Advances in Low-Resource and Endangered Languages

Evelyne Tzoukermann
The MITRE Corporation
7525 Colshire Dr.
McLean, VA 22102
tzoukermann@mitre.org

Jason D. Duncan
The MITRE Corporation
7525 Colshire Dr.
McLean, VA 22102
jduncan@mitre.org

Caitlin Christianson
DARPA Contractor
675 N Randolph Street
Arlington, VA 22203
cchristianson@gryphontechnologies.com

Boyan Onyshkevych
US Department of Defense
Ft Meade MD 20755
baonysh@tycho.ncsc.mil

Abstract

This paper reports on the approaches and results for the collection, analysis, and processing of low-resource and endangered languages carried out under the Low-Resource Languages for Emergent Incidents (LORELEI) Program[1]. LORELEI was a multi-year research and development program designed to discover new methods of quickly ramping up human language technology capabilities for low-resource languages, grounded in situations such as humanitarian and disaster relief use cases. The goal was to advance human language technology methods to better enable rapid, low-cost development of capabilities, with a focus on developing methods that apply to languages *of any type from any language family*, thus eliminating the need to tailor specific technologies to a narrow set of input languages with specific typological characteristics. We report in detail on evaluation scenarios developed for the program.

1. Goals and Challenges

The LORELEI Program was created to address and solve multiple challenges. The primary challenge was to improve response-time to future Humanitarian and Disaster Relief (HADR) situations in areas where low-resource languages are widely used [1]. Examples of HADR situations include the 2010 Haiti earthquake, where there was an urgent need to handle information in Haitian Creole, and the 2004 tsunami that affected over a dozen countries, with multiple languages in each country, including Burmese, Bengali, Malgachi, and many others. The challenge was to establish effective help in saving lives at the ground level and efficient means to restore infrastructure. Achieving these goals required communicating in local languages, many of which had not been previously addressed by developers of natural language processing technologies.

The LORELEI Program included the following three task areas:

1. *Machine Translation (MT)*: automatic translation of other languages into English
2. *Entity Discovery and Linking (EDL)*: automatic detection of key entity types in the language data and linking these extracted entities to a knowledge base
3. *Situation Frame Detection (SF)*: identifying the most acute disaster relief need types and locations for various types of situations, such

1 For more information,
see https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents.

Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 2 Resource Papers and Extended Abstracts, pages 24–30,
Online, March 2–3, 2021.

as those in which people require rescue, food, medical assistance, etc.

The program focused on the specific points listed below.

a) Lack of annotation and supervised training data combined with a short time frame

For LORELEI scenarios, research teams were assigned languages that they had not known beforehand, and thus had had no opportunity to train any system components prior to the release of "surprise" language evaluation data. Therefore, the researchers needed to develop algorithms to address the LORELEI evaluation tasks in languages for which they had little to no training data. Then, they were tested on three evaluation tasks at one day, one week, and one month in the first two evaluations and one day and one week in the last two evaluations. These points were chosen to mimic the needs of HADR mission planners, i.e. what needs are discoverable on the first day after a disaster, and then at later points to meet emerging relief needs.

b) Effective ways to use a limited amount of native speaker time

One of the few types of in-language resources that LORELEI did provide to the research teams during the evaluations was a certain amount of time to work with a native speaker of each target language, referred to as a "native informant." However, LORELEI did not provide any sort of direction as to how the native informant time was to be used, nor any ground-truth data to assess how useful the native informant input was when applied to various tasks. In general, native informant time seemed to be the most effective when applied to creating targeted annotations to support rapid development of references for evaluation tasks. However, other advantages of working with native speakers require more study. One of the questions that the program raised was the possibility of building a "Native Informant Library" of interactions from all the evaluations, which could then be applied to new incident language situations in the future.

c) Novel information extraction task: Situation Frame (SF) detection

The program developed a novel information extraction task, which was entitled "Situation Frame" detection. This task was designed to steer researchers towards creating systems capable of providing situational awareness for HADR responders about where needs exist, and which are most urgent, based on input data in any language [2]. The situation frame task was a complex semantic information extraction task, which LORELEI required to be performed on low-resource languages in the absence of any training data for supervised machine learning. Adding to the complexity of the task was the relatively low inter-annotator agreement on the situation frame annotations created as ground-truth data for the evaluations, indicating that there may be a high degree of ambiguity in the input data.

2. Data and Annotation for Low-Resource Languages

The languages used in LORELEI were divided in two groups: representative and incentive (see the language list in the Appendix). Representative languages were relatively widely-spoken, higher-resource languages to be used for research and development tasks like pivoting to related languages, whereas incentive languages were used as surprise languages in program evaluations. In total, there were 23 representative languages, which were chosen to provide broad typological coverage for as many language families as possible. The representative language data sets that were provided to LORELEI research teams included monolingual representative language text, parallel representative language and English text, several types of annotation, and tools for text processing, segmentation, and entity tagging, as well as lexicons and grammatical sketches. The incentive languages, or "incident languages," were the surprise languages revealed to the research teams at the beginning of each evaluation. Nine incident languages were selected to enable development and testing of LORELEI system capabilities. (Note that Uyghur both served as the first incident language and later was made available as a representative language.) Incident language data

sets were intended to reflect the kind of data that might be available at the outbreak of an incident involving a low-resource language. Each data set was relevant to a specific historical incident, such as a flood or an earthquake. Compared to representative language data, incident language data sets contained smaller amounts of monolingual text and found parallel text, as well as an assortment of grammatical resources.

For representative languages, multiple dictionaries of about 10,000 words were built from online resources designed to maximize monolingual text. These resources were supplemented by tools like Johns Hopkins University's Unimorph, which provided morphological segmentation for multiple languages [3]. Figure 1, shows a representation of the small, but rich, core data sets created during LORELEI. The data were labeled in the languages of LORELEI annotation tasks and translated into all LORELEI representative languages. The annotation tasks comprised Translation, Simple Named Entity, Full Entity, Entity Linking, Simple Semantic Annotation, and Noun Phrase Chunking. All together, the program developed various types of linguistic resources for 31 languages [4]. In contrast to most previous research efforts in this area, the majority of resources came from crowdsourced or found data, not MT-related technical advances that LORELEI researchers made in support of program-funded creation. Even though LORELEI was not an MT program, LORELEI researchers made technical contributions in MT, in support of the main objective of gaining situational awareness. This included:

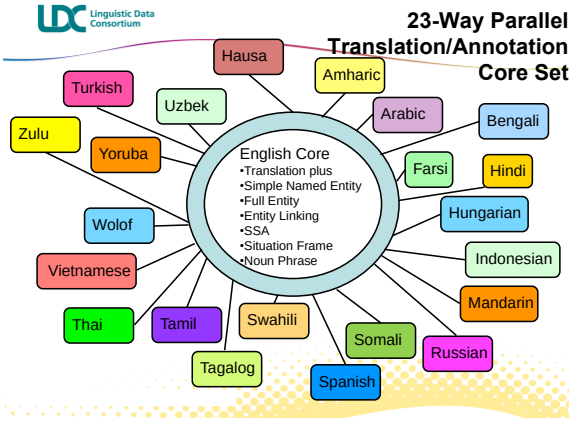

Figure 1: Representative Languages in the LORELEI Program

(a) neural machine translation which motivated the use of neural techniques elsewhere in the pipeline ([5] and [6]), and (b) the use of linguistic information from comparable languages. In the area of linguistic data innovations, LORELEI researchers successfully developed and implemented a new crowdsourcing platform, *Crowdtrans*, as well as novel methods for finding and leveraging existing parallel text, such as religious and government material, and comparable text in languages for which parallel text was unavailable.

3. Evaluation

Among the results of the program was the development of novel evaluation methods. There were two types of evaluation in the LORELEI Program. The first type, lab evaluation, consisted of the National Institute on Standards and Technology conducting annual evaluations, in which two surprise languages were revealed at the beginning of each evaluation to LORELEI participants, who then submitted their system outputs for assessment after 24 hours and after one week (and in some cases, also one month) of work. The final LORELEI lab evaluation, which was conducted with Odia, Ilocano, and English, showed that for the Named Entity Recognition task (part of entity discovery and linking task), LORELEI system performance on surprise languages was almost indistinguishable from performance of the same systems on English. Similarly, for situation frame detection, the LORELEI systems performed 85% as well on Odia and Ilocano as they did on English.

The second type of evaluation was a LORELEI-specific, task-completion evaluation aimed at showing the adequacy of LORELEI systems to support the kinds of situational awareness scenarios that arise in circumstances like disaster relief planning and incident management. For each of these evaluations, a language was specified in the context of an HADR exercise, and the exercise leaders attempted to respond based on information provided by LORELEI technology. LORELEI researchers, in partnership with MIT Lincoln Laboratory, participated in three exercises in Balkan countries (Bosnia,

Northern Macedonia, Montenegro). LORELEI information extraction technology was employed as an essential element of the system being used by all the participating nations to conduct their exercise activities. The 2019 exercise in Montenegro involved adding a new situation type to the SF inventory, report of a wildfire, which researchers were able to add in 24 hours. LORELEI's role was to process social media messages posted in Montenegrin by exercise participants, identifying relevant disaster-response information in order to create alerts showing need types and locations in the form of situation frames (Table 1 quantifies the corpus size). These LORELEI-produced alerts triggered the organizers to respond accordingly, for example, to send appropriate fire or rescue equipment to the reported location.

All Messages	**811**
Messages with Situation Frame Needs	332
Messages with No Relevant Content	479

Table 1. Simulated Social Media Message Contents for Montenegro Exercise

The LORELEI system allowed the incident organizers to identify all the disaster situations that were reported by social media, which is a significant success for the technology. Results achieved during the exercise are shown in Table 2. The numbers appear in terms of Precision, Recall, and F-measure (the weighted harmonic mean of precision and recall). When numbers are above 50%, we believe that it is sufficiently high to consider the technology for use in real-life applications. In this particular case, the numbers being between 60% and 70% is considered as evidence that the LORELEI systems are likely to yield meaningful operational results.

Detection Precision TP+TW / TP+TW+FP	60.2%
Detection Recall TP+TW / TP+TW+FN	72.0%
Detection F-Score 2PR / P+R	65.6%
Type Classification Accuracy TP / TP+TW	70.5%

TP = True Positive FP = False Positive TW = True Positive
TN = True Negative FN = False Negative Wrong Type

Table 2. LORELEI System Performance in Montenegro Exercise

Figure 2 shows images and descriptions, from the Montenegro field exercise.

Figure 2. Images and descriptions from Montenegro field exercise

4. Final Remarks

This paper presents a synthesis of the results of the LORELEI Program, which was an important effort for the low-resource and endangered language community, as well as for application of human language technology to response to any situation in which it is necessary to gain awareness of an incident based on a large amount

of low- and or high-resource language material. The results of the LORELEI lab and field evaluations validate the premise that even with extremely minimal data resources, human language technology can be used to address Humanitarian and Disaster Relief efforts or other emergent situations rapidly and effectively.

Moreover, LORELEI research has resulted in major advances likely to have a lasting impact in multiple areas, including (a) rapid ramp-up of natural language processing capabilities, (b) machine learning based on linguistic universals and inter-language relationships, (c) effective leveraging of imput native informants to maximize system development impact, (d) streamlining information extraction techniques for application to particular types of situations and linguistic landscapes, and (e) creating task-targeted repositories of low-resource language and related language data. Especially in places where languages with little or no documentation are spoken, LORELEI methods promise some concrete solutions to help in situations of emergency or disaster.

LORELEI was developed to enable research approaches focused on rapid adaptation through use of language universals and projection from related-language resources. The technologies resulting from LORELEI research are now capable of supporting situational awareness based on low-resource foreign language data sources within an extremely short time frame – starting as soon as 24 hours after new language and crisis requirements emerge. In addition, LORELEI has narrowed the gap between computational linguists working on endangered languages, native speakers of endangered languages, and field linguists who work on endangered-language documentation.

Acknowledgments

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

Selected References

[1] Christianson, Caitlin, Jason Duncan, and Boyan Onyshkevych. "Overview of the DARPA LORELEI Program." Machine Translation 32.1-2 (2018): 3-9.

[2] Malandrakis, Nikolaos, Ondrej Glembek, and Shrikanth S. Narayanan. "Extracting Situation Frames from Non-English Speech: Evaluation Framework and Pilot Results." INTERSPEECH. 2017.

[3] Hulden, Mans, and Ryan Cotterell. "Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection." Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection. 2018.

[4] LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low-Resource Languages Stephanie Strassel and Jennifer Tracey Linguistic Data Consortium 3600 Market Street, Suite 810 Philadelphia, PA 19104.

[5] Qi, Ye, et al. "When and why are pre-trained word embeddings useful for neural machine translation?." arXiv preprint arXiv:1804.06323 (2018).

[6] McNamee, Paul. "JHU LoResMT 2019 Shared Task System Description." Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages. 2019.

Appendix – Supplementary Material

This appendix lists the representative languages and incident languages used in the LORELEI Program. In each table, the first column shows the language and the following columns correspond to each corpus' Linguistic Data Consortium (LDC) reference number, by which the data set can be requested.

Representative Language	Monolingual Text	Annotation, Translation, Lexicon, & Tools	Monolingual Speech
Akan	LDC2018E06	LDC2018E07	LDC2017E84
Amharic	LDC2016E86	LDC2016E87	LDC2016E113
Arabic	LDC2016E88	LDC2016E89	LDC2016E123
Bengali	LDC2017E59	LDC2017E60	LDC2017E92
English	LDC2019E01	LDC2019E01	LDC2017E50
Farsi	LDC2016E92	LDC2016E93	LDC2016E124
Hausa	LDC2015E70	LDC2015E70	LDC2016E110
Hindi	LDC2017E61	LDC2017E62	LDC2017E88
Hungarian	LDC2016E98	LDC2018E99	LDC2016E125
Indonesian	LDC2017E65	LDC2017E66	LDC2017E91
Mandarin	LDC2016E100	LDC2016E101	LDC2016E108
Russian	LDC2016E94	LDC2016E95	LDC2016E111
Somali	LDC2016E90	LDC2016E91	LDC2016E126
Spanish	LDC2016E96	LDC2016E97	LDC2016E127
Swahili	LDC2017E63	LDC2017E64	LDC2017E86
Tagalog	LDC2017E67	LDC2017E68	LDC2017E89
Tamil	LDC2017E69	LDC2017E70	LDC2017E87
Thai	LDC2018E02	LDC2018E03	LDC2017E90
Turkish	LDC2014E115	LDC2014E115	LDC2016E109
Ukrainian	LDC2019E46	LDC2019E47	n/a
Uzbek	LDC2016E29	LDC2016E29	LDC2016E66
Vietnamese	LDC2016E102	LDC2016E103	LDC2016E128
Wolof	LDC2018E08	LDC2018E09	LDC2017E85
Yoruba	LDC2016E104	LDC2016E105	LDC2016E129
Zulu	LDC2018E04	LDC2018E05	LDC2017E93

Incident Language	Evaluation Text Data	Evaluation Text Reference Annotation	Evaluation Speech Data	Evaluation Speech Annotation
Ilocano	LDC2019E63	LDC2019R24	LDC2019E66	LDC2019E74
Kinyarwanda	LDC2018E55	LDC2018R16	LDC2018E60	LDC2018R11
Mandarin	LDC2016E30	n/a	LDC2016E115	n/a
Odia	LDC2019E62	LDC2019R23	LDC2019E65	LDC2019E73

Oromo	LDC2017E29	LDC2017R09	LDC2017E36	LDC2017E38
Sinhala	LDC2018E57	LDC2018R17	LDC2018E61	LDC2018R21
Tigrinya	LDC2017E27	LDC2017R08	LDC2017E35	LDC2017E37
Ukrainian	LDC2017E06	n/a	LDC2016E112	n/a
Uyghur	LDC2016E57	LDC2016R20	LDC2016E119/120	LDC2016E121
Uzbek	LDC2015E89	n/a	LDC2016E66	n/a