

MATERIAL

MACHINE TRANSLATION FOR ENGLISH RETRIEVAL OF INFORMATION IN ANY LANGUAGE

INTELLIGENCE VALUE

The MATERIAL program aims to revolutionize the way the IC consumes foreign language information, by turning multilingual text and speech media into useable intelligence information for analysts, regardless of their language expertise.

A large portion of the ever-increasing amounts of text, audio, and video data produced in today's world is being generated by populations of emerging importance in lower-resource languages. This rich source of data is of little value if the information cannot be effectively searched. Launched in October 2017, The MATERIAL program is a 47-month venture seeking to address this challenge by building robust, automated language capabilities with limited linguistic resources, expertise, and tools.

MATERIAL's ultimate goal is to build a Cross-Language Information Retrieval (CLIR) systems that find speech and text content in diverse lower-resource languages, using English search queries. This system will allow analysts to submit queries in English and receive short English summaries of relevant foreign language items that saliently display relevance to their information needs.

Success is measured by a novel end-to-end retrieval metric that will assess the system's ability to retrieve all relevant documents, while producing few false alarms. The MATERIAL program will provide:

- State-of-the-art Automatic Speech Recognition (ASR) and Machine Translation (MT) systems and models for Tagalog, Swahili, Somali, Lithuanian, Bulgarian, Pashto, Farsi, Kazakh, and Georgian
- Highly competitive models optimized for informal and formal speech and text
- Novel end-to-end CLIR systems available with dockerized, exchangeable component technologies
- Innovative ways of utilizing large amounts of diverse multilingual unstructured text and speech data to improve model performance
- Text processing tools to address morphology and divergent spelling
- Text, audio, and video data crawlers operationalized for U.S. government use
- Annotated, reusable datasets in multiple languages for CLIR, ASR, and MT research

- Open-source neural MT framework Marian (co-founded)
- Bitext harvesting tool Paracrawl (funded)
- Novel cross-language query-biased summary generation technique

PRIME PERFORMERS

- Raytheon BBN Technologies
- Columbia University
- University of Southern California Information Sciences Institute
- Johns Hopkins University

TESTING AND EVALUATION PARTNERS

- Massachusetts Institute of Technology Lincoln Laboratory
- National Institute of Standards and Technology
- University of Maryland Applied Research Laboratory for Intelligence and Security
- Tarragon Consulting Corporation

END-TO-END MATERIAL SYSTEM

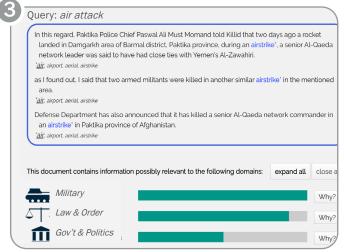


Analysts Query Foreign Language Speech and Text database using English

"air attack" in Pashto Speech and text documents

All domains

Query in English



Results Summarized in English



Must Momand told Killid that two days ago a rocket landed in Damgarkh area of Barmal کې دبې پېلوټه الوتکې ديوه <mark>هوايي بريد</mark> ترڅ کې د district, Paktika province, during an airstrike senior Al-Qaeda network leader was said to have had close ties with Yemen's Al-Zawahir

Aligned Bilingual Full Document Output



PROGRAM MANAGER

Tim McKinnon

Phone: (301) 243-2084 timothy.mckinnon@iarpa.gov



