Promoting Language Technology for Endangered Languages with Shared Tasks

Gina-Anne Levow
University of Washington
Seattle, WA USA
levow@uw.edu

Abstract

Although recent years have seen dramatic improvements in speech and language technologies, such systems are only available for the few hundred highest resource languages. These systems' reliance on large annotated datasets has limited their impact on the thousands of low-resource and endangered languages which could otherwise benefit. At the same time, it is estimated (Woodbury, 2019) that 50-80% of the world's languages are at risk of disappearing by 2100. Shared tasks have helped to drive rapid development of language technologies by providing shared, standardized datasets, evaluation metrics, and venues to communicate approaches and results. The current project aims to develop shared tasks and systems targeting language technologies that can accelerate documentation and facilitate revitalization of endangered languages, while advancing the state of the art.

Keywords: Shared Tasks, speech processing, speaker recognition, language identification

1. Introduction[1]

Speech and language processing techniques have made dramatic strides in the recent years. Human "parity", system performance at the level of people performing the same tasks, has been claimed for high profile tasks, such as automatic speech-to-text transcription (Stolcke and Droppo, 2017; Xiong et al., 2017) and machine translation (Hassan et al., 2018). However, these highly impressive results have typically been demonstrated on high-resuource languages such as English and the Chinese-English pair in the examples above. Furthermore, web-based tools for these applications are available for only 100-200 languages. Unfortunately, this situation is not surprising as the success of these systems relies not only on algorithmic advances, but also crucially on large-scale language resources, on the order of thousands or even tens of thousands of hours of transcribed speech or tens of millions of lines translated text.

As a result, it remains difficult to bring these new technologies to bear to benefit lower resource or endangered languages. At the same time it is feared that of the more than 6,000 languages spoken around the world, 50-80% could be lost by 2100(Woodbury, 2019). Language technologies have the potential to dramatically accelerate and facilitate efforts in language documentation and revitalization, if they can be effectively harnessed.

2. Approach: Shared Tasks

To bridge this gap between speech and language technologies and the needs of endangered language researchers and speaker communities, we plan to leverage the framework of Shared Task Evaluation Campaigns (STECs or Shared Tasks). STECs have been powerful drivers of speech and language technologies (Belz and Kilgarriff, 2006) and have contributed to the development of systems ranging from spoken dialog systems (Dahl et al., 1994) to document retrieval (Voorhees and Harman, 2005) to information extraction(Grishman and Sundheim, 1996). STECs provide standard datasets for training and testing systems, standard evaluation metrics, and venues for sharing results and techniques. As such, these tasks allow access to valuable data needed for system creation, direct comparisons across different methods, and transmisson of successful strategies. They also allow the community and organizers to focus research on tasks of interest, while pooling the costs of resource development.

Our planned STECs will focus research attention on tasks which will benefit endangered language researchers and speaker communities and will leverage growing archives of endangered language data. This setting will enable researchers in speech technology to assess their systems on a broader and more diverse range of languages than are typically employed (Bender, 2011). The tasks will also provide a more naturalistic setting in which to evaluate models for low-resource language processing, in contrast to simulations of low-resource settings by subsetting high resource language data.

3. Background & Design Principles

To help design tasks that would address crucial needs among researchers in endangered languages and speaker communities while challenging the state-of-the-art in language technology, a National Science Foundation-funded workshop, EL-STEC: Shared Tasks with Endangered Language Data, was held in September 2016, bringing together representatives of these different communitites. The discussions were driven by the identification of key pain points in the workflow of those striving to understand and document endangered languages, as well as capabilities desired by speaker communities. This process helped to define a suite of candidate tasks. In addition, it led us to articulate design principles that guide the structure of these and subsequent STECs. These criteria are described below.

[1]This paper summarizes and updates an original publication in (Levow et al., 2017).

Realism Our tasks should reflect the needs and usage environments of future users. We will also encourage participants to leverage any available resources for these tasks, rather than artificially restricting these resources as is very common in other shared task settings. Such sources range from linguistic repositories, such as ODIN (Lewis and Xia, 2010) or WALS (Haspelmath et al., 2008), to models derived from high resource languages that could be adapted to new tasks, such as pre-built Universal Background Models (Hasan and Hansen, 2011) or X-vector (Snyder et al., 2018) models for speaker identification. The structure of existing archive data will intrinsically impose a range of challenges, including limited data size, varied recording or collection conditions, differences in speakers or genres, and so on.

Typological diversity To truly focus on language technologies with broad effectiveness and applicability, our tasks will involve multiple typologically distinct languages in training and testing. In addition, evaluation will include previously unseen "surprise" languages from additional language families which will explicitly test new systems' portability to new languages.

Accessibility of shared tasks We hope to encourage broad participation in these tasks by lowering barriers to entry. In particular, baseline systems will be provided as part of the shared tasks to all participants, in addition to data and evaluation software. These baseline systems allow the organizers to validate the data and task design, as well as to provide a reference level of effectiveness. Since the baselines will be released publicly, they can also serve as starting points for participating teams with fewer resources, e.g. students, to build on to develop their own submissions. Finally, following the model of the Speech Recognition Virtual Kitchen (SRVK) project (Plummer et al., 2014), we will encapsulate these baselines and all needed software for the shared tasks in virtual machines to facilitate cross-platform development.

Accessibility of resulting systems It is crucial that the technologies developed under these shared tasks become available to benefit new user populations, including endangered language researchers and speaker communities. A key requirement is thus that participating teams provide detailed descriptions of their systems to enable replication and reimplementation. We also encourage the creation of systems based on free or open-source software.

Extensibility These initial tasks will establish a first set of multilingual resources for endangered languages as well as baseline performance against which to measure future progress. In addition, the tools developed for dataset preparation will provide a template for extension to additional new languages in future years.

Nuanced evaluation Multiple metrics can be incorporated in evaluation to better assess the strengths and weaknesses of different approaches or different facets of the tasks, rather than focusing on a single metric and leaderboard rank.

The EL-STEC workshop working groups defined three tasks which aimed to embody the above principles: one focused on speech processing, one on morpho-syntactic analysis and glossing through first-pass creation of interlinear glossed text, and one on orthographic normalization for endangered languages. Below, we describe the first set of tasks that we plan to field in upcoming challenges.

4. "Grandma's Hatbox": Speech Processing for endangered languages

A core challenge in documenting endangered languages is that although many hours of recordings between a linguist and their consultant(s) may be collected, those valuable recordings must still pass through a lengthy and time-consuming series of steps including segmentation, transcription, alignment, and glossing. Due to the time and effort required, each stage of this process yields less and less material analyzed in greater and greater detail, in a sort of funneling. We define a set of tasks that could help to increase the throughput of this process by automating key steps of speech processing, thereby freeing those working with endangered languages to focus their expertise on important areas of analysis, while making more material available to archives, and providing richer metadata to support easier access for speaker communities.

The "Grandma's Hatbox" task cascade envisions the process required to prepare and archive a new collection of speech recordings. The name evokes a scenario where a trove of recordings and field notes is discovered left behind in a box or donated by a former researcher. The subtasks involved are as follows:

- segmentation of the recordings by language, considering both automatic identification of known high and low resource languages and clustering of languages not known in advance,
- segmentation of the recordings by speaker, considering both automatic identification of known speakers and clustering of speakers not known in advance,
- automatic identification of the genre of the recording, to be drawn from a small fixed inventory, including narrative, conversation, elicitation, and ritual, and
- finally, automatic alignment of partial transcriptions to recorded audio.

Participating teams would be able to choose to work on any subsets of these tasks. Each step above can feed into a subsequent processing step to create an enriched audio archive. The resulting metadata can be provided as a template to augment an archive's database or converted to a standard format viewable in an interactive interface such as ELAN (Brugman and Russel, 2004). A graphical depiction of the automatic processing and enrichment appears in Figure 1.

5. Challenging the State of the Art in Speech Processing

The tasks outlined above simultaneously address key pain points in the work process of field linguists and others who work with endangered languages and push the state of the art in spoken language processing. By working with endangered language data, these tasks pose novel challenges and

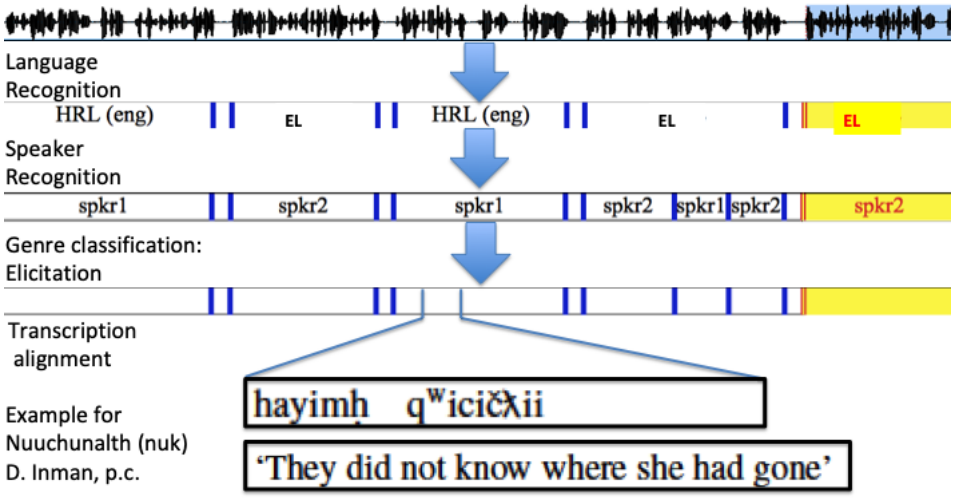

Figure 1: Speech processing cascade applied to audio and transcriptions.

present important opportunities. Broadly, they facilitate assessment of existing methods on new languages, many of which are typologically distinct from those on which most speech tools are trained and tested. Furthermore, the rise of neural network models has demonstrated effective techniques for exploiting large-scale data resources; however, there is substantial interest in low-resource techniques, such as those using unsupervised or semi-supervised learning. The fundamental resource constraints present in work with endangered language data will encourage exploitation and development of such methods.

Each of the steps of the speech processing cascade builds on existing technologies, spanning language identification, speaker diarization, speaker recognition, and automatic alignment. Although there are shared regimes in some of these areas, such as the NIST speaker (NIST, 2019) and language (NIST, 2017) recognition evaluations, the character of endangered language data poses new and important challenges beyond those typically addressed. This new data often involves multiple speakers, speaking in multiple languages or dialect varieties, often with short turns and possibly with fine-grained code-mixing. Rather than the broadcast or telephone conversational speech typically found in speech research corpora, recordings of endangered language data involve a range of genres and diverse, possibly noisy, recording conditions. These factors are often listed when discussing the limitations of existing tools. By creating tasks which drive development of systems to address these challenges, our STECs will advance these technologies.

6. Potential Benefits to Speaker Communities

In addition to aiding researchers working with endangered languages, these shared tasks hold the potential to also benefit speaker communities and aid in revitalization efforts. The speech processing tools presented above can facilitate access to new or existing recordings within communities. Speech, and video, recordings are notoriously slow and difficult to search or browse. By identifying the languages, speakers, and genres of recordings, the tools can enable some search and navigation within these valuable materials. In conjunction with speech alignment, content-based search and browsing could also be supported.

7. Conclusions & Future Work

We believe that Shared Task Evaluation Campaigns designed around endangered language data have the potential to benefit field linguists, endangered language researchers, language archives, and speaker communities while driving improvements in language technology. We anticipate that these enhancements will yield technologies that have applicability to a more diverse range of languages, both in terms of typology and in terms of availability of linguistic resources. We are currently preparing the datasets and software to launch the first iteration of our STECs in the coming year. We look forward to engaging with new partners and potential task participants.

8. Acknowledgements

This work has been supported by NSF #: 1500157 and NSF #: 1760475. Any opinions expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Great thanks to the members of the STREANLInED team at the University of Washington, especially my co-PI Emily M. Bender, Isaac Manrique, Jiani Chen, and Harita Kannan. We are also grateful for the contributions of Shobhana Chelliah, Joshua Crowgey, Dan Garrette, Jeff Good, Sharon Hargus, Mark Hasegawa-Johnson, Kristen Howell, Russ Hugo, David Inman, Jeremy Kahn, Lori Levin, Patrick Littell, Michael Maxwell, Alexis Palmer, Michael Tjalve, Laura Welcher, and Fei Xia during the EL-STEC workshop.

9. Bibliographical References

Belz, A. and Kilgarriff, A. (2006). Shared-task evaluations in HLT: Lessons for NLG. In *Proceedings of the Fourth International Natural Language Generation Conference*,

pages 133–135, Sydney, Australia. Association for Computational Linguistics.

Bender, E. M. (2011). On achieving and evaluating language independence in NLP. *Linguistic Issues in Language Technology*, 6:1–26.

Brugman, H. and Russel, A. (2004). Annotating multimedia/ multi-modal resources with ELAN. In *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*.

Dahl, D. A., Bates, M., Brown, M., Fisher, W., Hunicke-Smith, K., Pallett, D., Rudnicky, A., and Shriberg, E. (1994). Expanding the scope of the ATIS task: the ATIS-3 corpus. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 43–48. Morgan Kaufmann.

Grishman, R. and Sundheim, B. (1996). Message understanding conference - 6: A brief history. In *Proceedings of Coling 1996*, pages 466–471.

Hasan, T. and Hansen, J. H. L. (2011). A study on universal background model training in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1890–1899.

Martin Haspelmath, et al., editors. (2008). *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich. http://wals.info.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving human parity on automatic Chinese to English news translation.

Levow, G.-A., Bender, E. M., Littell, P., Howell, K., Chelliah, S., Crowgey, J., Garrette, D., Good, J., Hargus, S., Inman, D., Maxwell, M., Tjalve, M., and Xia, F. (2017). STREAMLInED challenges: Aligning research interests with shared tasks. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 39–47, Honolulu, March. Association for Computational Linguistics.

Lewis, W. D. and Xia, F. (2010). Developing ODIN: A multilingual repository of annotated language data for hundreds of the world's languages. *Journal of Literary and Linguistic Computing*, 25:303–319.

NIST. (2017). 2017 Language Recognition Evaluation Plan. `https://www.nist.gov/system/files/documents/2017/09/29/lre17_eval_plan-2017-09-29_v1.pdf`. Downloaded November 21, 2019.

NIST. (2019). 2019 NIST Speaker Recognition Evaluation Plan. `https://www.nist.gov/system/files/documents/2019/07/22/2019_nist_speaker_recognition_challenge_v8.pdf`. Downloaded November 21, 2019.

Plummer, A., Riebling, E., Kumar, A., Metze, F., Fosler-Lussier, E., and Bates, R. (2014). The Speech Recognition Virtual Kitchen: Launch party. In *Proceedings of Interspeech 2014*.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Stolcke, A. and Droppo, J. (2017). Comparing human and machine errors in conversational speech transcription. In *Proc. Interspeech 2017*, pages 137–141.

Ellen M. Voorhees et al., editors. (2005). *TREC: Experiment and Evaluation in Information Retrieval*. Digital libraries and electronic publishing series. The MIT Press, Cambridge, MA.

Woodbury, A. C. (2019). What is an endangered language? Accessed November 2019.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M. L., Stolcke, A., Yu, D., and Zweig, G. (2017). Toward human parity in conversational speech recognition. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 25(12):2410–2423, December.