NLP without Annotated Dataset

NLP for Endangered Languages

Sowmya Vajjala

Seminar für Sprachwissenschaft, University of Tübingen, Germany

27 January 2021

Why is this topic of interest to anyone?

What is different?

Isn't it just NLP with another new language? a case of multilingual NLP?

What is different?

Isn't it just NLP with another new language? a case of multilingual NLP?

- What is needed? can be different
- What is possible? can be different.
- Why is it needed? can be different.

What are some goals?

- Language documentation
- Language preservation
- Enabling communication (e.g., typing on phone)
- May be support teaching/learning of that language

"Systems constructed with zero expert resources [can] help field linguists document endangered languages, by providing tools to semi-automatically analyze and annotate audio recordings using automatically discovered linguistic units"
–Dunbar et al. (2017) The Zero Resource Speech Challenge 2017. Proc IEEE ASRU Workshop.
Source

What kind of stuff is useful?

Speech Technologies

- There may be several hours of audio recordings sometimes, for some language groups (due to field linguists efforts, community recordings etc.)
- However, accessing them like a human being can be difficult (how to search through them? are they transcribed? what languages are being spoken? etc)

What kind of stuff is useful?

Speech Technologies

- There may be several hours of audio recordings sometimes, for some language groups (due to field linguists efforts, community recordings etc.)
- However, accessing them like a human being can be difficult (how to search through them? are they transcribed? what languages are being spoken? etc)
- What is needed: Tools that can transcribe speech/label it/make it searchable somehow.

What kind of stuff is useful?

Text Technologies

- Typically, a lot of languages are morphologically complex, and current NLP methods may not exactly work
- So, from seemingly simple stuff like writing/speaking tools to grammar/spell checkers - everything can be much more challenging.

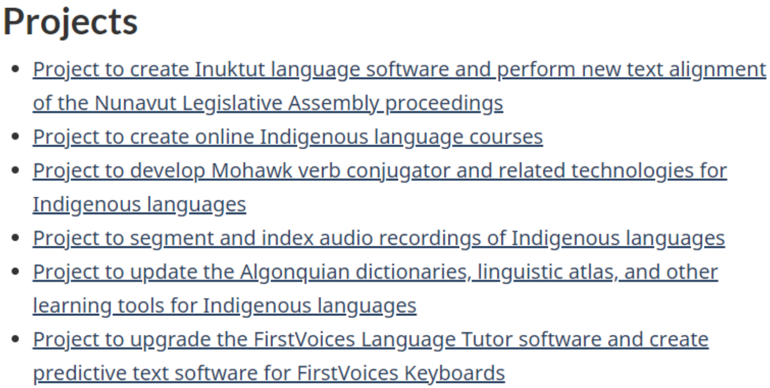
Where is NLP research connecting to this?

- Mobile applications for data collection related to language documentation (e.g., Bettinson & Bird, 2017
- Speech to text tools for field linguists (and others) (e.g., Foley et.al., (2019) tested this for Abui, an Indonesian language with 17000 speakers)
- Machine Translation and language revitalization Le and Sadat, 2020
- Developing language resources such as Universal Dependencies treebanks (https://universaldependencies.org/)

... ...

Some current research projects

Projects at my organization (NRC, Canada)



Projects

- Project to create Inuktut language software and perform new text alignment of the Nunavut Legislative Assembly proceedings
- Project to create online Indigenous language courses
- Project to develop Mohawk verb conjugator and related technologies for Indigenous languages
- Project to segment and index audio recordings of Indigenous languages
- Project to update the Algonquian dictionaries, linguistic atlas, and other learning tools for Indigenous languages
- Project to upgrade the FirstVoices Language Tutor software and create predictive text software for FirstVoices Keyboards

(note: I am not involved with these. I can find out details if any of you are interested.)

Text Input tools

"Interactive Word Completion for Morphologically Complex Languages"

- developed a method for morphologically-aware text input in Kunwinjku, a polysynthetic language of northern Australia.
- tested its portability through a relatively more known morphologically complex language - Turkish.
- deployed it as a usable tool and reported on its use.

Lane, W., & Bird, S. (2020, December). Interactive Word Completion for Morphologically Complex Languages. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 4600-4611).

Note: Other efforts of this kind exist for other languages.

Speech Transcription tools

"Enabling Interactive Transcription in an Indigenous Community"

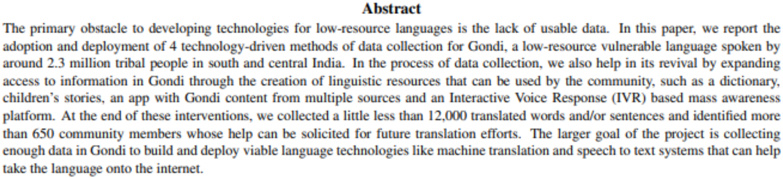
- Focus on Kunwinjku- Language spoken by 1500 people
- Task: speech transcription
- Result: Early stage deployment
- No training required.
- Participation of Indigenous people in the transcription process

Le Ferrand, E., Bird, S., & Besacier, L. (2020, December). Enabling Interactive Transcription in an Indigenous Community. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 3422-3428).

Note: Other efforts of this kind exist for other languages.

Other tools

"Learnings from Technological Interventions in a Low Resource Language: A Case-Study on Gondi"



Abstract

The primary obstacle to developing technologies for low-resource languages is the lack of usable data. In this paper, we report the adoption and deployment of 4 technology-driven methods of data collection for Gondi, a low-resource vulnerable language spoken by around 2.3 million tribal people in south and central India. In the process of data collection, we also help in its revival by expanding access to information in Gondi through the creation of linguistic resources that can be used by the community, such as a dictionary, children's stories, an app with Gondi content from multiple sources and an Interactive Voice Response (IVR) based mass awareness platform. At the end of these interventions, we collected a little less than 12,000 translated words and/or sentences and identified more than 650 community members whose help can be solicited for future translation efforts. The larger goal of the project is collecting enough data in Gondi to build and deploy viable language technologies like machine translation and speech to text systems that can help take the language onto the internet.

Mehta, D., et.al. (2020, May). Learnings from Technological Interventions in a Low Resource Language: A Case-Study on Gondi. In Proceedings of The 12th Language Resources and Evaluation Conference (pp. 2832-2838).

However ....

"The State and Fate of Linguistic Diversity and Inclusion in the NLP World"

- There are still a lot of languages in the world without any sort of NLP support.
- We still don't know much about how "language agnostic" are our language agnostic technologies.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020, July). 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 (pp. 6282-6293).

Some Caveats

It is more than "doing NLP on another new language".

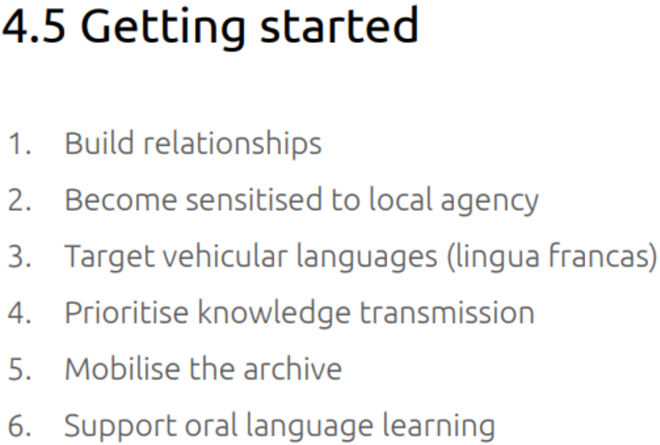
- "The research must be done respectfully, in collaboration with the communities who are the custodians of these languages. There should be no room in this field for scholars whose main goal is extracting "interesting" research without considering the linguistic needs of communities" (Kuhn et.al., COLING 2020)
- "While working with well-resourced languages, the main problem in designing language technologies is engineering. For low-resource languages, however, the main problem is one of designing methods for data collection upon which the language technology can be built." (Mehta et.al., 2020)

Some Caveats

It is more than "doing NLP on another new language".

- "The research must be done respectfully, in collaboration with the communities who are the custodians of these languages. There should be no room in this field for scholars whose main goal is extracting "interesting" research without considering the linguistic needs of communities" (Kuhn et.al., COLING 2020)
- "While working with well-resourced languages, the main problem in designing language technologies is engineering. For low-resource languages, however, the main problem is one of designing methods for data collection upon which the language technology can be built." (Mehta et.al., 2020)

Getting Started



4.5 Getting started

1. Build relationships
2. Become sensitised to local agency
3. Target vehicular languages (lingua francas)
4. Prioritise knowledge transmission
5. Mobilise the archive
6. Support oral language learning

Resources to explore further

- Indigenous Protocol and AI - position paper
- CMU's low resource NLP bootcamp, 2020
- Indigenous Language Technologies project
- `https://github.com/LowResourceLanguages/`
- Workshops on the use of Computational Methods in the Study of Endangered Languages
- Reading List on NLP for Endangered Languages (scroll to the end to see the list)
- Follow work by Steven Bird (a co-author of NLTK/NLTK book)

To get a taste of it

Do the first part of the exercise described in Chintang.pdf :-) It is from NACLO, again.