Swahili and Sinhala. This may be due to the better representation of these languages in the pre-training corpus of XLM-R (CommonCrawl Corpus). This strongly indicates the importance of the pre-training dataset size for developing LT, both in terms of absolute size as well as relative size to other languages (Wu and Dredze, 2020). As expected, the performance in *Train on MT Translations* setup is the worst among the three; except for Korean in LUIS, all values lie below the acceptable limit, which could be a compounded effect of poor translation quality and inferior NLU solutions. To conclude, all languages in class 3-5 had at least one solution yielding an acceptable accuracy, while all languages in class 0-2, except Gujarati, Sinhala and Zulu, had no acceptable solution.

Lost in Translation: Table 4 shows the intent misclassification errors due to the errors in MT translations. The manual translation in the target language correspond to the 'Actual Example' in English, and the phrase translated back to English for the *Train on English Data* setup is reported under the 'Misclassified Translated Example.' We categorize the translation errors as Terminology Mismatch, Fluency and Relevance (Li et al., 2020). We find that domain-specific terms often get translated incorrectly into English (4). In a few cases, the translations result in unnatural queries resulting in loss of fluency, such as *Does SARS-CoV-2 sit in the air?*. All these factors lead to poor performance of *Train on English Data* setup for low-resource languages. We find that *Terminology Mismatch* is the most common issue affecting the performance[9]. Interestingly, technical terms like *incubation* does not exist in a few of our target languages, hence the manually written test queries in these languages just had the English term written in that language's script. In such cases, we found lesser performance drop compared to languages when equivalent vocabulary exists in the target language. E.g., high drops in F1-score for intents like "`Quarantine`" and "`Incubation`" in Hausa (76%, 100% respectively) and Amharic (56%, 100%) justify this, whereas for Zulu, where the human translator used English terms in their queries resulted in much lower drop in F1 scores (20%, 0%). See appendix for the intent-wise F1 scores for different languages.

Implications: Based on our experimental results, we wish to explore how to prioritize the resource-investment strategies to push the state of current LT forward. Resource-poor languages mostly underperform across all the three set-ups, so then *should we invest more towards developing better translation systems or focus more on improving the current NLU solutions for different languages?* We observe that a good quality translation system can support building bots from scratch in a new language, and often performs on-par with the *Train on Manual Translations* setup for high-resource languages (e.g., Korean, Hindi) and sometimes even for low-resource languages (e.g., Gujarati). Building bots from scratch in a new language is resource-intensive, requiring rapid prototyping, which may be infeasible during a crisis since massive data collection efforts need to be made. Therefore, a generic way to ensure pandemic-readiness in a language is by ensuring reasonably accurate MT systems similar to that for class 4 languages. Improving representation of low-resource languages in the pre-training datasets of existing multilingual models (specifically on domain-specific corpora as done by (Gu et al., 2021; Zhang et al., 2020)) is yet another way to ensure preparedness, as it can lead to improved performance of the MMLMs on these languages (Wu and Dredze, 2020). Unlabelled language data for low resource languages can also be leveraged to build Machine Translation systems in these languages when used in conjunction with the parallel data in high resource languages for training massively multilingual models. (Siddhant et al., 2022; Bapna et al., 2022).

3.5 Entity Recognition

We also evaluate the developed chatbots on another core task of NLU, i.e. entity recognition (Ali, 2020) on English, Hindi and Bengali. To train and evaluate different COVID bots on this task, we use a set of 200 user related queries (obtained by augmenting existing dataset of 147 queries). Entity types were identified from a subset of labels from the

[9]While investigating the correlation between the translation quality (BLEU (Papineni et al., 2002)) between the reference original English query and the back-translated English string from each manually written query in target language) and the intent classification results, we did obtain a positive spearman correlation coefficient with the value of 0.36. and a p-value 0.21. With a p-value of 0.21 the correlation is not statistically significant which we suspect might be due to BLEU not being a great measure of translation quality (Sulem et al., 2018) to measure the subtleties discussed above.

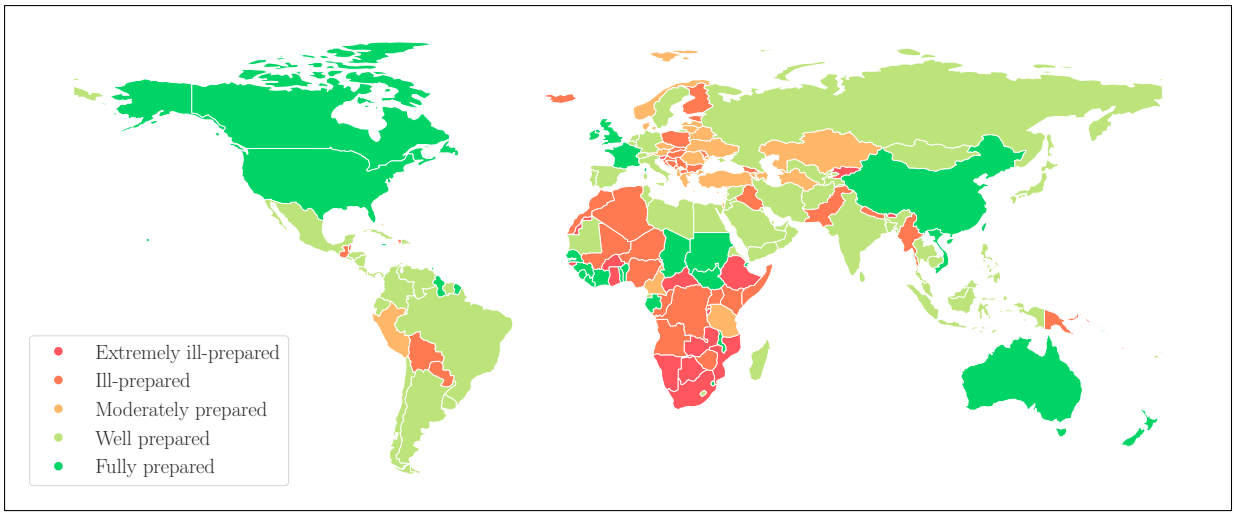


Figure 3: World map showing the Readiness of each country in terms of fighting the next pandemic using LT.

CORD-19 NER dataset (Lu Wang et al., 2020), and the queries were accordingly tagged by two native speakers of Bengali and Hindi. Overall, our dataset had a mix of *medical* and *non-medical* entities. The final set of *medical* entity types consists of: `Covid` (COVID-related entities), `PhysicalScience` (technical terms related to bio-molecular mechanism of the disease), and `Disease` (any form of illness or symptoms). The *non-medical* entity types are: `BodyPart` (name of the body part), `Country` (country name), `Duration` (length in days), `Protection` (ways to protect against COVID, such as 'mask', 'gloves'), and `InfoSource` (source of information). `Country` (country name). For generating the equivalent translations, we manually aligned the entity tags in two languages: Hindi (supported by DialogFlow and LUIS) and Bengali (supported by DialogFlow). In majority of the cases, we observe that domain-specific entities such as `incubation, ACE-2 Cells, biochemical assays` are hard to predict by these models on languages other than English. For instance, for `Covid` entity, we observe significant F1-score drop of 24.6% for Hindi and 42.9% for Bengali. However, for non-medical entities, these models were found to perform comparatively better, e.g., drop in F1-score on `Country` tag was 5.2% for Hindi and 8.9% for Bengali.

4 Measuring Global Readiness

Although our current work focuses on analyzing pandemic-preparedness of only 16 languages, here we try to generalize our findings to other languages by introducing a *Readiness Score* for every language which empirically measures the preparedness of current LT to serve its speakers in a pandemic-like emergency situation. The definition of readiness is based on the assumption that one has access to the best available LT by considering the highest intent detection accuracy of A_l^* for a language across different frameworks and training setups. We then define the readiness of a language l as its relative accuracy with respect to English as $r_l = \frac{A_l^* - A_{random}}{A_{en}^* - A_{random}}$, where A_{en}^* denotes the best case accuracy on English, and A_{random} is the accuracy of a random classifier: $A_{random} = 100/numberOfIntents$.

We would like to interpolate r_l for all the languages of the world, and hence would need more training examples than the 16 languages that we currently have. We select a set of 11 proxy languages (details in Appendix A.5). This has been done in order to ensure the coverage of the features of major language families in the world[10] while training the model. For these languages, we compute *proxy* accuracies $\tilde{A}_l^*$ by building and evaluating chatbots on MT translated data. We then train a Gaussian Process Regression model for predicting readiness scores with the r_l values for the 27 languages as our training set. We use geographical and genetic features from the URIEL database (Littell et al., 2017) to represent the languages. The predictive model, which has an average absolute prediction error of 5%, is then used to estimate the readiness scores of 116 new languages supported by major MMLMs (mBERT and XLM-R) and/or translators (Google and Microsoft). For all

[10] *Ethnologue 24 (2021)*: `https://en.wikipedia.org/wiki/List_of_language_families`

other languages, we set $r_l = 0$, as one can expect near random performance without any LT, as we did see for Kikuyu (Table 3). The estimated final r_l scores for all the languages were used to extrapolate the pandemic-readiness of each country c, as follows. We use the country-wise language and speaker demographic data[11] to calculate the country-wise readiness (similar to Blasi et al. (2021)), $r_c = \sum_{l \in \mathcal{L}_c} s_{c,l} r_l$, where $\mathcal{L}_c$ is the set of languages spoken in country c, and $s_{c,l}$ is the fraction of c's population forming native speakers of the language l. The r_c values were clustered to generate five classes (Extremely ill-prepared: 0-0.33, Ill-prepared: 0.33-0.74, Moderately prepared: 0.74-0.83, Well prepared: 0.83-0.92, Fully prepared: 0.92-1) using Jenks' natural breaks optimization (Jenks, 1967). These classes were used to generate a readiness heatmap of the world (Fig 3).

Observations: From Figure 3, one can observe that South and East African countries are Extremely ill-prepared, due to the high dominance of low-resource languages. For instance, people in Zambia's speak Bemba, Chewa and Luzi, all of which are severely under-resourced. As pointed out in Anastasopoulos et al. (2020), these regions might also be worse-hit in a pandemic situation, and therefore, require immediate attention. For Ill-prepared regions such as Bolivia in South America, and Guatemala in Latin America, r_l values are slightly better due to the abundance of Spanish speakers, however there is a sizeable population speaking under-served languages such as Q'eqchi and Guarani. Countries that fall within fully to moderately prepared categories typically have large native speaker population of one or more of the class 5 languages (English, French, Chinese, Arabic) and/or well-supported languages (e.g., Korean, Bengali, Malay). It is important to note that while approximating readiness of a language, we assumed same value for all its diverse linguistic variants and dialects, which in certain cases results in overestimation of r_c. High r_c for north and central African countries (e.g., Libya, Egypt and Sudan) might be due to sizeable population of a resource-rich language Arabic. However, Arabic has several dialects, which vary from the Modern Standard Arabic at various linguistic levels, and consequently the performance of LT systems for such dialects also vary considerably (Zbib et al., 2012; Alsharhan and Ramsay, 2020). It holds true for Spanish and Portuguese spoken in Latin America (Lipski, 2014) and French dialects of Western Africa.

5 Conclusion and Recommendation

From our chatbot development experiences, we uncover a set of interesting insights to arrive at the following recommendations which can improve the state of preparedness of languages to develop useful technologies during the next pandemic.

— Our experiments showed that low-resource Indian languages (such as Marathi, Bengali) were benefited due to the presence of a geographically and/or linguistically closely related well-resourced language (Hindi). This notion of such "bridge" languages has been explored before in the context of MT (Paul et al., 2013) and zero/few-shot transfer in MMLMs (Lauscher et al., 2020). We recommend the community to target bridge languages for the regions that are currently poorly prepared from an LT perspective.

— Drawing insights from the brittleness of MT for domain-specific terms (*airborne, incubation*) or newly-coined terms (*COVID*), we believe that commercial and open-source bot frameworks can benefit from domain adaptation techniques (Chu and Wang, 2018), or techniques to inject new terms to existing solutions.

Our study confirms that except English, only a few European and Asian languages push forward the state-of-the-art research in LT for healthcare. Our preliminary investigation suggests that instead of demographic demand, it is the economic prowess of the users of a language that drives the investment towards developing sophisticated LT solutions for a given language. For instance, Swahili, even though considered as the *lingua franca* of Africa, is still under-served by commercial chatbot frameworks. Similar trends were observed for Hausa which has a considerably large speaker base compared to Dutch (resource-rich)[12].

We believe that these findings will play a crucial role in making the community aware of the disparity that needs to be addressed before the next pandemic hits.

[11] Infoplease Languages Spoken in Each Country of the World: https://bit.ly/3HoAs9K

[12] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Nazakat Ali. 2020. Chatbot: A conversational agent employed with named entity recognition model using artificial neural network. *CoRR*, abs/2007.04248.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Eiman Alsharhan and Allan Ramsay. 2020. Investigating the effects of gender, dialect, and training size on the performance of arabic speech recognition. *Language Resources and Evaluation*, 54(4):975–998.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Alexandra Balahur and Marco Turchi. 2012. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 52–60, Jeju, Korea. Association for Computational Linguistics.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Moumita Basu, Anurag Shandilya, Prannay Khosla, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. Extracting resource needs and availabilities from microblogs for aiding post-disaster relief operations. *IEEE Transactions on Computational Social Systems*, 6(3):604–618.

Karthik S Bhat, Mohit Jain, and Neha Kumar. 2021. Infrastructuring telehealth in (in)formal patient-doctor contexts. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic inequalities in language technology performance across the world's languages.

Daniel Braun, Adrian Hernandez Mendez, Florian Matthes, and Manfred Langen. 2017. Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185, Saarbrücken, Germany. Association for Computational Linguistics.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP 2018*, pages 2475–2485.

UN Department of Operational Support. 2020. Covid-19 frequently asked questions.

Sharmishta Suhas Desai. 2021. Chatbot for covidvaccine using deep learning.

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.

Thiago Castro Ferreira, Milena Soriano Marcolino, Isaias Ramos, Raquel Oliveira Prates, Leonardo B. Ribeiro, Zilma Reis, Adriana Silvina Pagano, Wagner Meira, and Antônio Luiz Pinho Ribeiro. 2020. Ana: A brazilian chatbot assistant about covid-19.

Promila Ghosh, M. Raihan, Md. Tanvir Islam, and Md. Ekhlasur Rahaman. 2019. Safeguard: A prototype of an application programming interface to save the disaster affected people. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Ankur Gupta, Yash Varun, Prarthana Das, Nithya Muttineni, Parth Srivastava, Hamim Zafar, Tanmoy Chakraborty, and Swaprava Nath. 2021. Truthbot: An automated conversational tool for intent learning, curated information presenting, and fake news alerting.

George F Jenks. 1967. The data model concept in statistical mapping. *International yearbook of cartography*, 7:186–190.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and E. Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *ArXiv*, abs/2110.15621.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Timothy J Judson, Anobel Y Odisho, Jerry J Young, Olivia Bigazzi, David Steuer, Ralph Gonzales, and Aaron B Neinstein. 2020a. Implementation of a digital chatbot to screen health system employees during the COVID-19 pandemic. *Journal of the American Medical Informatics Association*, 27(9):1450–1455.

Timothy J. Judson, Anobel Y. Odisho, Jerry J Young, Olivia Bigazzi, David J. Steuer, Ralph Gonzales, and Aaron B. Neinstein. 2020b. Implementation of a digital chatbot to screen health system employees during the covid-19 pandemic. *Journal of the American Medical Informatics Association : JAMIA*, 27:1450 – 1455.

Dipesh Kadariya, Revathy Venkataramanan, Hong Yung Yip, Maninder Kalra, Krishnaprasad Thirunarayanan, and Amit Sheth. 2019. kbot: Knowledge-enabled personalized chatbot for asthma self-management. In *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 138–143.

Takeshi Kamita, Tatsuya Ito, Atsuko Matsumoto, Tsunetsugu Munakata, and Tomoo Inoue. 2019. A chatbot system for mental healthcare based on sat counseling method. *Mobile Information Systems*, 2019:9517321.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *NeurIPS*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. 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. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240.

William Lewis. 2010. Haitian Creole: How to build and ship an MT engine from scratch in 4 days, 17 hours, & 30 minutes. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation*, Saint Raphaël, France. European Association for Machine Translation.

Yunyao Li, Tyrone Grandison, Patricia Silveyra, Ali Douraghy, Xinyu Guan, Thomas Kieselbach, Chengkai Li, and Haiqi Zhang. 2020. Jennifer for COVID-19: An NLP-powered chatbot built for the people and by the people to combat misinformation. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

John M. Lipski. Castile and the hydra: the diversification of spanish in latin america.

John M. Lipski. 2014. *2. The Many Facets of Spanish Dialect Diversification in Latin America*, pages 38–75. University of Chicago Press.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya

Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*, page arXiv:2004.10706v2. 32510522[pmid].

Aymen Ben Elhaj Mabrouk, Moez Ben Haj Hmida, Chayma Fourati, Hatem Haddad, and Abir Messaoudi. 2021. A multilingual african embedding for faq chatbots. *ArXiv*, abs/2103.09185.

Theodora A. Maniou and Andreas Veglis. 2020. Employing a chatbot for news dissemination during crisis: Design, implementation and evaluation. *Future Internet*, 12(7).

Alistair Martin, Jama Nateqi, Stefanie Gruarin, Nicolas Munsch, Isselmou Abdarahmane, and Bernhard Knapp. 2020a. An artificial intelligence-based first-line defence against covid-19: digitally screening citizens for risks via a chatbot. *bioRxiv*.

Alistair Martin, Jama Nateqi, Stefanie Gruarin, Nicolas Munsch, Isselmou Abdarahmane, Marc Zobel, and Bernhard Knapp. 2020b. An artificial intelligence-based first-line defence against covid-19: digitally screening citizens for risks via a chatbot. *Scientific Reports*, 10(1):19012.

Fahad Mehfooz, Sakshi Jha, Sahil Singh, Shreya Saini, and Nidhi Sharma. 2020. Medical chatbot for novel covid-19. *ICT Analysis and Applications*, 154:423 – 430.

Do June Min, Veronica Perez-Rosas, Shihchen Kuo, William H. Herman, and Rada Mihalcea. 2020. Upstage: Unsupervised context augmentation for utterance classification in patient-provider communication. In *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of Machine Learning Research*, pages 895–912. PMLR.

Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. 2018. Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *J Med Internet Res*, 20(6):e10148.

Trang Nguyen and Maxim Shcherbakov. 2021. Enhancing rasa nlu model for vietnamese chatbot. 9:33–36.

David Oniani and Yanshan Wang. 2020. A qualitative evaluation of language models on automatic question-answering for covid-19. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '20, New York, NY, USA. Association for Computing Machinery.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of ACL 2017*, pages 1946–1958.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ho-yeon Park, Gun-doo Moon, and Kyoung-jae Kim. 2020. Classification of covid-19 symptom for chatbot using bert. *Solid State Technology*, 63(6):19185–19188.

SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, and Bongwon Suh. 2019. Designing a chatbot for a brief motivational interview on stress management: Qualitative case study. *J Med Internet Res*, 21(4):e12231.

Michael Paul, Andrew Finch, and Eiichrio Sumita. 2013. How to choose the best pivot language for automatic translation of low-resource languages. *ACM Transactions on Asian Language Information Processing*, 12(4).

Praveen Prasannan, Stephy Joseph, and Rajeev R R. 2020. A chatbot in Malayalam using hybrid approach. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): System Demonstrations*, pages 28–29, Patna, India. NLP Association of India (NLPAI).

Ban Quy Tran, Thai Van Nguyen, Thang Duc Phung, Viet Tan Nguyen, Dat Duy Tran, and Son Tung Ngo. 2021. Fu covid-19 ai agent built on attention algorithm using a combination of transformer, albert model, and rasa framework. In *2021 10th International Conference on Software and Computer Applications*, ICSCA 2021, page 22–31, New York, NY, USA. Association for Computing Machinery.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation.

Koustav Rudra, Subham Ghosh, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh. 2015. Extracting situational information from microblogs during disaster events: A classification-summarization approach. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 583–592, New York, NY, USA. Association for Computing Machinery.

Shijing Si, Rui Wang, Jedrek Wosik, Hao Zhang, David Dov, Guoyin Wang, and Lawrence Carin. 2020. Students need more attention: Bert-based attention model for small data with application to automatic patient message triage. In *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of Machine Learning Research*, pages 436–456. PMLR.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *CoRR*, abs/2201.03110.

J.Jinu Sophia and T.Prem Jacob. 2021. Edubot-a chatbot for education in covid-19 pandemic and vqabot comparison. In *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1707–1714.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Nguyen Thi Thu Trang and Maxim Shcherbakov. 2021. Enhancing rasa nlu model for vietnamese chatbot.

Meng-Han Tsai, James Yichu Chen, and Shih-Chung Kang. 2019. Ask diana: A keyword-based chatbot system for water-related disaster management. *Water*, 11(2).

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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

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.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of EMNLP 2019*, pages 3685–3690.

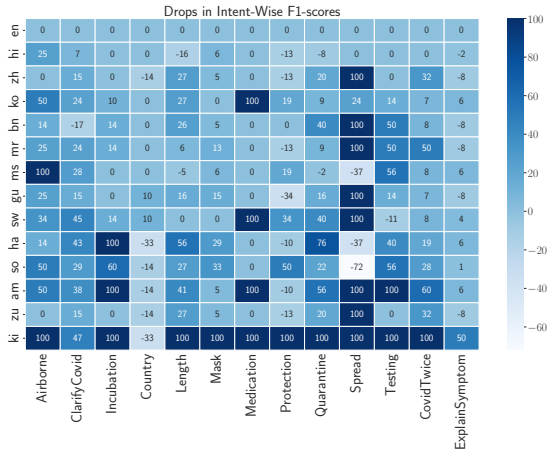


Figure 4: Relative drops (relative to English) in intent wise F1 scores for different languages in the *Train on English* setup (in LUIS). Negative values indicate increase in the scores relative to English.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.

Rong Zhang, Revanth Reddy Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avirup Sil, and Todd Ward. 2020. Multi-stage pretraining for low-resource domain adaptation. *ArXiv*, abs/2010.05904.

A Example Appendix

A.1 Intent Definitions and Descriptions

The different intents used for our experiments are described in table 5. We provide definitions and examples for each of the different intents used.

A.2 Bot Building Strategies

A.3 MMLM Training Setup

For our experiments with Multilingual Pre-trained Transformers we consider mBERT (*bert-base-multilingual-cased*) and XLMR (*xlm-roberta-base*) for training intent classifiers. As mentioned in the main text we explore two methodologies to train and evaluate these MMLMs, a detailed description with hyperparameters is given below:

1. KNN using Pre-trained Embeddings:
Since the scale of our data is on the lower side, training an end-to-end classifier might be prone to

Intent Type	Example in English	Definition
Airborne	Can the virus that causes COVID-19 be transmitted through the air?	Queries related to how much COVID is carried by air
ClarifyCovid	How do I know if it is COVID-19 or just the flu?	Queries related to difference betwen COVID and other diseases
Incubation	Can someone in incubation infect other people?	Queries related to situations where a person is infected with COVID and is going through incubation phase
Length	How long does the illness make you poorly for?	Queries regarding longevity of COVID infection
Mask	Should I wear a mask while exercising?	Queries about wearing mask
Protection	Ways to keep safe from COVID-19	Queries about the ways of protection from COVID
Quarantine	Will I avoid coronavirus, if I self-isolate?	Queries about the effect of quarantining after getting infected with COVID
Spread	Aside from inhalation, are there other ways coronavirus can spread?	Queries about the spreading process of COVID
Testing	Where can I get my test done?	Queries about the testing process of COVID
CovidTwice	If you get COVID-19, can you get it again?	Queries about whether COVID can infect someone more than once
ExplainSymptom	I have a sharp pain here in the chest	User explaining COVID related symptoms
Country	How many people in Italy have COVID-19?	Querying about the statistics of infection in different countries
Medication	Do any of the drugs reduce mortality?	Querying about the medication to survive from COVID
Treatment	Which vaccines are good to protect against the virus?	Querying about the treatment strategies associated with COVID

Table 5: Different intents with definitions and examples present in our dataset.

Bot Building Setup	Training Strategy	Testing Strategy
Train on English Data	Train set comprises of the English queries	Test set comprises of English queries where the manually written queries in target language are translated to English using MT system
Train on MT Translations	Train set comprises of the English queries translated to target language using MT System	Test set comprises of manually written queries in target language
Train on Manual Translations	Train set comprises of manually written queries in target language	Test set comprises of manually written queries in target language

Table 6: Different strategies for building the chatbots.

over-fitting. We fit a k-Nearest Neighbors (KNN) classifier on the sentence embeddings obtained using the pre-trained model for the queries in training data. At test time, we similarly obtain the representation of the user query and find its nearest neighbors among the training queries to predict its intent. The optimal value for k was empirically found to be 1 and for sentence embeddings, we take the average of the representation of each token of the sentence in the last layer of MMLM.

We also tried fine-tuning the pre-trained model with the training queries using a Masked Language Modelling (MLM) objective. Additionally, we also fine-tuning on a much larger COVID-19 queries dataset in english : COQB (Li et al., 2020) along with our training queries, as has been pointed by Lauscher et al. (2020) can be an effective strategy for few shot transfer. We use 3 epochs to fine-tune the models with a learning rate of 5e-5 and Adam-W optimizer (Loshchilov and Hutter, 2019). A masking probability of 15% was used during the MLM training and maximum sequence length was taken to be 32.

2. Fine-tuning an End to End Classifier :
We also try fine-tuning the MMLMs end-to-end by adding a classification head on top of the pre-trained network to classify the input query into one of the 14 intents. We adapt the sequence classification scripts for GLUE benchmark (Wang et al., 2018) provided by hugging face[13] on our dataset. We fine-tune the classifier for 20 epochs, with the same learning rate and optimizer as the MLM fine-tuning in the first point with a batch size of 8.

For every language we use the best accuracies

[13] https://huggingface.co/transformers/v2.9.1/examples.html

		Medical			Non-Medical				
Lang	Bot	Covid	PhysicalScience	Disease	BodyPart	Country	Duration	Protection	InfoSource
Hi	DF	24.6	50.1	+3	32	7.52	5.2	11.3	+30.1
	LUIS	24.56	+41	94	21	5.31	+4	12.06	+72.72
Bn	DF	42.9	43.1	52	34	6.3	8.9	20.4	8.34

Table 7: Relative drop in entity-type wise F1-score in *Entity Recognition* task using DialogFlow (DF) and LUIS.

obtain from either of these two strategies [14]. All the experiments were run on 4 NVIDIA V-100 GPUs with 32 GB memory.

A.4 Language Readiness Analysis

Results and Analysis

Initially, we have plotted the readiness measures of each language used in our training data on the scatter plot in Figure 5 with the language class on X-axis and readiness measure in Y-Axis. It clearly shows that the African languages such as `Somali`, `Amharic`, `Hausa`, `Zulu` are below the trend-line in terms of readiness. In fact, some of the European languages such as `Icelandic`, `Hungarian`, `Estonic`, `Finnish` also require some attention. Primarily, we observe that the readiness measure is not a direct function of the language class from this plot. As we can see that even though majority of the class 4 and 5 are near the trend line, the observation is similar for Class 1 as well.

Therefore, we also resort to understanding how much does the trend hold true for the language families of these corresponding languages? So, we approximate each of the language family by taking the average scores of each language falling into that class and plot those in Figure 6. It was interesting to observe that the English-major language families such as `Austroasiatic`, `Koreanic` and `Sino-Tibetan` are well-served, and consequently lie above the trend line. Overall, `Indo-European` language families are well near the trend line and then the resource-poor language families are `Afroasiatic`, `Niger-Congo` and `Uralic`, the worst being the `Afroasiatic` language family.

A.5 Details on Language Readiness Prediction

In section 4, we discussed the estimation of readiness values of different languages. We first extended our 16 languages that we considered for intent recognition experiments with proxy scores for an additional 11 languages, namely, `French (fr)`, `Arabic (ar)`, `German (de)`, `Spanish (es)`, `Portuguese (pr)`, `Vietnamese (vi)`, *Hungarian (hu)*, `Finnish (fi)`, `Czech (cs)`, `Estonian (et)`, and `Icelandic (is)`. Finally, it covers six primary language families in the world, such as: 1) `Indo-European`, 2) `Sino-Tibetan`, 3) `Afroasiatic`, 4) `Niger-Congo`, 5) `Koreanic` and 6) `Austroasiatic`. To estimate the readiness values of the remaining 116 languages supported by the Translators (Google and Bing) and MMLMs (mBERT and XLMR), we used the available readiness data for the 27 to build a regression model. We used Gaussian Processes to model the readiness prediction problem, due its efficiency on the small sized datasets. Radial Basis Function (RBF) with added noise level for each instance (White Kernel), was used, and the *length scale* of RBF and noise level were tuned using L-BFGS algorithm with 5 restarts for the optimizer. The model selection was done using a Leave One Out strategy, where we move one language to validation set and train on the remaining, repeating this for all the languages and measuring average accuracy. Besides Gaussian Process Regression (GPR), we also experimented with Linear Regression, Lasso Regression and XGBoost (Chen and Guestrin, 2016), but observed inferior validation accuracies.

A.6 Global Pandemic Readiness Measurement

In section 4 of the paper, we have talked about how to actually take the speaker-base values into account while calculating the readiness scores for each country in the world and the final r_l scores obtained on all the languages are used to extrapolate the readiness of each country c in the world. We had also experimented in a way such that all the languages spoken in the country are weighted

[14] technically 4, as in the KNN case we consider no fine-tuning, fine-tuning on Train Queries, and fine-tuning on Train and COQB queries

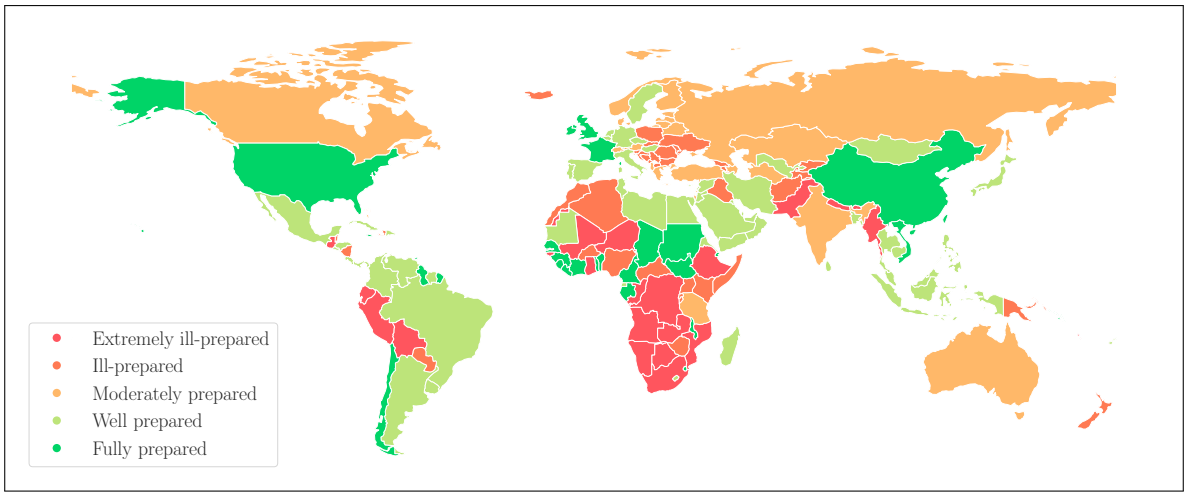


Figure 5: World Map showing the Readiness of each country in terms of combating the next pandemic using LT. Their Levels of Preparedness are shown as legends in the bottom left corner. This map was generated by providing uniform weightage to all the languages spoken in a country, i.e. excluding the percentage of speaker-base for a particular language in a country (Readiness measure calculated using Equation 2).

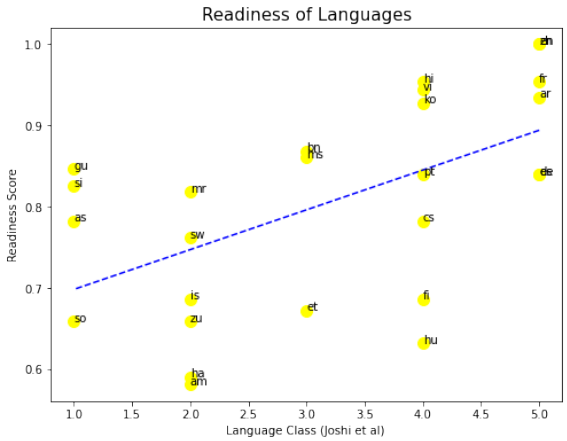


Figure 6: shows the readiness scores of the languages which are used in our training data for readiness measurement using GPR

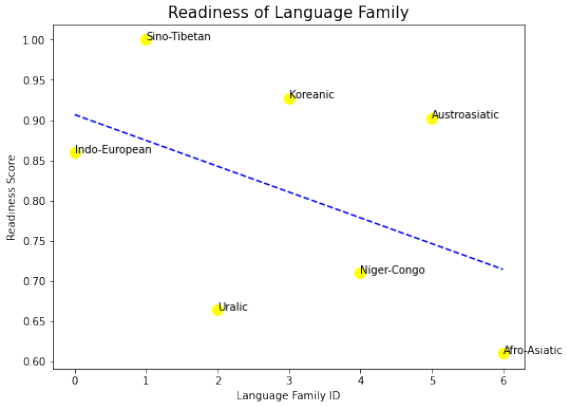


Figure 7: shows the readiness scores of the language families of the corresponding languages which are used in our training data for readiness measurement using GPR

equally while calculating the readiness of a country. This is similar to the linguistic utility defined by Blasi et al. (2021) in their work for a country c we calculate linguistic readiness r_c^{ling} as:

$$r_c^{ling} = \frac{1}{|\mathcal{L}_c|} \sum_{l \in \mathcal{L}_c} r_l \quad (1)$$

The r_c^{ling} values have been plotted in Figure 5. Based on our observations on these values we make the following observations highlighting the difference between demographic and linguistic readiness of different countries.

Observations: The map shown in 5 provides us an idea of how each country in the world would be able to effectively combat the pandemic by leveraging LT solutions. However, this is when we are actually considering uniform speaker-base for each language in a country. Overall, it can be observed that some of the Asian countries like India falls in the ***moderately prepared*** zone now which was initially treated as ***well prepared***. This is due to the presence of class 4 language Hindi (having a readiness score of 0.9536) with a considerably high speaker-base (46.19%). Also, similar trend is observed in Canada (home to the speakers of various languages like `English`, `French`, `Punjabi`, `Italian`, `Spanish`, `German`, `Cantonese`, `Arabic`, `Tagalog`).