MATERIAL:
PUBLICLY AVAILABLE PROGRAM ARTIFACTS

Intelligence Advanced Research Projects Activity
IARPA
Creating Advantage through Research and Technology

TITLE	DESCRIPTION	CREATOR	REFERENCE	LOCATION
Kaldi models for MATERIAL Languages	ASR models for Swahili, Somali and Tagalog	JHU		*https://github.com/kaldi-asr/kaldi/tree/master/egs/material*
Sentence Filtering for noisy data	Data and tools related to research and development of novel sentence filtering methods	JHU	*https://arxiv.org/pdf/1805.12282.pdf*	*http://www.statmt.org/wmt19/parallel-corpus-filtering.html*
Paracrawl	The main goal of the ParaCrawl project is to create the largest publicly available corpora by crawling hundreds of thousands of websites, using open source tools. As part of this effort, several open source components have been developed and integrated into the open-source tool Bitextor, a highly modular pipeline that allows harvesting parallel corpora from multilingual websites or from preexisting or historical web crawls such as Common Crawl or the one available as part of the Internet Archive. The processing pipeline consists of the steps: crawling, text extraction, document alignment, sentence alignment, and sentence pair filtering.	JHU, University of Edinburgh, among others		*https://paracrawl.eu/about*
Marian NMT	An efficient, free Neural Machine Translation framework written in pure C++ with minimal dependencies. It is mainly being developed by the Microsoft Translator team. Many academic (most notably the University of Edinburgh and in the past the Adam Mickiewicz University in Poznań) and commercial contributors help with its development. It is currently the engine behind the Microsoft Translator Neural Machine Translation services and being deployed by many companies, organizations and research projects, including some USG agencies. Its main features include (i) an efficient pure C++ implementation, (ii) fast multi-GPU training and GPU/CPU translation, (iii) state-of-the-art NMT architectures: deep RNN and Transformer, and (iv) permissive open source license (MIT).	Microsoft, University of Edinburgh, among others		*https://marian-nmt.github.io*
The Large-Scale CLIR Dataset	A retrieval dataset built for Cross-Language Information Retrieval (CLIR). The dataset is derived from Wikipedia and contains more 2.8 million English single-sentence queries with relevant documents from 25 other selected languages. It supports cross-lingual learning to rank.	JHU	*https://www.cs.jhu.edu/~kevinduh/papers/sasaki18letor.pdf*	*https://www.cs.jhu.edu/~kevinduh/a/wikiclir2018/*
TFMTL: Tensor Flow Framework for Multitask learning	TFMTL is a full-pipeline framework for multi-task learning text classification tasks developed in TensorFlow. You can download formatted text datasets, preprocess datasets, configure the embedding/encoding/FFN architectures by running a few scripts and modify the configurations in some JSON files. You can also easily add your own modules following the standard input/output.	JHU	*https://aclanthology.org/D19-6105.pdf*	*https://github.com/felicitywang/TFMTL*
RTG NMT	Reader-Translator-Generator (RTG) is a Neural Machine Translation toolkit based on PyTorch which supports Transformer NMT models and is the foundation of the 500-to-English and 600-to-English MT systems (*http://rtg.isi.edu/many-eng/*), as well as fine-tuned MT models for the nine MATERIAL languages (*http://rtg.isi.edu/many-eng/models/MATERIAL/*)	USC-ISI	*https://isi-nlp.github.io/rtg/*	*https://github.com/isi-nlp/rtg*
Vista NMT	Convolutional sequence-to-sequence software for machine translation based on Tensorflow.	USC-ISI		*https://github.com/isi-vista/VistaMT*

MATERIAL:
PUBLICLY AVAILABLE PROGRAM ARTIFACTS

Intelligence Advanced Research Projects Activity
IARPA
Creating Advantage through Research and Technology

TITLE	DESCRIPTION	CREATOR	REFERENCE	LOCATION
Universal Romanizer (UROMAN)	A tool implemented in Perl that converts text in any script to the Latin alphabet.	MATERIAL SARAL Team	*https://www.isi.edu/~ulf/uroman.html*	*https://github.com/isi-nlp/uroman*
Universal Tokenizer (UTOKEN)	A multilingual tokenizer implemented in Python that divides text into words, punctuation and special tokens such as numbers, URLs, XML tags, email-addresses and hashtags. The tokenizer comes with a companion detokenizer.	MATERIAL SARAL Team		*https://github.com/uhermjakob/utoken*
MCSS: Multi-lingual Common Semantic Space	Knowledge-empowered multilingual common semantic space framework that had been applied to Machine Translation, name tagging, unsupervised cross-lingual entity linking, Cross-lingual Structure Transfer for relation and event extraction, and Knowledge Graph Completion. Released are several resources for up to 282 languages: Knowledge-empowered Embeddings, parallel sentences mined from our embeddings, automatically generated training data and name taggers trained from them. Given a document in any of the supported languages, this framework is able to identify name mentions, assign a coarse-grained or fine-grained type to each mention, and link it to an English Knowledge Base if it is linkable.	UIUC	*https://blender.cs.illinois.edu/paper/multilingual_commonspace.pdf*	*http://blender.cs.illinois.edu/software/saral/*
Morphagram	An unsupervised morphological segmentation framework implemented in Python that is language independent. The input to the framework is an unlabeled list of words, and the output is the morphological segmentation of the input words and a grammar that parses unseen words. Adding linguistic knowledge (in terms of prefixes and suffixes) is an option.	MATERIAL SCRIPTS Team	*https://aclanthology.org/2020.lrec-1.879.pdf*	*https://github.com/rnd2110/MorphAGram*
Morphological Analyzer and POS Tagger	A morphologicalanalysis neural system implemented in Python that does morphological segmentation and morphological tagging for all MATERIAL languages (Swahili, Tagalog, Somali, Lithuanian, Bulgarian, Pashto, Farsi, Kazakh and Georgian). The analyzer can be executed through a runnable JAR or as a Docker image, in either a standalone mode or a clientserver one. For each word in a given context, the analyzer produces the following information: (i) word part-of-speech (universal POS tagset), (ii) word segmentation (prefixes, stem and suffixes), (iii) word Tense (past/present), and (iv) word Number (singular/plural). The tagger is also available in a non-neural version that is based on Averaged Perceptron.	MATERIAL SCRIPTS Team	*https://aclanthology.org/2020.emnlp-main.391.pdf*	*https://github.com/rnd2110/unsupervised-cross-lingual-POS-tagging*
Text Normalization	A normalization system implemented in Python that performs text cleanup, transliteration and a set of other operations to control numbers, punctuation marks, repetitions and foreign text in the nine MATERIAL languages (Swahili, Tagalog, Somali, Lithuanian, Bulgarian, Pashto, Farsi, Kazakh and Georgian)	MATERIAL SCRIPTS Team		*https://github.com/rnd2110/SCRIPTS_Normalization*
Summarizer	This component takes as input a query and a set of foreign language documents (and their translations) for any one of the MATERIAL languages (Swahili, Tagalog, Somali, Lithuanian, Bulgarian, Pashto, Farsi, Kazakh and Georgian), that have been determined by CLIR to be relevant to the query. The summarizer produces a 100 word English word summary of the document for each component of the query. It must be called from within the SCRIPTS CLIR component.	MATERIAL SCRIPTS Team		*https://github.com/eturcan/scripts*