Proceedings of SIGUL2022 @LREC2022, pages 92–97
Marseille, 24-25 June 2022

Language Technologies for Low Resource Languages: Sociolinguistic and Multilingual Insights

A. Seza Doğruöz, Sunayana Sitaram
Universiteit Gent, Microsoft Research India
Belgium, India
as.dogruoz@ugent.be, sunayana.sitaram@microsoft.com

Abstract

There is a growing interest in building language technologies (LTs) for low resource languages (LRLs). However, there are flaws in the planning, data collection and development phases mostly due to the assumption that LRLs are similar to High Resource Languages (HRLs) but only smaller in size. In our paper, we first provide examples of failed LTs for LRLs and provide the reasons for these failures. Second, we discuss the problematic issues with the data for LRLs. Finally, we provide recommendations for building better LTs for LRLs through insights from sociolinguistics and multilingualism. Our goal is not to solve all problems around LTs for LRLs but to raise awareness about the existing issues, provide recommendations toward possible solutions and encourage collaboration across academic disciplines for developing LTs that actually serve the needs and preferences of the LRL communities.

Keywords: Low Resource Languages, Multilingualism, Sociolinguistics, Language Technologies

1. Introduction

Low resource languages (LRL) refer to the languages spoken in the world with less linguistic resources for language technologies (LTs) (Cieri et al., 2016). Endangered and/or minority languages also overlap with the LRLs in terms of lacking resources for LTs. Different than endangered languages, not all LRLs and minority languages suffer from low numbers of speakers (cf. Pandharipande (2002)) for the situation of minority languages in India).

Joshi et al. (2020) categorize the languages of the world into six categories based on the resources available in terms of labeled and unlabeled data. More than 88% of the world's languages belong to the lowest resource class, with only 25 languages belonging to the two high resource classes. In other words, a majority of the world's languages count as LRLs even when they have large numbers of speakers (e.g. Gondi (Mehta et al., 2020) and Odia (Parida et al., 2020) spoken in India).

Data collection, annotation and analyses remain as challenges for LTs involving LRLs due to limited resources. Even when the data challenges are resolved, the resulting LTs may still not be favored and adopted by the LRL communities.

Recent advances in massive contextual language models (particularly multilingual versions) (Devlin et al., 2018; Conneau et al., 2020) give the impression that LTs for LRLs are solved based on their performance on some benchmarks (mainly covering high resource languages and a few NLP tasks) (Ruder et al., 2021; Liang et al., 2020). However, the majority of (approx. 100) languages covered by these models remain untested by these benchmarks, and the models are not trained on the majority of the world's languages.

Our goal is to highlight the dangers of viewing LRLs the same as high resource languages (HRLs) but only with less data and limited budget. To do this, we provide examples of well-intended but failed LTs for LRLs and explain the reasons through insights from sociolinguistics and multilingualism. Next, we describe the challenges with data (e.g. dangers of focusing on "purity" in LRLs). Lastly, we provide guidelines for different parts of the pipeline (i.e. data collection, annotation and evaluation) to develop better and more informed LTs for LRLs and their speakers.

2. Issues about Sociolinguistic Variation

We start this section with an example of a failed LT for an LRL community in India and explain the reasons with links to sociolinguistics. Voice-based systems are potentially useful LTs for LRL communities with low literacy rates. Spoken Dialogue Systems or Interactive Voice Response (IVR) systems rely on carefully designed prompts and a vocabulary list that needs to be recognized by a speech recognizer. For example, VideoKheti (Cuendet et al., 2013) is a speech and graphics based application targeting speakers of Malvi language in India (a sub-dialect of the Rajashtani dialect of Hindi (Bali et al., 2013)). The application was created to help (illiterate) farmers in rural India to access agricultural online videos by spoken search. During the data collection, a local non-governmental organization (NGO) assisted the project to develop a vocabulary list for the speech recognition system. However, the list contained many technical words which were borrowed from Hindi (and in a formal register) and did not exist in the linguistic repertoires of Malvi speakers in daily and informal communication. Example (1) illustrates one of the problematic technical terms which could (roughly) be translated as "chemical pesticide" (Bali, 2020).

1. **Rasaayanik tarike se kharpatwaar niyantran**
   Chemical technique for weed control

Instead of example (1), Malvi speakers would normally use example (2) in the same context.

2. **keede maarne ki dawaai**
   pest killing medicine

As illustrated with examples above, the terminology used for the linguistic prompts in the app did not match with the daily language use in the Malvi community. This mismatch led to errors in both the recognition and understanding of the prompts produced by the system.
Another mismatch in language use was observed in terms of gender differences. More specifically, female Malvi speakers had more difficulty than male speakers using this app. There could be a few reasons behind this observation. Although the new terminology in the app (see example (1)) was unfamiliar to both male and female speakers of the same community, (some) male speakers eventually got familiar with the new terminology through attending the meetings organized by the NGO. For female members, on the other hand, it is not always socially acceptable to attend such public meetings. Second, female members may not always feel comfortable to voice their opinions freely in presence of males or elderly relatives (e.g. parents-in-law) even if they attend such meetings.
Despite the well-intended efforts, the particular app ended up relying mostly on the graphic interface and the speech part was underused by the LRL community members. In other words, it did not serve its development purpose not to mention the unfortunate use of resources and (possible) disappointment among developers and LRL community members who spent time and energy on it. Both of the challenges explained above could have been avoided by a thorough analysis of sociolinguistic variation in the respective LRL community. More concretely, specific farming terminology for the app should not have been developed in a top-down fashion but in a bottom-up way through observing and collecting informal and conversational data from the community members across different backgrounds (e.g. genders, ages, educational background) and in different contexts (e.g. from males and females on different occasions). In this way, the app would have reflected the language used by the LRL community members and it would have served its development purpose.

3. Issues about Multilingualism

Considering that multilingualism is the norm in majority of the world (Dorian, 2014), it is also reasonable to assume (at least some) LRL speakers and communities to be multilingual. In that case, there is a need to analyze their attitudes toward LRLs as well as the power and prestige hierarchies in those contexts before developing any LTs for these communities. For example, speakers of endangered and/or minority languages who had disadvantages in social life (e.g. finding a job) due to lack of language abilities in the dominant language may prefer not to speak LRLs with their children (Dorian, 2014). Pandharipande (2002) gives an example of a housemaid who is a native speaker of Tulu (a LRL) and works in Mumbai (India). She declined to teach and speak Tulu with her children since English and Marathi are the languages that they should be learning for upward mobility (e.g. better education and jobs) according to her.
Similarly, it is quite normal for multilingual LRL community members to switch across languages/dialects in their daily communication. Although there is plenty of research about multilingual language use and code-switching across languages in the world (e.g. an extensive survey by Doğruöz et al. (2021)), multilingualism is not always taken into account while developing LTs for these communities.
For example, in Automatic Speech Recognition (ASR) systems, Srivastava et al. (2018) and Shah et al. (2020) observe that it is not possible to remove all the utterances with foreign words (e.g. code-switching into English) in Hindi since some of these words are already borrowed and got integrated into the language over time. Besides, Hindi has already quite a few borrowings from Persian and Arabic due to centuries-long contact (Jain and Cardona, 2007). Since the distinction between code-switching and borrowing is often blurry (Doğruöz et al., 2021), filtering either of these from the system arbitrarily will lead to system failures. As a result, LRL communities will not approve and adopt the system for which valuable time, energy and resources were invested. Therefore, aiming to create monolingual data sets even for comparisons or benchmarking purposes is not a meaningful effort for LRLs which inherently contain many borrowed words in highly multilingual areas (e.g. India, Africa, Polynesian islands).

4. Issues about Data

Data-driven studies in NLP and speech processing rely on large datasets of text and speech to build models or gain insights automatically. These datasets are curated from naturally occurring data (e.g. social media and/or recorded conversations among humans), or they are created specifically targeting the intended use case scenario.
In general (for most HRLs), there is a tendency to collect only monolingual data, in its standard dialect and with a formal register so that a "pure" target language (e.g. ignoring the inherent sociolinguistic variation in the community) would benefit the accuracy of the system. As a result, the data set becomes very small and artificial in the sense that it does not represent the language spoken in the community anymore (cf. Nguyen et al. (2016)). Although these flaws could be improved for HRLs with enough resources over time, there is (usually) not a second chance for LRLs with limited

manpower, budget and resources. As a result, the LRL communities are left with LTs that do not reflect their language use and do not serve their needs and preferences.

5. Recommendations for Building LTs for LRLs

In the previous sections, we explained how lack of insights in sociolinguistics and multilingualism leads to flaws in developing LTs for LRLs. In this section, we provide guidelines and solutions about how to avoid these pitfalls for the LT pipelines targeting LRLs.

Preparation: Before building any type of LTs and collecting data, making sufficient sociolinguistic inquiries about the dynamics and language use practices among a LRL community is crucial. For example, literacy status of the users, availability of written scripts in a LRL, multilingual and mixed language practices in the community should be researched extensively. In addition, existing data sets (albeit small or not of high quality) for endangered and minority languages (e.g. Pangloss collection by Michailovsky et al. (2014) for endangered Asian, Oceanic, Caucasian, European languages, ELAR (Endangered Languages Archive) collection described by Nathan (2013)) could serve as starting points for LTs in LRLs. They usually come with a description of the meta-data (e.g.participants/community, context) which could give some preliminary insights about the community dynamics. Before collecting any type of data in the LRL community, it is recommended to connect with the Linguistics Department of a local university for their help on available literature, on-going or completed projects on the local LRLs as well as training and employing their students for field work. Instead of allocating resources in a top-down fashion, it is more feasible, less expensive and less time-consuming to start bottom-up with the existing resources and collaborate with fellow researchers in linguistics/sociolinguistics who may already have insights about the LRLs and their communities in depth.

Data Collection: Given that NLP models are becoming larger and require more data than ever before, data collection remains the backbone on which LTs are built upon today. Ideally, the data for LRLs should be collected from the speakers that LTs will benefit. However, this is not always practiced. Instead, it often results in approximating the target LRL by using existing HRL data which is often not representative of the LRLs spoken in the community.

LRL communities should not be expected to adapt to language of the LT developed through random and approximate data sets. Instead, it is crucially important to send multi-disciplinary (e.g. computational (socio)linguists,engineers, multilingualism experts, social workers) teams to spend extended periods of time with the target LRL community with the goal of understanding their (multilingual) needs and preferences as well as the the sociolinguistic variation operating in the particular LRL context. If this is considered a challenge (which should not be), it is at least desirable to collect better approximate data (instead of random ones) which would reflect LRLs in real-life like conversational situations (e.g. movies and soap operas reported by (Biswas et al., 2022)).

Data Cleaning: Language technologies are usually built with monolingual assumptions about character sets, vocabulary and lexicons. The limited amount of data available for LRLs is further reduced if it is also cleaned or filtered to make it (often unnaturally) monolingual. In addition to ignoring the dynamic and multilingual aspects of the data, there is also the danger of not being able to make the best use of naturally occurring data with all its deficiencies and variation (aka "bad language" (Eisenstein, 2013)), or collecting data that does not reflect the real use in the given community.

Prior work on dealing with linguistic variation in the data focused on normalization and domain adaptation. Both of these approaches are problematic. Normalization processes assume that there is a default norm in every language and this norm is often associated with the monolingual, standard dialect and formal register. This assumption results in ignoring communities and speakers who may not use the standard dialect and formal register in their daily communication. Similarly, domain adaptation is not ideal for shifts in medium of expression, like social media. Languages are dynamic and they constantly change even in (supposedly) monolingual contexts. Therefore, LTs should also change and handle linguistic variation simultaneously instead of ignoring and cleaning the data through extensive normalization processes (cf. Nguyen et al. (2016)).

In multilingual contexts and communication, this translates as avoiding to clean the data from foreign influences (e.g. code-switching) to make it "pure" or monolingual, avoiding to create artificial datasets by collecting data in the wrong register (e.g. "formal" instead of "informal"), and avoiding to ignore the foreign language influences (e.g. code-switching and borrowing) during the processing phase.

Annotation: Labeling sociolinguistic variation (e.g. multilingualism, variation in styles, registers, variation across contexts and social variables of users) in the data is challenging due to the lack of standardization. In fact, tailor made solutions are probably more feasible than standard solutions that are assumed to apply across all LRLs. In addition to code-switching, it is also common to switch across scripts in India. For example, annotators use multiple scripts (Devanagari and Latin) to transcribe Hindi-English code-switched speech (Srivastava and Sitaram, 2018) and they may end up transcribing the same word in both scripts in different instances in the corpus. Although it may seem that this problem can be avoided by training the annotators or providing instructions to them, it remains an extremely challenging problem because the distinction

between switching and borrowing is blurry (Doğruöz et al., 2021). A related issue is the lack of standardized spellings for borrowed words. Inconsistencies in transcription lead to less training data per word during the model building. As a result, a vicious cycle is created with difficulties in using automated tools to bootstrap labeling due to inconsistently labeled data.

Model building: Models built with monolingual assumptions may produce errors while processing inherently multilingual LRLs and this leads to lower performance of the model. Systems may either ignore content that is not in the expected language, or perform poorly on multilingual utterances. Massive multilingual models such as multilingual BERT (Devlin et al., 2018) and XLMR (Conneau et al., 2020) can process around 100 languages in a single model, however they tend to perform worse on LRLs compared to HRLs (Wu and Dredze, 2020). There is also evidence to show that these models perform poorly on mixed languages (Khanuja et al., 2020). Using these models through few-shot or zero-shot techniques on LRLs may not lead to desired outcomes, since the data they are pre-trained on (e.g. Wikipedia texts or randomly crawled data from the web) does not represent the language use within the LRL community. Adaptation techniques can be explored if the standard variety of the language is used to build the model. However, the adaptation data needs to be collected considering the sociolinguistic variation in the LRL. During the system design phase, there is a need to carefully examine which models are best suited for the intended purpose, instead of assuming that the largest, latest and most accurate models on HRL will also perform best on LRLs. If multiple languages are being served by the same model, there is a need to consider whether the model is fair to all languages (Choudhury and Deshpande, 2021) and that some languages do not benefit at the cost of others. Models that are explainable and easy to debug will also benefit from the feedback provided by the users of LRL communities.

Evaluation: Evaluation benchmarks do not exist for most LRLs (Bhatt et al., 2021). The few and available test datasets may not reflect the way language is used in LRL communities and decrease the usefulness of the benchmarks. Many of these benchmarks (very expensive to create), turn out to be brittle to spurious patterns learned by NLP models (Glockner et al., 2018). Due to over optimization on a small set of benchmarks, it is likely that the performance of NLP models (even on HRLs) is an overestimate. This situation is even more stark in case of LRLs (Wu and Dredze, 2020), where benchmarks do not exist for most languages and tasks.

Code-switching and borrowing make it harder to evaluate systems due to cross-script transcription, multiple ways of conveying the same meaning and the issues (mentioned earlier) with data collection and annotation. Although some NLP benchmarks (e.g. XTREME-R by (Ruder et al., 2021)) cover 50 languages and a set of diverse tasks, each language is still assumed to be strictly monolingual. Currently, there are only two benchmarks that deal with mixed languages (i.e. code-switching, GLUECoS (Khanuja et al., 2020) and LinCE (Aguilar et al., 2020)). However, even these benchmarks cover very few LRLs and a small set of tasks.

Metrics to evaluate LTs are flawed, since they are usually created for HRLs and they do not always reflect the nuances of how the LTs will actually be used in other languages. There is an urgent need to create better metrics and benchmarks, preferably by collecting evaluation data directly from the speakers of the target LRL communities. Accurate and meaningful evaluation of LTs for LRL users can only happen through their participation.

6. Discussion

To conclude, building LTs for LRLs is not a solved problem and there are no simple and quick recipes. LTs built without enough understanding of the LRL communities may not serve their purposes. Therefore, all the aspects mentioned above regarding the data collection, cleaning, annotation, model building and evaluation should be considered by multi-disciplinary teams before building any LTs for LRLs.

Experimenting with LRLs in computational linguistic domains can be a commendable scientific endeavour to test the limits of NLP models (e.g. massive multilingual models), explore new modeling techniques and may also lead to significant improvements in performance for these languages. However, improvements in performance should not be conflated with usefulness of the LTs for the target LRL community without making sure that the factors mentioned above are taken into account, and appropriate evaluation (when possible, including the LRL community members) is carried out.

If the goal is to develop LTs that are actually useful for the LRL community members, there is a need to slow down and understand the social and linguistic dynamics operating in a LRL community through a careful examination. After involving all the stakeholders, appropriate data that reflects real-life language use in the LRL community should be collected without being exposed to a cleaning/normalization phase to increase the accuracy of the models. Fair and explainable models which could also integrate feedback should be favored and evaluated by using the appropriate benchmarks or by testing with the LRL community members.

Ignoring the above-mentioned challenges and pitfalls will only lead to LTs which will remain as experimental trials without any prospects for successful adoptions by the LRL communities. Any serious attempts on building LTs for LRLs can be only be realized through inter-disciplinary collaboration across fields and after following above-mentioned steps closely. We hope that the guidelines in our paper could serve as footprints for the researchers and developers to build better LTs for LRLs and their communities.

7. Bibliographical References

Aguilar, G., Kar, S., and Solorio, T. (2020). Lince: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1803–1813.

Bali, K., Sitaram, S., Cuendet, S., and Medhi, I. (2013). A hindi speech recognizer for an agricultural video search application. In *Proceedings of the 3rd ACM Symposium on Computing for Development*, pages 1–8.

Bali, K. (2020). The giant leaps in technology - and who is left behind. TEDxMICA.

Bhatt, S., Goyal, P., Dandapat, S., Choudhury, M., and Sitaram, S. (2021). On the universality of deep contextual language models. *arXiv preprint arXiv:2109.07140*.

Biswas, A., Yılmaz, E., van der Westhuizen, E., de Wet, F., and Niesler, T. (2022). Code-switched automatic speech recognition in five south african languages. *Computer Speech & Language*, 71:101262.

Choudhury, M. and Deshpande, A. (2021). How linguistically fair are multilingual pre-trained language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12710–12718.

Cieri, C., Maxwell, M., Strassel, S., and Tracey, J. (2016). Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Cuendet, S., Medhi, I., Bali, K., and Cutrell, E. (2013). Videokheti: Making video content accessible to low-literate and novice users. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2833–2842.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Multilingual bert readme document. *Library Catalog: github. com*.

Doğruöz, A. S., Sitaram, S., Bullock, B. E., and Toribio, A. J. (2021). A survey of code-switching: Linguistic and social perspectives for language technologies. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*. Association for Computational Linguistics.

Dorian, N. (2014). *Small-language fates and prospects: Lessons of persistence and change from endangered languages: Collected essays*. Brill.

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 359–369.

Glockner, M., Shwartz, V., and Goldberg, Y. (2018). Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655.

Jain, D. and Cardona, G. (2007). *The Indo-Aryan Languages*. Routledge.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (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.

Khanuja, S., Dandapat, S., Srinivasan, A., Sitaram, S., and Choudhury, M. (2020). Gluecos: An evaluation benchmark for code-switched nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585.

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., et al. (2020). Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Mehta, D., Santy, S., Mothilal, R. K., Srivastava, B. M. L., Sharma, A., Shukla, A., Prasad, V., Venkanna, U., Sharma, A., and Bali, K. (2020). Learnings from technological interventions in a low resource language: A case-study on gondi. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2832–2838.

Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., and Adamou, E. (2014). Documenting and researching endangered languages: the pangloss collection. *Language Documentation & Conservation*, 8:119–135.

Nathan, D. (2013). Access and accessibility at elar, a social networking archive for endangered languages documentation. *Oral literature in the digital age: archiving orality and connecting with communities*, pages 21–40.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., and De Jong, F. (2016). Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593.

Pandharipande, R. V. (2002). Minority matters: Issues in minority languages in india. *International Journal of Multicultural Societies*, 4:213–235.

Parida, S., Dash, S. R., Bojar, O., Motlicek, P., Pattnaik, P., and Mallick, D. K. (2020). OdiEnCorp

2.0: Odia-English parallel corpus for machine translation. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 14–19, Marseille, France, May. European Language Resources Association (ELRA).

Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Garrette, D., Neubig, G., et al. (2021). Xtreme-r: Towards more challenging and nuanced multilingual evaluation. *arXiv preprint arXiv:2104.07412*.

Shah, S., Sitaram, S., and Mehta, R. (2020). First workshop on speech processing for code-switching in multilingual communities: Shared task on code-switched spoken language identification. *WSTC-SMC 2020*, page 24.

Srivastava, B. M. L. and Sitaram, S. (2018). Homophone identification and merging for code-switched speech recognition. In *Interspeech*, pages 1943–1947.

Srivastava, B. M. L., Sitaram, S., Mehta, R. K., Mohan, K. D., Matani, P., Satpal, S., Bali, K., Srikanth, R., and Nayak, N. (2018). Interspeech 2018 low resource automatic speech recognition challenge for indian languages. In *SLTU*, pages 11–14.

Wu, S. and Dredze, M. (2020). Are all languages created equal in multilingual bert? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130.