IARPA's Contribution to Human Language Technology Development for Low Resource Languages

Carl Rubino
IARPA
4600 Sangamore Road, Bethesda MD 20816 USA
Carl.Rubino@iarpa.gov

Abstract

With the goal of advancing the state of natural language processing development in constrained resource conditions, using machine learning and repeatable methodologies, the Intelligence Advanced Research Projects Activity (IARPA) launched two multi-year research programs: Babel and MATERIAL. This paper details the philosophy and objective behind each program to advance technologies in this area and introduces the program corpora available to stimulate related research.

Keywords: Low Resource Languages, Cross-language Information Retrieval, Machine Translation, ASR, corpora, Cross-language summarization, machine learning, speech technology, AQWV

Résumé

Iti gandat ti panagpadur-as iti agdama a kasasaad ti automatiko a panagproseso ken panagsursuro kadagiti pagsasao no agkurang ti datos a maikurri, insayangkat ti IARPA ti dua a dakkel a programa ti panagsukisok a nagpaut iti sumagmamano a tawen: Babel ken MATERIAL. Iburayko ditoy ti panggep ken pilosopia dagiti dua a programa iti panangparang-ay kadagiti kastoy a teknolohia. Ilawlawagko met ti amin a korpora dagiti programa nga imbunong ti IARPA tapno makadur-as pay ti panagsukisok iti daytoy a tay-ak.

1. Introduction

Advances in natural language processing (NLP) have continued to progress substantially since the advent and application of deep learning methods. Several applications in this realm have benefited from breakthroughs resulting from sustained refinements and methodological evolution. However, despite impressive progress in performance for multiple human language technology areas, the gap in performance between English and other languages suggests that the application of these novel techniques to new languages does not necessarily portend success. Deep learning methods consistently require much more data than is usually available for the majority of the world's languages. Moreover, deep learning methods are often impacted by noise in training data more than traditional machine learning methods. Recognizing the need to improve human language technologies for lower resource languages, the Intelligence Advanced Research Projects Activity (IARPA) invested in multiple relevant research endeavors.

To address unsolved problems deemed important to the Intelligence Community, IARPA, the research wing of the U. S. Office of the Director of National Intelligence, employs a competitive bid process to mobilize the best talent worldwide to work on our research programs. That competition process is our solicitation of proposals in response to a Broad Agency Announcement (BAA) of our research programs. Typically, we fund multidisciplinary teams to address the research comprehensively. Selected teams work collaboratively and competitively to tackle the challenges and advance our understanding of the problems, frequently with cross-disciplinary solutions. To propel research toward solutions that work, IARPA measures Performer Team progress via a rigorous and well-defined metric-based evaluation that is often a product of both initial foresight and lessons learned from active engagement in the research activities.

In the realm of human language technology, IARPA has launched a variety of research initiatives ranging from small-term studies and seedlings (research efforts of less than a year) to five multi-year research programs. Of the five programs, three are complete: SCIL (Socio-cultural Content in Language for social role and goal discovery), METAPHOR (analysis of metaphor to gain insight into interpreting cultural norms), Babel (Speech Recognition); and two are currently ongoing: MATERIAL (Machine Translation and Cross-language information Retrieval) and BETTER (Cross-language Information Extraction and Retrieval). We will only cover MATERIAL and Babel here, as they involve research with low resource languages in low resource conditions. These endeavors greatly expanded the IARPA portfolio of complex, multidisciplinary programs[1], but most importantly for the NLP community, pioneered a new evaluation paradigm to measure NLP progress, and provided several large annotated datasets accessible to the community to encourage continued research in this area.

[1] IARPA has a diverse research portfolio encompassing research from a variety of disciplines, including math, physics, linguistics, biology, neuroscience, political science, and cognitive psychology.

2. The Babel Program

The Babel program, the brainchild of Dr. Mary P. Harper, was launched in 2011, to address two technological gaps in speech technology (Harper 2011). At the program's inception, mature technologies for low resource languages to process speech in a meaningful way for keyword search (KWS) was non-existent, and the time and resources required for system development to address this problem were beyond the reach of most research institutions. To address these shortcomings, IARPA contracted multiple multinational "performing teams" of experts who competed to develop agile and robust methods for supporting effective keyword search over massive amounts of recorded- speech in foreign languages.

	BP	OP1	OP2	OP3
Practice Languages	**4 languages:** Cantonese, Pashto, Tagalog, Turkish	**5 languages:** Assamese, Bengali, Haitian Creole, Lao, Zulu	**6 languages:** Cebuano, Kazakh, Kurdish, Lithuanian, Telugu, Tok Pisin	**7 languages:** Amharic, Dholuo, Guarani, Igbo, Javanese, Mongolian, and Pashto (revisited)
Surprise Language	Vietnamese in **4 weeks**	Tamil in **3 weeks**	Swahili in **2 weeks**	Georgian in **1 week**
Training Data Limits	80 hours, 10 hours (with dictionary)	**60 hours**, 10 hours (with dictionary)	**40 hours, 3 hours, Select 3 hours (no dictionary)**	40 hours (no dictionary)
Recordings	Mixed Environment Telephone	Mixed Environment Telephone **& Microphone (2 types)**	Mixed Environment Telephone **& Microphone (2 types)**	Mixed Environment Telephone **& Microphone (7 types)**
ATWV Target	0.3 or greater	0.3 or greater	0.3 or greater	**0.6 or greater**
WER Target	N/A	N/A	N/A	**50% or less**
Meet Target with:	80 hours	**60 hours, 10 hours**	**40 hours, 3 hours, Selected 3 hours**	40 hours
What was in the BAA?	80 hours (with dictionary)	60 hours (with dictionary)	40 hours (with dictionary)	40 hours (with dictionary)

Figure 1 Babel languages and targets per period

As a way to ensure efficacy of the search tools, and portability of the methods used to build them, IARPA used diverse languages from multiple language families, and real-world recording conditions in the training and evaluation. Data were collected and consistently transcribed in-country using normalized conventions. Extensive vetting of the resources was performed by the University of Maryland's Center for Advanced Study of Languages (CASL), with particular attention to languages with less standardized orthographies. IARPA contracted MIT Lincoln Labs as a Test and Evaluation partner to process the corpora for effective evaluation, and develop speech recognition reference systems to baseline pre-program state-of-the-art performance to set meaningful metric thresholds for the performing teams. The National Institute of Standards and Technology (NIST) was tasked with developing the program evaluation metric and with conducting regular evaluations of the performing systems.

Automatic Speech Recognition (ASR) training data created for the program included 40-80 hours of transcribed speech in each language, provided to program participants upon the kickoff of each new language to jump start their research. Datasets used in Babel consisted of natural conversations of at least two thousand speakers, recorded under various microphone and service provider conditions (See Figure 1).

The metric used to evaluate progress towards the program goal was ATWV (Actual Term Weighted Value), a detection metric that, for each foreign keyword query, awards true hits and penalizes false alarms and misses, with relative weights set by IARPA, then averages the scores over an entire query set[2]. This detection metric was deemed most suitable to measure performance as it ensured each query would have equal weight regardless of its relevance probability. System developers would choose a decision threshold of their probability computations for their "Actual" decisions to maximize this metric, and IARPA would use these scores to determine progress and the status of each team in the program. In the fourth and final program period, KWS systems were also expected to achieve a word error rate (WER) of fifty percent or less.

As Babel progressed, teams faced increasingly more challenging program goals in shorter lengths of time and with less training data, leveraging approaches they learned to mitigate issues associated with resource and transcription dearth. Multilingual features and acoustic models proved effective, as well as grapheme based acoustic modeling and cascaded adaptor grammars to better resolve new words not seen in the training data.

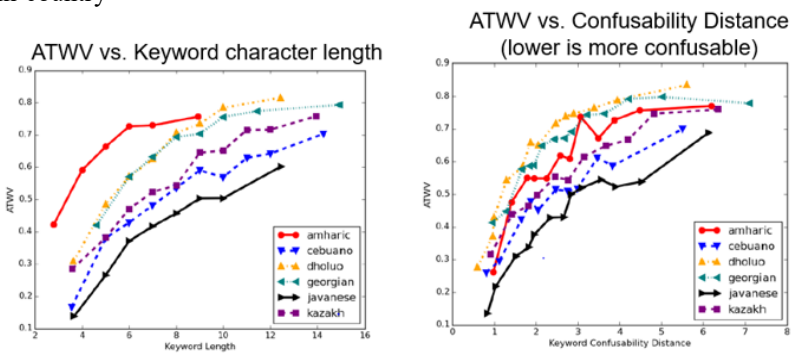

Figure 2 ATWV Correlations as reported by BBN

IARPA and Performing Teams investigated multiple issues to understand differential progress between systems, configurations and languages (Hartmann et al 2017). A number of variables seemed to correlate with overall system performance. Variables investigated included word error rate performance, average word length, keyword

[2] OpenKWS Evaluation Plan, https://bit.ly/36z6BsR

confusability distance (weighted keyword length where lower values are more confusable, See Figure 2), graphemic error rate calculated against the training data (Figure 3), speaking rate (phones per second), keyword frequency, and inter-annotator agreement in the transcriptions. Measuring these details was not just critical to understanding of performance degradations to allow research teams to appropriately compensate for them, but also to inform the program manager of optimal language choices for subsequent periods of the program based on predicting expected performance, as well as query construct design.

To optimize the methods used to both advance the science and evaluate performance, a good program design was critical. A strategic and meaningful choice of languages proved to be most important. A number of criteria were considered prior to language selection, ranging from language-specific features (phoneme inventory, morphological complexity, orthography, syntax, dialectal variation, typological uniqueness, and genetic relationships) to accessibility and the cost of the collection. For the latter measure, Dr. Harper as the Program Manager took into account the number of native speakers, quality of phone connections, availability of linguistic expertise for transcription, and the political stability of the region to assure a safe in-country collection. In addition, it was strongly desired that the languages released in Babel were not spoken by members of the Performer Teams.

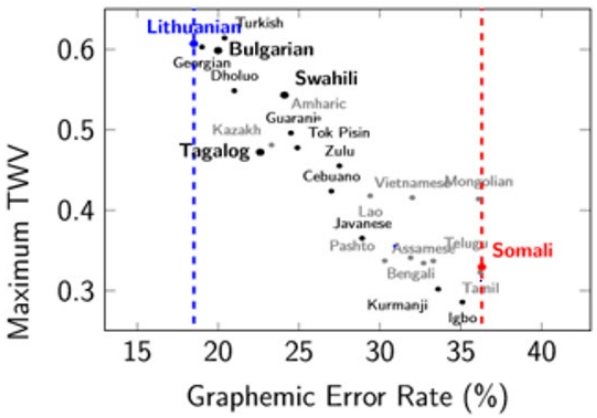

Figure 3 GER was a strong predictor of performance

Prior to each data collection, a "Language Specific Peculiarities" (LSP) document describing unique and essential properties of each language was produced to ensure consistency in translation, a well-balanced collection in terms of dialect coverage, and to optimize keyword selection for the final evaluation. Teams were also provided with pronunciation lexicons for words in the training set. These lexicons were converted to a finite state transducer (FST) so each orthographic word would be mapped onto a set of reasonable pronunciations. The Test and Evaluation team also created confusion matrices for words whose pronunciations were similar but not identical. The LSP documents, pronunciation lexicons, and training data for the twenty six languages released in this program are available from IARPA, and via the University of Pennsylvania Linguistic Data Consortium, http://-ldc.upenn.edu

3. The MATERIAL Program

Leveraging lessons learned from Babel, the MATERIAL (Machine Translation for English Retrieval of Information in Any Language) program was launched in 2017 to evaluate performance on a much wider array of human language technologies to include cross-language information retrieval and summarization (Rubino 2017). Performers on this program built systems that can retrieve foreign language speech and text documents responsive to domain-constrained English queries, and provide evidence or relevance, in English, to both the query string and its domain. This novel evaluation measured performance, not on each underlying technology involved, but on two functional and unified End-to-End capabilities in a way these technologies were had not been evaluated before. The first performance score, AQWV (Averaged Query Weighted Value), measured the effectiveness of cross-language information retrieval systems. The second, measured the End-to-End performance (E2E AQWV) of systems as judged by humans on their cross-language summarization capability with a novel crowd-sourced evaluation methodology utilizing the Amazon Mechanical Turk platform designed and executed by Tarragon Consulting[3]. CLIR AQWV scores could be improved in the E2E evaluation if summarization systems provided enough evidence for the human judges to reject false alarms. However, the CLIR AQWV scores could also degrade if the summaries were not of sufficient quality to convince the judges to retain true positive documents.

To effect this evaluation paradigm, unique datasets were compiled to include domain-annotated documents in six genres, queries of various types designed to probe various analytic dimensions, and relevance decisions for each query against the program documents. The query typology developed for MATERIAL allowed IARPA to probe each system's ability to handle ontological concepts, and to resolve ambiguity resulting from polysemy, homophony, and/or named entities.

As of December 2019, five languages were released for evaluation. For the Phase I period languages (Swahili, Tagalog and Somali), systems were also required to identify eight domains: Government and Politics, Lifestyle, Business and Commerce, Law and Order, Physical and Mental Health, Military, Sports, and Religion. For Phase II, three additional languages will be evaluated: Lithuanian, Bulgarian, and a surprise language to be released in January 2020. IARPA plans to release 3 more languages in the final Phase III of the program.

Like the Babel program, language expertise procurement was disallowed to drive approaches that leverage machine learning. Likewise, each period of the program had a development stage in which performing teams worked with multiple "practice languages" to develop their methods. The development stage was followed by an evaluation stage, in which progress was officially evaluated on a

[3] MATERIAL Evaluation Plan, https://bit.ly/2oRT923

"surprise language". The evaluation dataset was partitioned into three temporal epochs, corresponding to different query sets and new domain releases. Cross-language query-biased summaries, providing relevance justification for each retrieved document, were evaluated at the end of each phase by English-speaking crowd sourced judges. This fundamental development and evaluation cycle is illustrated in Figure 4. For Phase I of the program, IARPA released Swahili and Tagalog as the practice languages. The surprise language was Somali.

MATERIAL documents were collected in two modes (text and speech), and six genres (news text, monologic social media text, and topical text, conversational speech, broadcast news audio, and topical audio). A challenging aspect of the evaluation was the planned mismatch condition between the training and evaluation conditions.

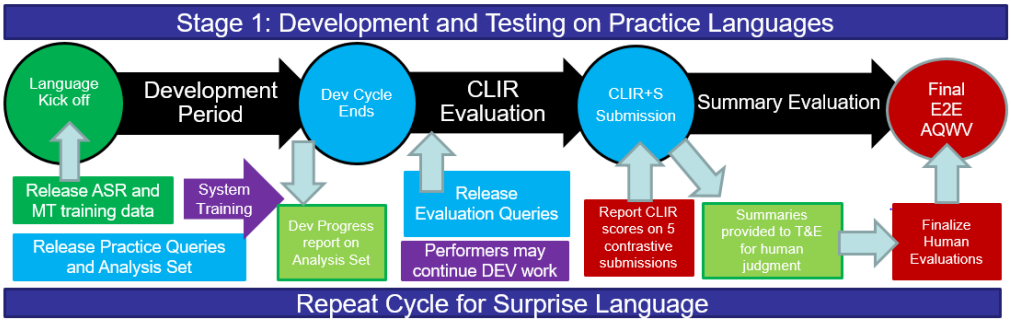

Figure 4 Development and Evaluation Cycle for each MATERIAL Language

At the release of each new language, IARPA distributed "Training Build Packs". These consisted of fifty hours of transcribed telephony for ASR development, phonetic lexica, a parallel corpus for MT training consisting of approximately 800K English words from news texts translated and aligned at the sentence level, and a language description that detailed language-specific peculiarities pertaining to both the text and audio collection, transcription, and the language itself. To challenge teams to create portable solutions, not all speech and text genres in the evaluation condition were represented in the training data. Teams developed novel methods to cope with the mismatch. Considering the dearth of training provided for each language kickoff, teams were expected to complement the IARPA-provided build packs with their own data harvests.

Beyond the training datasets, the remaining documents were partitioned by NIST and MIT Lincoln Laboratories into three other sets, optimized to achieve a reasonable training/evaluation balance. A "devtest" set of documents with relevance annotations for each query was provided to each team to allow them to locally score their systems on CLIR via the AQWV metric. Teams were discouraged from using this set for training to better understand the results of component engine work and joint process optimization. An analysis set of documents was provided with the devtest and after each evaluation for glass box testing to analyze errors. The analysis set was fully translated; audio documents in this set were transcribed with the conventions employed in the training build pack. Teams were allowed to closely scrutinize this set to help diagnose ASR, MT and CLIR errors and understand progress. Finally, a blind evaluation set was divided temporally into three epochs, and scored at NIST to reveal final results.

Although language-independent machine learning approaches are desirable to effect quick capability ramp up, it was obvious that not all languages are created equal. Different strategies were employed per language to optimize AQWV. Each system analysis yielded interesting observations. We found that vowel removal strategies that helped Tagalog, hurt Swahili and Somali. Translation lattices that helped Somali and Tagalog did little to improve results on Somali. Stemming during retrieval and translation benefited Swahili and Somali but not Tagalog. IARPA also investigated dataset peculiarities per language. Vocabulary growth, OOV (out of vocabulary) words in evaluation not in training, sentence perplexity and audio clipping were measured as factors that may affect performance. It was immediately evident that the high vocabulary growth rate in Lithuanian did not result in lower AQWV scores. The relative wealth of resources available for this language compensated for this perceived difficulty of the expanding vocabulary. The next language to be released will offer additional challenges to include rampant orthographic and dialectal variation.

4. Conclusion

With proper design resulting from well-informed planning and sustained collaboration with expert scientific teams, the U.S. Government can launch programs to tackle seemingly impossible challenges in NLP. Teams in the Babel program proved that an effective keyword search capability can indeed be developed in one week. MATERIAL teams are exploring novel methods to enable cross-lingual semantic search of text and audio outside the traditional realm of machine translation under realistic constraints that mirror the current problem space. It is our goal that the insights and lessons we have learned from our investments and work are applied by the community as we propel the research to take on more ambitious problems in the future. IARPA invites the research community to learn from both our progress and mistakes, and to profit from the datasets we will disseminate to see how much further they can go. IARPA also welcomes the community to propose high-risk high-reward ideas for new research in the NLP domain.

5. Acknowledgements

My thanks to Ilya Zavorin and Catherine Cotell for their helpful comments on a previous draft of this paper.

6. Bibliographical References

Harper, M. (2011). Babel BAA. https://bit.ly/2KvdDoY.

Hartmann, W, D. Karakos, R. Hsiao, L. Zhang, T, Alum. and Gertz, M. Alumäe, S. Tsakalidis and R. Schwartz (2017). Analysis of Keyword Spotting Performance Across IARPA Babel Languages, ICASSP'17, pages 5765-5769.

Rubino, C. (2017). Material BAA. https://bit.ly/37gKhV9.