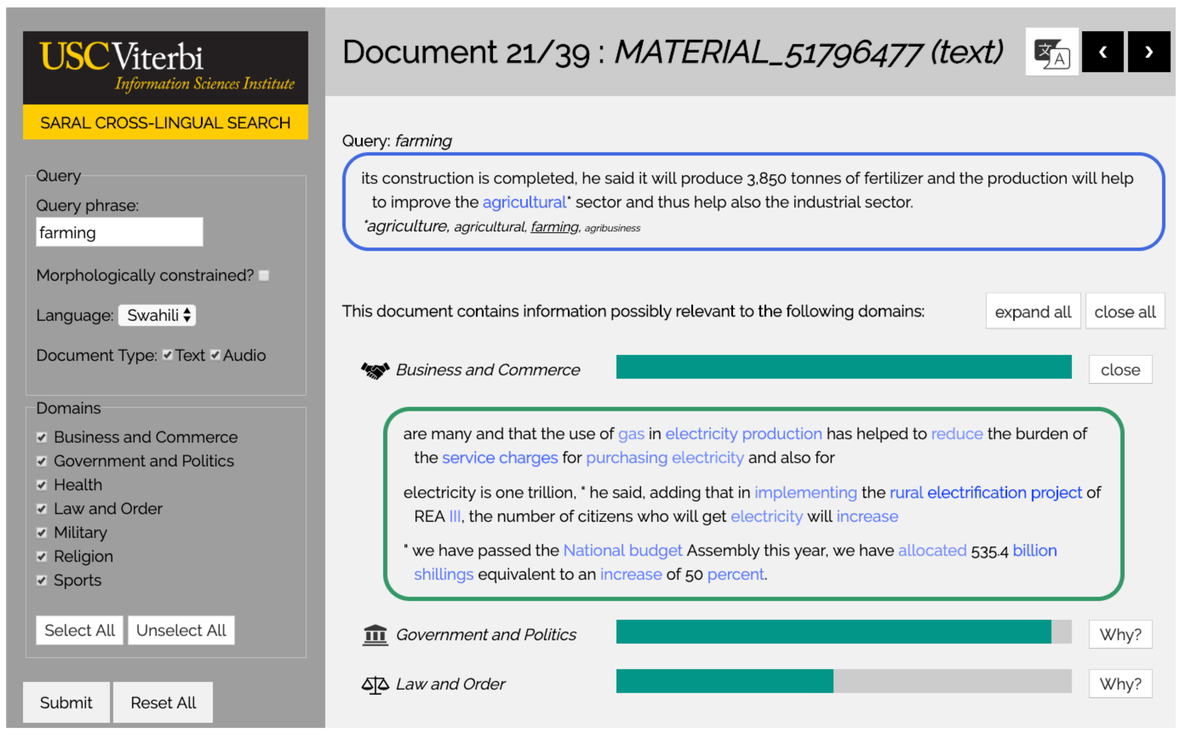
SARAL: A Low-Resource Cross-Lingual Domain-Focused Information Retrieval System for Effective Rapid Document Triage

Elizabeth Boschee, Joel Barry, Jayadev Billa, Marjorie Freedman, Thamme Gowda, Constantine Lignos, Chester Palen-Michel, Michael Pust, Banriskhem Kayang Khonglah, Srikanth Madikeri, Jonathan May, and Scott Miller

University of Southern California Information Sciences Institute, Idiap Research Institute

Goal: Enable English search for text and audio from lower-resourced languages

- Give supporting evidence in English for document relevance to query and domains (topics)
- Provide complete transcriptions and translations
- Models trained on 1.6m words of parallel and 40 hours of transcribed data per language pair



Querying

- Select source language and text/audio/both
- Queries can be morphologically constrained, e.g. to require plural nouns or past tense verbs
- Can filter to select documents estimated to be relevant to pre-specified domains (e.g. health)

Retrieval

- Retrieval matches simultaneously using English via multiple MT outputs, source language via translation tables, and shared embedding space via SEARCHER
- Variety of term expansions used for English retrieval, e.g. stemming, WordNet, paraphrases (PPDB), nearest neighbors in embeddings, etc.
- Can match multi-term queries by using multiple mechanisms simultaneously, e.g. query *rainy season* might match *rainy* in MT and *xilli* (tr: *season* or *time*) in source language via translation tables

Speech Recognition (ASR)

- Two Kaldi-based LF-MMI ASR systems, trigram LM rescored with RNN-LM
- Idiap system combines 3 CNN-BLSTM models
- ISI system combines 8 TDNN-F models

Machine Translation

- System combination of Transformer and syntax-based MT, training on <2M parallel words
- Transformer uses additional 14.5M words of backtranslated region-relevant scraped English

SEARCHER

- Maps terms from queries and documents into shared embedding space
- Uses convolutional encoder to contextually encode source terms and attention mechanism to match them to query terms
- Trained using sentence relevance paradigm, where source language sentence is relevant to term *t* if the parallel English sentence contains *t*

Domain Classification

- Use NYT Corpus for topic annotations, mapping corpus topics to domains of interest
- Identify uni/bi/tri-grams that are representative of each domain by TF-IDF-like measure

Evidence Generation

- Use document excerpts with highest CLIR scores to concisely highlight why document is relevant
- Use footnotes providing alternate translations for key words to compensate for noisy MT

Demo Capabilities

- Currently loaded indices: Somali and Swahili
- Other tested languages: Bulgarian, Lithuanian, and Tagalog (more to come over time)
- Top end-to-end performance in the most recent IARPA MATERIAL CLIR+summarization evaluations

Adding a New Language

- Add basic support for new language in ~3 days
- Improved system in <10 days, additional time primarily needed for text/audio scraping

Try it out!

- Open https://material.isi.edu/ on a desktop/laptop
- Register using token PpnOMgavHR3j

Contact: boschee@isi.edu

Acknowledgments

Thanks to Heng Ji for fruitful discussions. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract FA8650-17-C-9116. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA,or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

USC Viterbi Information Sciences Institute

idiap RESEARCH INSTITUTE