AFRL-RH-WP-TR-2019-0115

QUICKLY ADAPTED SPEECH AND TRANSLATION ENABLED INFORMATION RETRIEVAL (QuickSTIR)

Philipp Koehn
Sanjeev Khudanpur / Kevin Duh
Ben Van Durme / Jonathan Wintrode

The Johns Hopkins University
3400 North Charles Street
Baltimore MD 21218-2608

December 2019

FINAL REPORT

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING
AIRMAN SYSTEMS DIRECTORATE
WARFIGHTER INTERFACE DIVISION
WRIGHT-PATTERSON AFB OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

Qualified requestors may obtain copies of this report from the Defense Technical Information Center (DTIC).

AFRL-RH-WP-TR-2019-0115 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

SLYH.RAYMOND.E.1230265794 Digitally signed by SLYH.RAYMOND.E.1230265794 Date: 2020.11.10 09:08:34 -05'00'

RAYMOND E. SLYH, DR-IV, Ph.D.
Work Unit Manager
Mission Analytics Branch
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

MURDOCK.WILLIAM.P.1048742161 Digitally signed by MURDOCK.WILLIAM.P.1048742161 Date: 2020.11.10 10:55:43 -05'00'

WILLIAM P. MURDOCK, DR-IV, Ph.D.
Chief, Mission Analytics Branch
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

CARTER.LOUISE.ANN.1230249128 Digitally signed by CARTER.LOUISE.ANN.1230249128 Date: 2020.11.23 15:32:57 -05'00'

LOUISE A. CARTER, DR-IV, Ph.D.
Chief, Warfighter Interface Division
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704–

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
13-12–2019	Final	26 September 2017 – 31 October 2019

4. TITLE AND SUBTITLE

Quickly Adapted Speech and Translation Enabled Information Retrieval (QuickSTIR)

5a. CONTRACT NUMBER

FA8650–17–C–9115

5b. GRANT NUMBER

N/A

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Philipp Koehn
Sanjeev Khudanpur
Kevin Duh
Ben Van Durme
Jonathan Wintrode

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

H0UU

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The Johns Hopkins University
3400 North Charles Street
Baltimore MD 21218-2608

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Materiel Command
Air Force Research Laboratory
711 Human Performance Wing
Airman Systems Directorate
Warfighter Interface Division
Wright-Patterson AFB, OH 45433

10. SPONSOR/MONITOR'S ACRONYM(S)

711 HPW/RHCML

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

AFRL-RH-WP-TR-2019-0115

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

This work was declared to be Fundamental Research prior to contract award. 88ABW-2020-3288; Cleared 23 Oct 2020

14. ABSTRACT

This report provides research results in the areas of Automatic Speech Recognition (ASR), MT (MT), Cross-Lingual Information Retrieval (CLIR), Domain Classification, and Summarization in the context of low-resource training settings for the IARPA MATERIAL Program.

15. SUBJECT TERMS

Automatic Speech Recognition (ASR), MT (MT), Cross-Lingual Information Retrieval (CLIR), Domain Classification, Summarization

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Raymond E. Slyh
U	U	U	SAR	32	19b. TELEPHONE NUMBER *(include area code)* N/A

Standard Form 298 (Rev. 8–98)
Prescribed by ANSI Std. Z39.18

TABLE OF CONTENTS

LIST OF FIGURES

SUMMARY

This document provides a summary of work completed by researchers at Johns Hopkins University (JHU) and Raytheon in the project "*Quickly adapted Speech and Translation enabled Information Retrieval"* (QuickSTIR), as part of the Intelligence Advanced Research Projects Activity (IARPA) Machine Translation for English Retrieval of Information in Any Language (MATERIAL) program. This work was performed over the period 1 October 2017 to 30 September 2019 under contract FA8650–17–C–9115.

Basic research was carried out on automatic speech recognition (ASR), machine translation (MT), cross-lingual information retrieval (CLIR), domain classification, and summarization under low-resource conditions. An integrated open-source system was developed that allows the training of all components. The system was evaluated in two evaluation campaigns (dry run and final evaluation), achieving the third best result among four teams in the final evaluation. The project also assisted in data collection for the final phase of the MATERIAL program.

1.0 INTRODUCTION

This document provides a summary of work completed by researchers at Johns Hopkins University (JHU) and Raytheon in the project "*Quickly Adapted Speech and Translation enabled Information Retrieval"* (QuickSTIR), as part of the Intelligence Advanced Research Projects Activity (IARPA) Machine Translation for English Retrieval of Information in any Language (MATERIAL) program. This work was performed over the period 1 October 2017 to 30 September 2019 under contract FA8650–17–C–9115.

Basic research was carried out on automatic speech recognition (ASR), Machine Translation (MT), cross-lingual information retrieval (CLIR), domain classification, and summarization under low-resource conditions. An integrated open-source system was developed that allows the training of all components. The system was evaluated in two evaluation campaigns (dry run and final evaluation), achieving the third best result among four teams in the final evaluation. The project also assisted in data collection for the final phase of the MATERIAL program.

2.0 EXPERIMENTS AND ACCOMPLISHMENTS

2.1 Automatic Speech Recognition (ASR)

We developed ASR systems for Swahili, Tagalog, and Somali using the state-of-the-art algorithms we researched in the program. Our recipe is publicly available at GitHub

https://github.com/kaldi-asr/kaldi/tree/master/egs/material

The highlights of our accomplishments include Time Delay Neural Networks Factored (TDNN-F) training for acoustic modeling, semi-supervised acoustic modeling, and Recurrent Neural Net Language Model (RNNLM) rescoring for language modeling.

2.1.1 TDNN-F training for Acoustic Modeling

We introduced TDNN-F, a new neural network architecture for speech recognition. TDNN-F is a factored form of TDNNs which is an efficient and well-performing architecture for speech recognition. TDNN-F is the same as a TDNN --- with layers compressed via Singular Value Decomposition (SVD) --- that is trained from a random start with one of the two factors of each matrix constrained to be semi-orthogonal. The Word Error Rate (WER) results from TDNN-F models are often better than the TDNN- Long Short-Term Memory Network (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) results, while being much faster to decode.

- Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi and Sanjeev Khudanpur: *Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks*, Interspeech 2018

2.1.2 Semi-Supervised Acoustic Modeling

In addition to supervised training, we ran semi-supervised training of acoustic models using the extension of lattice-free Maximum Mutual Information (MMI) to semi-supervised scenarios (Manohar et al., 2018). This method involves various ways of creating supervision from lattices obtained from decoding of unsupervised data. We added unlabeled audio --- from the EVAL sets --- to the labeled audio in the training set to train the acoustic model. Semi-supervised acoustic modeling improved our results for 12-16% absolute WER.

2.1.3 RNNLM Rescoring for Language Modeling

We proposed methods to support neural-based language modeling for ASR. RNNLM rescoring for language modeling improved our results for 0.5-2% absolute WER.

In one study, we combined the use of subword features (letter n-grams) and one-hot encoding of frequent words so that the models can handle large vocabularies containing infrequent words. We proposed a new objective function that allows for training of unnormalized probabilities. An importance sampling-based method is supported to speed up training when the vocabulary is large:

- H. Xu, Ke Li, Y. Wang, J. Wang, S. Kang, X. Chen, D. Povey and S. Khudanpur: *Neural network language modeling with letter-based features and importance sampling*, Institute of Electrical and Electronics Engineers (IEEE) International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018

In another study, we introduced a pruned lattice-rescoring algorithm that improves ASR speed

and accuracy. In a common lattice-rescoring approach in ASR, a word-lattice is generated from 1st-pass decoding and the lattice is then rescored with a neural model, and an n-gram approximation method is usually adopted to limit the search space. Our pruned lattice-rescoring algorithm improved the n-gram approximation method to further limit the search space and uses heuristic search to pick better histories when expanding the lattice.

- H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey and S. Khudanpur: *A pruned RNNLM lattice-rescoring algorithm for automatic speech recognition*, IEEE ICASSP, 2018

Besides the research results we directly used in developing our ASR systems in the MATERIAL recipe, we conducted further research on speaker recognition and end-to-end ASR systems, which were published in top-tier conferences of the field.

- David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey and Sanjeev Khudanpur. *Speaker Recognition for Multi-Speaker Conversations Using X-Vectors*, IEEE ICASSP, 2019
- Adithya Renduchintala, Shuoyang Ding, Matthew Wiesner and Shinji Watanabe: *Multi-Modal Data Augmentation for End-to-end ASR*, Interspeech 2018
- Junyi Yang, Lucas Ondel, Vimal Manohar and Hynek Hermansky. *Towards Automatic Methods to Detect Errors in Transcriptions of Speech Recordings*, IEEE ICASSP, 2019
- Hainan Xu, Shuoyang Ding and Shinji Watanabe. *Improving End-to-end Speech Recognition with Pronunciation-assisted Sub-word Modeling*, IEEE ICASSP, 2019

2.2 MT

We developed MT models for low resource conditions, using a wide range of statistical and neural MT architectures, with a special focus on how data is prepared and presented to training, and attention to how MT interfaces with speech recognition and information retrieval.

2.2.1 MT Architectures

During the project we developed MT systems using various architectures:

(1) statistical phrase-based models,
(2) statistical hierarchical models,
(3) recurrent neural network models, and
(4) Transformer neural network models.

At the start of the project, neural models under-performed statistical models but improved throughout, especially with the advent of the Transformer model. Nevertheless, our final system still draws more heavily on the statistical model due to its greater transparency and diversity, allowing us to extract richer output (n-best lists, bag of phrases) that is beneficial to the downstream information retrieval task.

MT technology has been mostly developed with an eye towards high resource scenarios (the most popular benchmark language pairs are German-English, French-English and Chinese-English), so we spent significant engineering effort on adapting them to low resource conditions. This involved experimentation to find appropriate hyper-parameters, such as the number of

subword units. We report on this experimentation in detail in our monthly reports but also in published system descriptions on related tasks.

- Shuoyang Ding, Adithya Renduchintala and Kevin Duh. *A Call for Prudent Choice of Subword Merge Operations*, MT Summit 2019
- Philipp Koehn, Kevin Duh and Brian Thompson. *The JHU MT Systems for WMT 2018*, Conference on MT (WMT) 2018
- Kelly Marchisio, Yash Kumar Lal and Philipp Koehn, *Johns Hopkins University Submission for WMT News Translation Task*, WMT 2019

2.2.2 Data Refinements

Our main focus in improving MT to low resource conditions was targeted on developing novel methods to modify existing data and augment it in creative ways. Not all the methods we developed throughout the project were integrated into our final evaluation system since they were on a longer-term research trajectory.

Supervised and Unsupervised Morphology Most languages in the world are morphologically much richer than English, creating a much larger vocabulary of surface word forms. Our ambitious Unimorph project (http://www.unimorph.org/) aims at developing morphological analyzers for hundreds of languages. We employed the tools and resources built by this project to the MATERIAL languages. In addition to such supervised methods, we also used unsupervised morphology, broadly based on Morfessor but refined to our needs for proper handling of named entities and numbers. Since we are translating from a morphologically rich language, we split it into morphemes using these analyzers in a pre-processing step.

Character-Based Embedding. The motivation for using morphological analysis also applies to character-based models. By inducing the embedding for words from their spelling, we hope to exploit the fact that similarly spelled words often have similar meaning by automatically learning the patterns that match word form with their function.

- Adithya Renduchintala, Pamela Shapiro, Kevin Duh and Philipp Koehn, *Character-Aware Decoder for Translation into Morphologically Rich Languages*, MT Summit 2019

Iterative Back-Translation. A classic data augmentation technique in MT is back-translation, i.e., the creation of synthetic parallel data by translating target side monolingual text using a MT system trained in the reverse direction. We developed iterative MT by iterating this processing, thus taking advantage of both source and target side monolingual text.

- Cong Duy Vu Hoang, Philipp Koehn, Gholamreza Haffari and Trevor Cohn: *Iterative Back-Translation for Neural MT*, Workshop on Neural MT 2018

Multi-Language Training. Another data resource that has been proven successful is related language pair data. We took note of research on training a single MT system for multiple language pairs that includes ideas such as zero-shot translation. In our case, we are interested on improving the translation quality of a targeted language pair by the augmentation with related language pair data. We showed improvements even with relatively unrelated language pairs (e.g., French-English for the MATERIAL languages) and including training data in the reverse

direction (English-Swahili for Swahili-English). For instance, when adding French-English to Swahili-English data, we observed a gain of 2.4 Bilingual Evaluation Understudy (BLEU) points (Monthly report 2018-05).

Adaptation Methods. Since our data augmentation techniques (back-translation, related language data, crawl\-ed data) give us diverse parallel data, we need to be careful to consider how to use each type of data during training. Due to the centrality of this problem, we dedicated a large amount of effort towards adaptation methods.

The most popular method for balancing a large set of less relevant data (out of domain) and a typically smaller set of more relevant data (in domain) is continued training, i.e., additional training iterations just on the in-domain data. However, this often leads to over-fitting, so we developed regularization methods to overcome such catastrophic forgetting. We also more closely explored curriculum methods which lay out at what stage to train on what data. We also explored other aspects of domain adaptation, such as adapting to individual documents, controlling aspects of style, and analysis of these adaptation methods.

- Huda Khayrallah, Brian Thompson, Kevin Duh and Philipp Koehn: *Regularized Training Objective for Continued Training for Domain Adaption in Neural MT*, Workshop on Neural MT 2018
- Brian Thompson, Huda Khayrallah, Jeremy Gwinnup, Kevin Duh, Philipp Koehn. *Overcoming Catastrophic Forgetting During Domain Adaptation of Neural MT*, North American Chapter of the Association for Computational Linguistics (NAACL) 2019
- Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, Kevin Duh. *Curriculum Learning for Domain Adaptation in Neural MT*, NAACL 2019
- Sachith Sri Ram Kothur, Rebecca Knowles and Philipp Koehn: *Document-Level Adaptation for Neural MT*, Workshop on Neural MT 2018
- Kelly Marchisio, Jialiang Guo, Cheng-I Lai and Philipp Koehn, *Controlling the Reading Level of MT Output*, MT Summit 2019
- Brian Thompson, Huda Khayrallah, Antonios Anastasopoulos, Arya D. McCarthy, Kevin Duh, Rebecca Marvin, Paul McNamee, Jeremy Gwinnup, Tim Anderson and Philipp Koehn: *Freezing Subnetworks to Analyze Domain Adaptation in Neural MT*, WMT 2018

Analysis. One core challenge of neural methods is their lack of transparency. It is hard to gain insight into their inner workings and what lead to specific predictions, making it hard to engage in error-driven optimization. Lack of model transparency remains one of the biggest problems for deep learning research in general. Our main angles to approach this problem was to develop methods that allow us to track what inputs led to specific decisions using saliency and comparison with models that lack some inputs, probing and visualization of internal states, as well as linking specific behavior (say, copying of words) to their form and context.

- Shuoyang Ding, Hainan Xu and Philipp Koehn, *Interpreting Word Alignment in Neural MT with Saliency*, WMT 2019
- Xutai Mai, Ke Li, and Philipp Koehn: *Source Context Dependency in Neural MT*,

European Association for MT (EAMT) 2018

- Rebecca Marvin and Philipp Koehn: *Exploring Word Sense Disambiguation Abilities of Neural MT Systems*, Association for MT in the Americas (AMTA) 2018
- Rebecca Knowles and Philipp Koehn. *Context and Copying in Neural MT*, Empirical Methods in Natural Language Processing (EMNLP) 2018

Integration with Speech and Retrieval. MT is part of the processing pipeline of the QuickSTIR platform. In the case of speech documents, it is a processing step after speech recognition and before information retrieval. Instead of treating these components as purely modular, we developed methods to more closely adapt MT to the upstream speech recognition component and the downstream information retrieval component.

One aspect of this tighter integration is to preserve ambiguity and diversity which is especially important in view of the need for high recall during information retrieval. We do this by processing n-best lists and producing bag of phrases (more on that in Section 2.4 on cross-lingual information retrieval).

Another aspect is the proper handling of tokenization, including truecasing. We maintain the same tokenization throughout the process. Typically, we prefer the use of truecased text (say, lowercased *the* at the beginning of a sentence, but uppercased *Smith* anywhere) but speech recognition is typically build to produce lowercased text (say, *smith*). Other aspects are missing punctuation and the spelling out of numbers (say, *three* instead of *3*).

2.2.3 Web Crawling of Parallel Data

A major part of our MT effort is the acquisition of additional parallel text through the targeted crawling of web sites that contain translated text.

The processing pipeline that we refined throughout the project and adapted to the MATERIAL languages consists of six steps: (1) identification of web sites for crawling, (2) downloading the content of the web site by following all links within the same webdomain, (3) extraction of text from Hypertext Markup Language (HTML) pages and language identification, (4) document alignment, (5) sentence alignment, and (6) sentence pair filtering. Maintaining the crawling infrastructure is a major engineering effort that requires continuous processing of data across over a dozen machines and specialized hardware such as Solid State Drives (SSD) drives for intermediate storage during crawling.

The main challenge that were presented with in MATERIAL was to adapt the existing infrastructure to the low resource scenario. The pipeline draws in various components on the existence of translation dictionaries and baseline MT systems. In the low resource scenario, these are less reliable.

A major contribution during the project was the discovery and close examination of the impact of noisy data that led to the development of novel sentence filtering methods, including the organization of two shared tasks on sentence filtering at the WMT conference (http://www.statmt.org/wmt18/parallel-corpus-filtering.html and http://www.statmt.org/wmt19/parallel-corpus-filtering.html). We also collaborated with Facebook to create a public benchmark for low resource languages which was used in one of the shared tasks. Our advances in sentence alignment based on neural sentence representations led to

a new open source tool which significantly outperforming existing methods.

- Huda Khayrallah and Philipp Koehn: *On the Impact of Various Types of Noise on Neural MT*, Workshop on Neural MT 2018 (outstanding paper award)
- Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk and Philipp Koehn, *Low-Resource Corpus Filtering using Multilingual Sentence Embeddings*, WMT 2019
- Huda Khayrallah, Hainan Xu and Philipp Koehn: *The JHU Parallel Corpus Filtering Systems for WMT 2018*, WMT 2018
- Francisco Guzm\'an, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary and Marc'Aurelio Ranzato, *Two New Evaluation Datasets for Low-Resource MT: Nepali-English and Sinhala English*, EMNLP 2019
- Brian Thompson and Philipp Koehn, *Improved Sentence Alignment in Linear Time and Space*, EMNLP 2019

2.3 CLIR

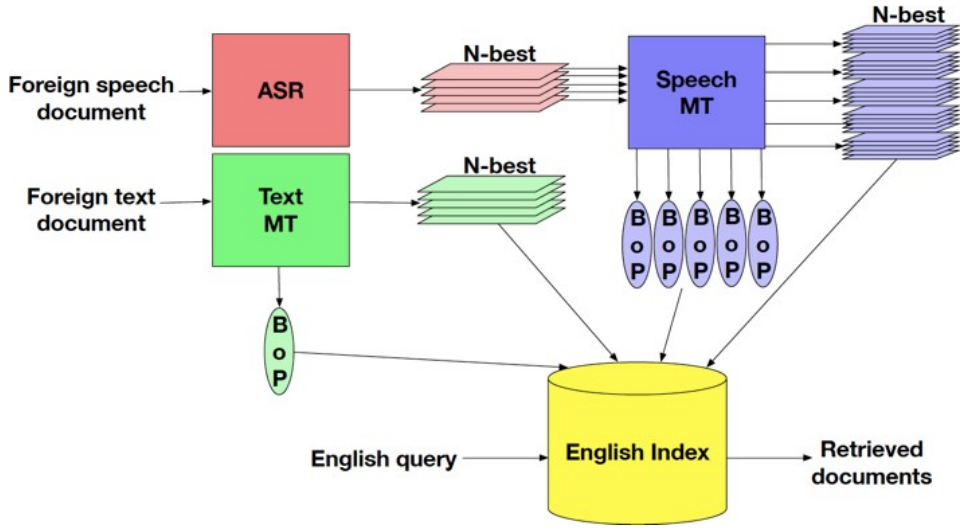

Figure 1. Multi-View Document Representation for Robust CLIR

We developed a flexible CLIR system based on the idea of multi-view document representations. Each foreign document to be indexed is transcribed with multiple speech recognition engines and translated with multiple MT engines (Figure 1).

The diversity of engines, together with different N-best list and bag-of-phrase (BoP) representation of the transcription and translation outputs, lead to more robust CLIR results. The size of the n-best list generated from ASR and then the size of the M-best list generated from MT varied are hyperparameters that we optimized, by using values of up to ten for each of them.

Our CLIR software is implemented as a thin layer on top of ElasticSearch, an open-source and scalable search engine.

As seen in Figure 2, the use of multiple views consistently improve information retrieval (IR) metrics across our three test collections.

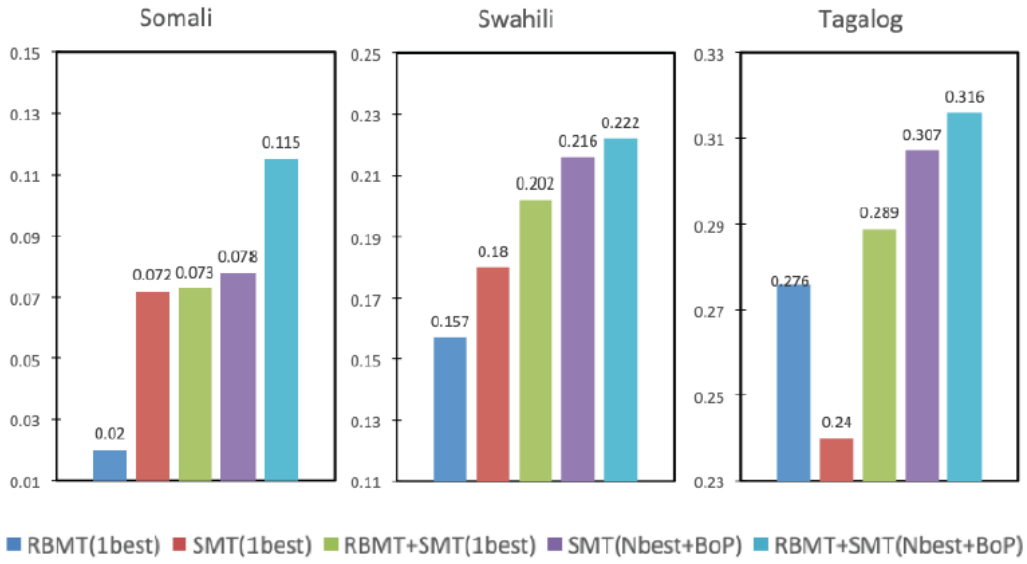

Figure 2. Maximum Query Weighted Value (MQWV) Improvement Due to Multi-View Representation on Three CLIR Test Collections.

For example in Tagalog, combining the output of rule-based MT (RBMT), statistical MT (SMT) with N-best lists, and SMT with BoP, achieves 0.316 MQWV, significantly outperforming CLIR systems that using only one of the individual document representations.

We also explored the novel research area of cross-lingual learning-to-rank. The idea is to build a supervised training set that contains English queries, foreign documents, and their relevance judgments.

Figure 3 illustrates our process for automatically extracting a silver-label training set of English queries and foreign documents from Wikipedia.

This large-scale dataset contains more than 2.8 million English queries with relevant documents from 25 other selected languages. We demonstrate that it is sufficiently large to support the research of neural network methods for CLIR.

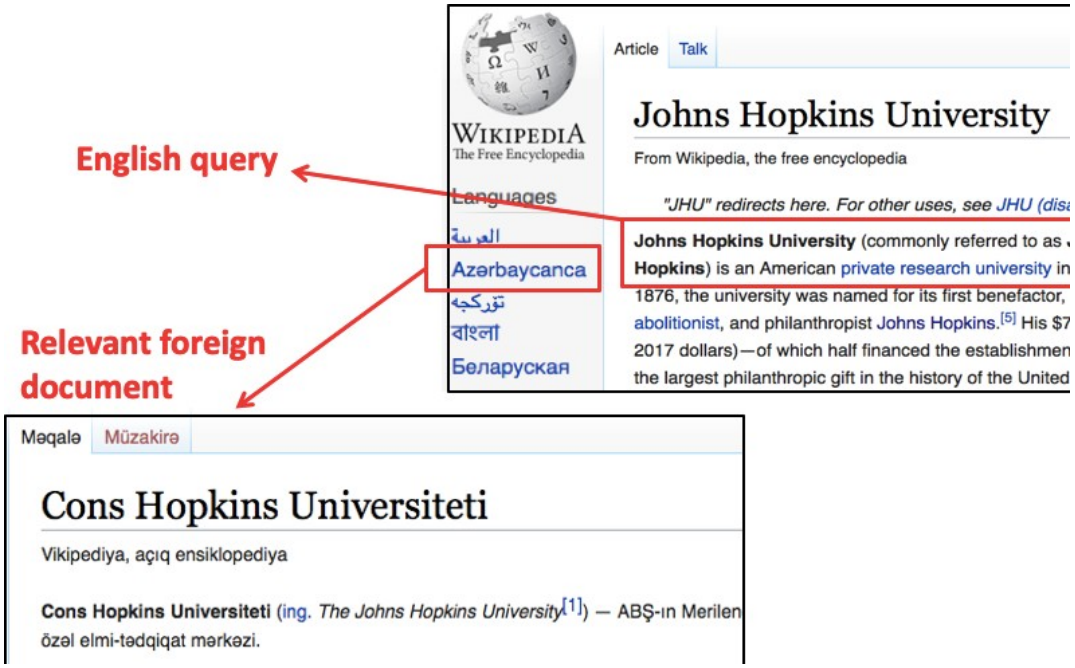

Figure 3. WikiCLIR Training Data Creation Process

We assume the first sentence of an English Wikipedia article can be converted into a meaningful query, and exploit the inter-wiki links to extract relevant foreign documents. This is silver-label data in the sense that it relies on automatic heuristics and some amount of noise is expected.

The CLIR efforts are summarized in these publications:

- Mahsa Yarmohammadi, Xutai Ma, Sorami Hisamoto, Muhammad Rahman, Yiming Wang, Hainan Xu, Daniel Povey, Philipp Koehn and Kevin Duh, *Robust Document Representations for Cross-Lingual Information Retrieval in Low-Resource Settings*, MT Summit 2019

This work describes in detail our multi-view CLIR system and presents results on MATERIAL collections.

- Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh and Kentaro Inui: *Cross-lingual Learning-to-Rank with Shared Representations*, NAACL 2018

This work describes how we constructed the WikiCLIR dataset that supports cross-lingual learning-to-rank. The dataset is publicly available at:

https://www.cs.jhu.edu/~kevinduh/a/wikiclir2018/

2.4 Domain

2.4.1 Model

We train or finetune binary classifiers for each domain in each language using synthetic data (for language 1A and 1B) plus gold annotations of other languages if allowed (for language 1S). Our best performing model is Deep Averaging Networks (DAN), a simple, fast but effective deep unordered model. In particular, we use the concatenation of max-pooling and mean-pooling of

the embeddings of all the words in a document as the representation layer and feeds it into a dense layer with a nonlinear activation function (ReLu for speed) for training. We normalize the output layer with a softmax layer and use the softmax value of the positive label as the domain score. We use pre-trained word embedding Global Vectors for Word Representation (GloVe) and a word dropout of 0.1 for regularization. Thresholds tuned on the F1 of P-Miss and P-FalseAlarm on the dev set for each classifier are used for the final evaluation.

For synthetic training data, we extract 1,000 relevant documents using Lucene from the English Wikipedia according to the domain description for each domain as positive example and 1,000 irrelevant documents as negative ones. We then sample 1,000 sentences from the documents, stratified by a preliminary domain classifier's score on a sentence as positive example for fine-tuning. Granting some assumptions on the preliminary classifiers, these data sets are relatively balanced, facilitating training of future domain classifiers. The sentences for each domain were then annotated with scalar judgments of in-domainess using three-way redundancy on Amazon Mechanical Turk using Efficient Annotation of Scalar Labels (EASL).

Our work exploring the power of non-contextual simple models for text classification is described in

- Seth Ebner, Felicity Wang and Benjamin Van Durme: *Bag-of-Words Transfer: Non-Contextual Techniques for Multi-Task Learning*, Deep Learning for Low-Resource Natural Language Processing (NLP) Workshop (DeepLo) 2019

In this paper we study three simple but effective bag-of-words techniques for text classification in the multi-task setting: pooling encoders(the DAN model, in which we use a concatenation of mean-pooling and max-pooling with non-linear rectified linear Unit (ReLU) transformation and word dropout), pre-trained word embeddings, and unigram generative regularization (we reconstruct the input sequence using a conditional unigram language model).

For the multi-task setting, we train the main task with auxiliary tasks empirically picked by previous works. These tasks share the same word embedding and encoding (linear layer(s) of the DAN model) layers and have their own Multilayer Perceptron (MLP) and softmax layers which map the encoded representations to their predictions. Weighted sum of loss of each task is used for backpropagation for all parameters and the weights are a hyper-parameter to tune.

We followed a state-of-the-art paper with BiLSTM and complex shared labeling embeddings and tested on eight datasets with pairwise input sequences on (topic/target/aspect-based) sentiment analysis and textual inference (Multilingual Natural Language Inference [MultiNLI]). Experiment results showed that simple methods can achieve comparable results, with easier implementation, greater speed and less memory.

Our code is publicly available at

https://github.com/felicitywang/tfmtl

2.4.2 Data Collection

We collected scalar judgments of the domains relevant to each sentence using the EASL interface, a crowd annotation interface for eliciting bounded, real-valued (scalar) scores on relatively simple items. The interface shows five items to the crowd worker at once, eliciting a separate score for each. For the domain likelihood judgment task, the crowd worker is given the

name and description of a domain and instructed to score each sentence according to how likely it belongs in that domain. This interface is illustrated here, Figure 4, instantiated with a domain likelihood judgment task for the ``military'' domain:

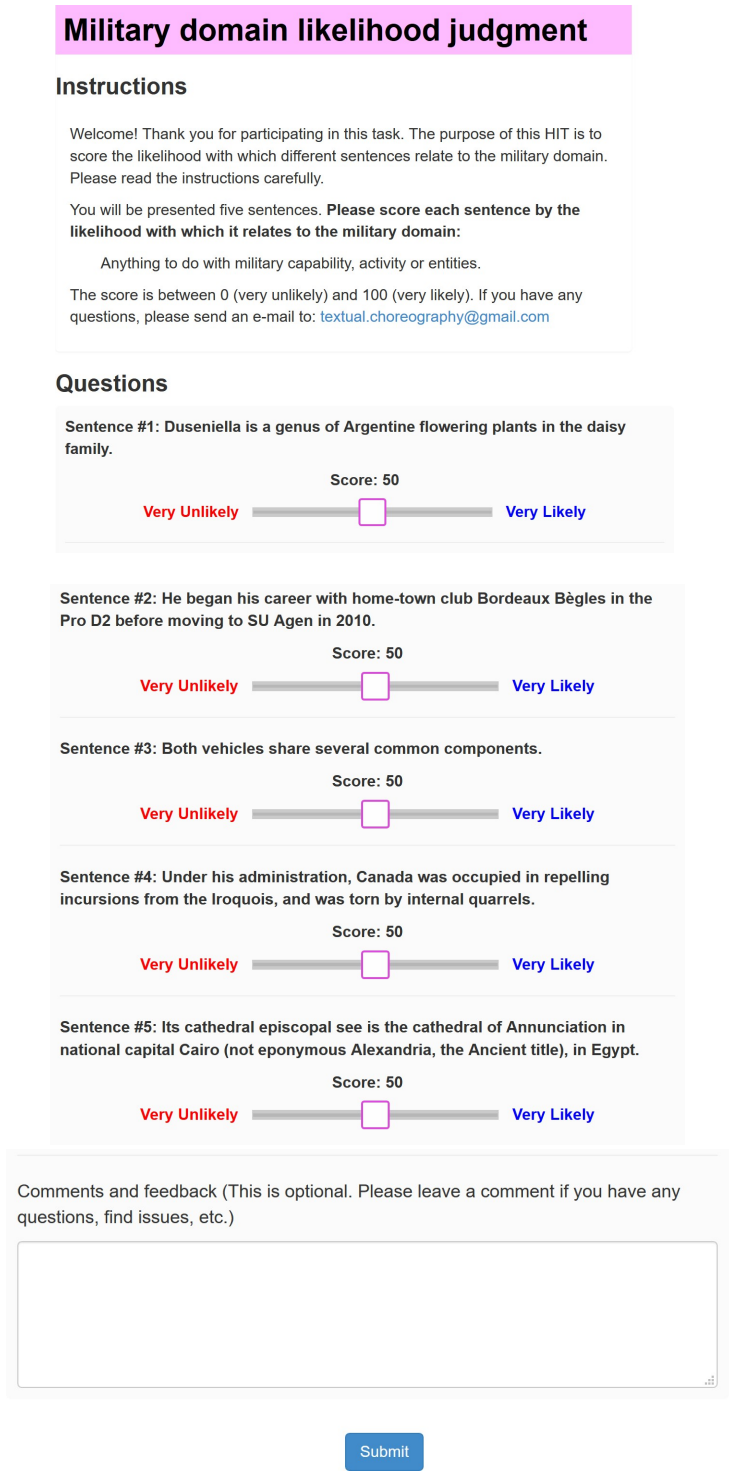

Figure 4. Military Domain Likelihood Judgement

The position of the slider for each item is encoded as a number between 0 and 100 (inclusive). To reduce variance in these elicited scores, redundant annotations are collected (and averaged) for each item.

We conducted four pilot studies to test the annotations elicited through EASL for the domain annotation task: A comparison of crowd annotations to keyword matches, a comparison of crowd annotations to expert annotations, a comparison of document-level expert annotations to sentence-level expert annotations, and finally a dry run for the MATERIAL evaluation in which sentences are sampled based on an automatic classifier's predictions and then annotated by crowd workers.

In the first study, we compared crowd annotations collected with EASL to the presence of domain-relevant keywords provided by the MATERIAL program. Specifically, we elicited annotations on sentences from Wikipedia articles matching the domain-relevant keywords as well as on sentences from random Wikipedia articles. In a preliminary analysis on the military domain, scores were clustered around 0 and 100, meaning most documents were scored as either very unlikely related to the military domain or very likely related to it. However, in a subsequent analysis on six domains (including a second round of annotations on the military domain), scores were much more evenly distributed. The dramatic shift in the shape of the military domain's score distribution between rounds points to a potential need to control variance in the annotations due to time and/or annotator selection.

In the second study, we compared crowd annotations collected with EASL to expert annotations collected through another interface. Specifically, we elicit EASL crowd annotations for the expert document-wise domain annotations for language 1B provided by the MATERIAL program; we then use the expert annotations to guide an exploration of annotation/annotator filtering heuristics. Every sentence from a sample of expert-annotated documents in each domain was annotated using EASL. Ranking each document by its highest-scoring sentence under the crowd annotations and using the correlation between that ranking and the expert annotations to measure crowd annotation quality, we investigated four crowd annotation filtering approaches: filtering out annotations that were far from the median, filtering out annotations that were completed too quickly, filtering out annotations that were completed too slowly, and filtering out annotators who picked similar scores regardless of the sentence. Overall, we found that the rank correlation between aggregate (per-document maximum) crowd annotations and expert annotations was better after filtering out annotations completed in less than 20 seconds (right) than it was before filtering (left):

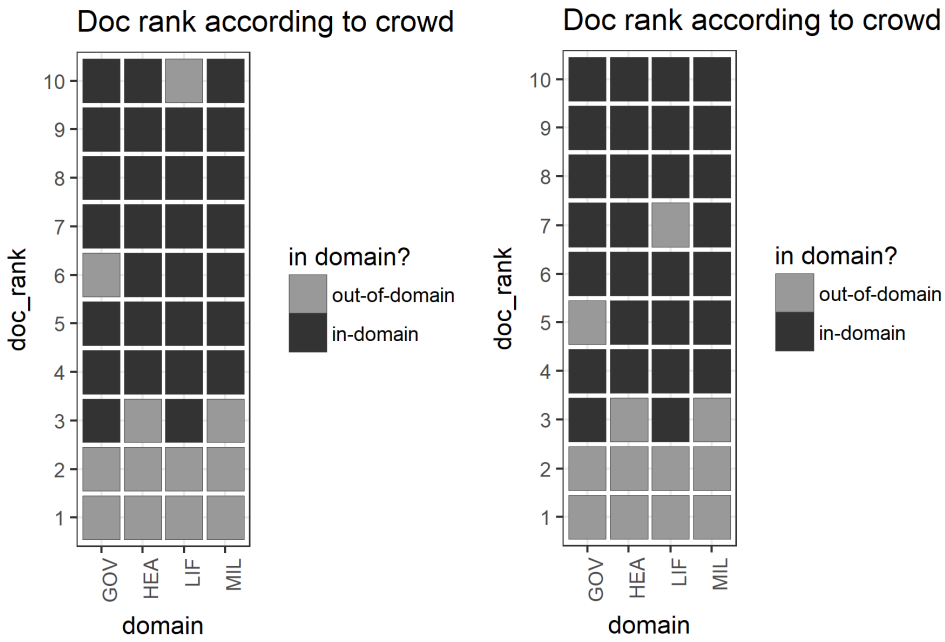

Figure 5. Doc Rank According to Crowd

In the third study, we compared a set of sentence-wise expert annotations to the document-wise expert annotations from before. There is an imperfect correspondence between these two sets of annotations; for example, three sentences annotated as in-domain for government (GOV) and five annotated as in-domain for lifestyle (LIF) do not appear in the documents annotated as in-domain for government and lifestyle (respectively). The following plots show the distributions of crowd-annotated sentence scores when they are grouped by sentence-level expert annotations (left) and document-level expert annotations (right). The notches of the boxplots overlap under the sentence-level grouping (left) but not for the document-level grouping for GOV (right, first panel), suggesting that *sentence-level* crowd annotations do not correlate well with *sentence-level* expert annotations and may counterintuitively correlate better with *document-level* expert annotations.

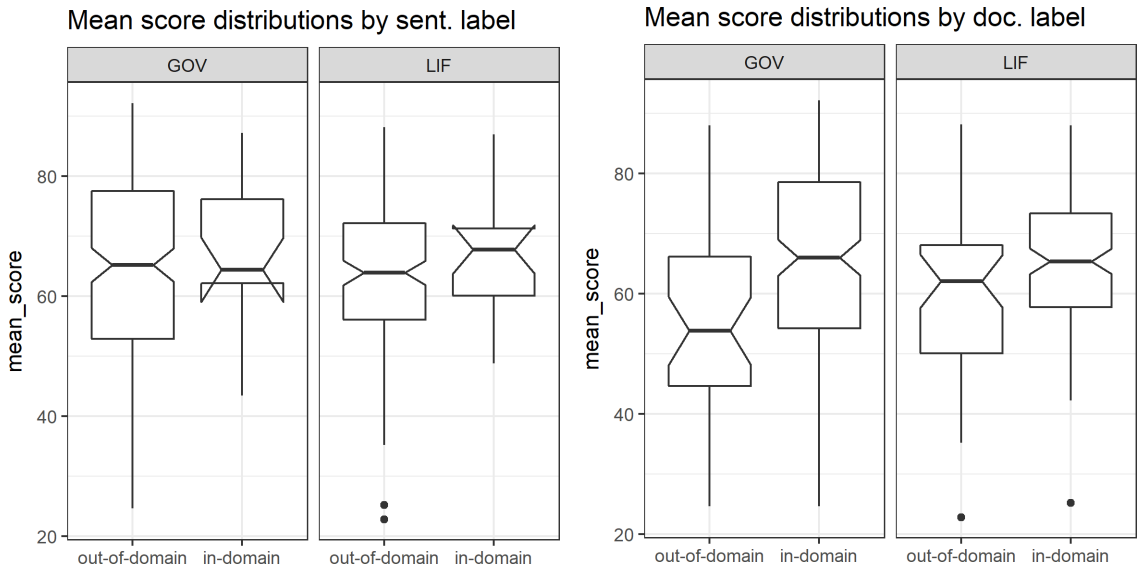

Figure 6. Mean Score Distributions by Sent. and Doc. Label

In the final study, we constructed a balanced crowd-annotated data for a MATERIAL dry run. To create a sample of sentences for each domain that spans a range of scores, we first trained a rudimentary domain classifier and used it to produce initial estimates of domain scores for a set of candidate sentences. We then sampled from those sentences, stratifying the sample uniformly by estimated score; we investigated non-uniform sampling strategies but did not find sufficient empirical support to deviate from uniform sampling. We then elicited crowd annotations of scores for that stratified sample, finding a relatively balanced distribution of scores within each domain, as desired. A multimodality is exhibited in the scores for most domains; we leave its investigation to future work:

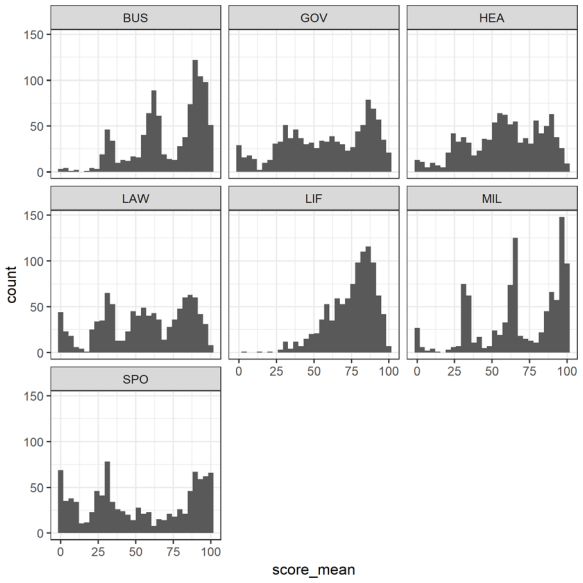

Figure 7. A Multimodality Exhibited in the Scores

2.5 Summarization

Initial summarization work entailed supporting Mechanical Turk experiments to optimize summarization system hyperparameters. The team generated sample summaries consisting of word cloud (see Figure 4) and extractive snippet images from CLIR output based on the CLIR team's initial system built on training, development, and analysis data sets.

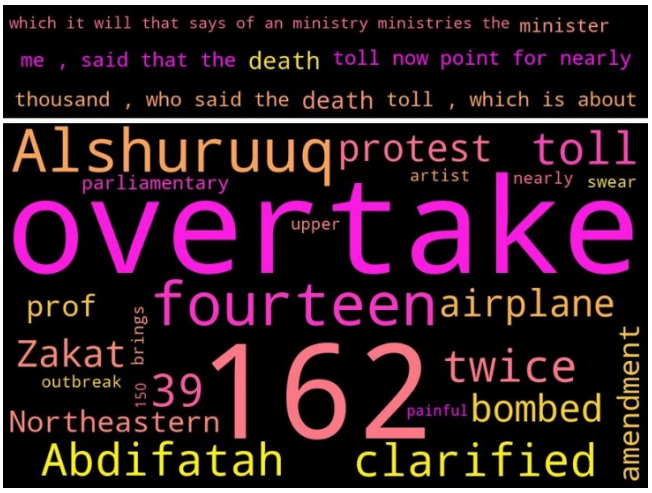

Figure 8. Example summary Image for the Query "Minster of the Interior, Death of Foreigners"

Subsequent summarization work consisted incorporating changes to CLIR inputs and upstream processes under development by the ASR (automatic speech recognition), MT, and CLIR teams. These included ASR N-best best output and MT bag-of-phrase output, which required significant modifications to the processing of CLIR output. Additionally, the summarization team handled MATERIAL end-to-end (E2E) system submission, which turned out to be a moving target given numerous changes by IARPA to the expected output format.

The summarization team also put additional software infrastructure in place to manage efficient execution and generation of summaries and E2E submission metadata for supporting the submission process (see Figure 9). This work focused on productization and hardening of the summarization pipeline. This included equipping the system to parallelize execution in a grid environment as well as hardening of the server components to allow configuration for simultaneous instances of the summarization module. Additionally, summarization focused on minor changes to our pipeline to handle additional upstream N-best output embedded in the CLIR raw results as well as output validation scripts for the evaluation period.

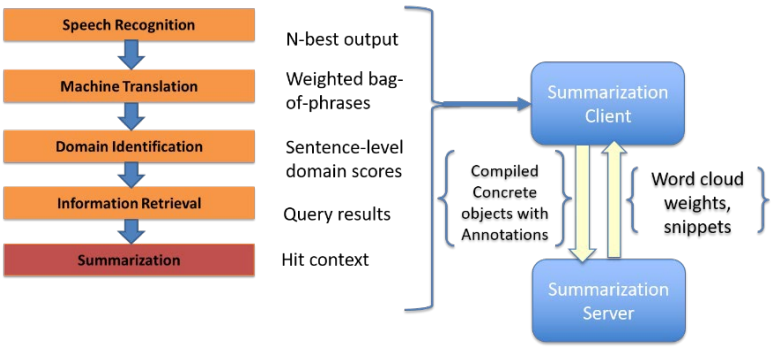

Figure 9. Summarization Pipeline System Diagram

2.6 System Integration

2.6.1 Experiment Tracking

Initial system integration work focused on adapting a PHP-based experiment tracking web application originally built for MT experiments to report progress, output, and statistics for ASR and CLIR experiments as well. This effort involved a significant amount of code porting from PHP to a modern Python Representational State Transfer (RESTful) framework. The resulting system can be integrated with the subsequent experiment execution pipeline.

2.6.2 Experiment Pipeline

To facilitate a common picture of CLIR component progress, the AST and JHU team created a unified pipeline to execute ASR, MT, and IR from a unified experiment configuration file. The outputs and performance analysis (if appropriate ground truth is available) are then exposed through the above-mentioned web dashboard enabling cross-system and cross-corpora comparisons of results (building off a previously deployed interface for MT experiments). In addition to facilitating internal experiments, by isolating code, data, and derived components of the pipeline, and by defining common output formats at each stage, the framework enables rapid prototyping and testing of arbitrary compatible components in any execution environment.

The idea behind this approach is best explained by illustrating (as shown in Figure 10) . The dependencies between the CLIR system components, separated into input data items (blue), source code controlled items (red) , and derived outputs (models – purple, data – green)

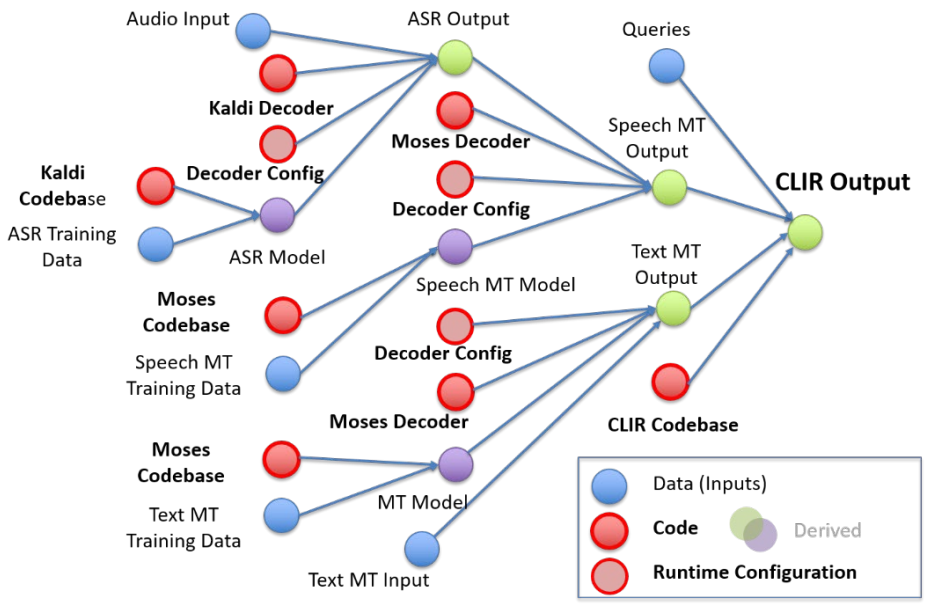

Figure 6:

Figure 10. Experiment Provenance Graph – Source Code Controlled Items Highlighted

Given a well-defined set of code versions, including configuration files with specific parameter settings for each training and decoding component, (specified as git commits for example) and a well-defined set of data inputs, a system with repeatable output and performance can be defined.

Additionally, as illustrated by Figures 11 and 12, this clear delineation of component input and output allows for easy system comparison.

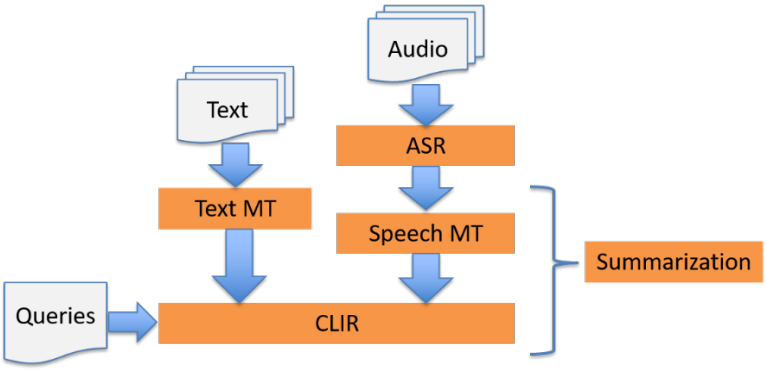

Figure 11. Simplified CLIR+S Process Flow

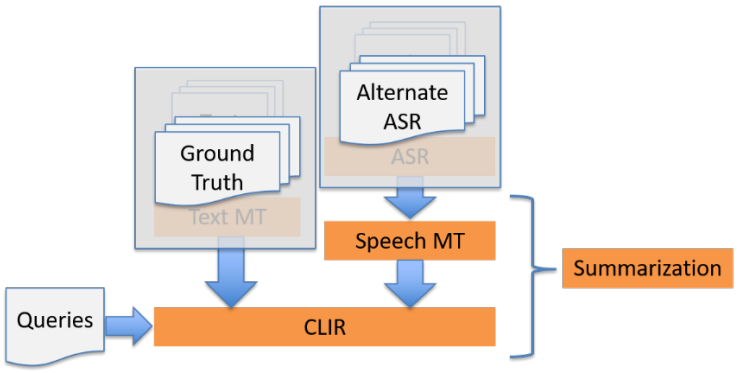

Figure 8:

Figure 12. CLIR System Flow with External Inputs for System Comparison

The current implementation assumes a batch processing model, which easily leverages already existing batch execution modules in Kaldi and Moses, however we encapsulate the decoding functionality of ASR and MT such that the pipeline could easily be extended to interface with Docker endpoints for any of the component technologies.

The "technical debt" that must be incurred to build a single pipeline is fourfold. First, the isolation of decoding components and parameters from the complex training recipes within which they are typically embedded. Second, the identification of parameters to expose to the top-level configuration. Third, maintaining the "decoder" module as new decoding techniques and/or sequences of operations are developed during the course of the research program. Fourth, the standardization of output formats so that the pipeline can be configured to run all components or a subset of components (only MT and IR, not ASR, for example) using the output of previous experiments.

By way of example, we can describe the ASR decoder component as containing models (which change based on different training data, hyper-parameter choices, and training operation sequences, Deep Neural Network (DNN) architectures, etc.), sequence of operations (e.g., BLSTM decoding with the HCLG (Hidden Markov Model [HMM], Context, Lexicon, Grammar) Finite State Transducer (FST)followed by RNNLM re-scoring) The Kaldi approach to capturing a unique decoder is purely through source code versioning. The training data, models, and decoding output are tightly coupled. This has proven quite effective at ease of replication of ASR experiment results. Given a version number and access to training data. In the course of creating our full CLIR pipeline, the decoupling of training from decoding has been absolutely essential. This, combined with standard output formats, ensure researcher B can use models built by researcher A, or simply the output of researcher A's pipeline.

This decoupling of training from decoding, and isolation of key decoding model components and configuration was also necessary to the Dockerization of ASR and MT. We have completed a Docker-image endpoint for the first and our pipeline provides a roadmap for the second.

2.7 Data Collection for Option Period 2

Our final contribution to MATERIAL is the collection of documents in the three languages targeted in Phase 3. These are documents crawled from the web, with a focus on news sites and blogs.

The goal for the data collection was to find 500-1500 word documents of coherent text. News stories or blog posts are typical examples of that, while arbitrary web pages that may contain hotel information, product descriptions, image captions other mostly non-text content are a less likely match.

News Stories. We identified the major online news sites for the languages by consulting web directories. These may be online newspapers, the online presence of television channels, or web-only media.

Blogs. We used statistics from CommonCrawl to identify web sites with content in the targeted languages. These statistics were originally created to divide up all the text in CommonCrawl into languages and release a set of corpora, split by language. For our purpose, we assembled statistics of language distributions for the documents under one webdomain, measured in bytes of text in that language. If the amount of text for one of the targeted languages exceeded a minimum (e.g., 10,000 or 1 million characters), we selected it for deep crawling. To constrain to blog content, we crawled only sites with the string *blog* in it.

Non-Blog. For one of the targeted languages (Kazakh), we did not collect sufficient news or blog data, so we also collected content from any webdomain with a minimum amount of text detected in the CommonCrawl corpus.

2.7.1 Processing Pipeline to Extract Text

We employed the robust processing pipeline that has been developed under a umbrella project called *Paracrawl* which received funding from diverse sources, i.e., Google, Bloomberg, the European Union, in addition to IARPA MATERIAL.

The processing pipeline has been developed for parallel corpus collection, but a truncated and modified version of the pipeline was used here for monolingual corpus collection. It consists of the processing steps, web crawling, language identification, and text extraction.

The result of this processing pipeline as used here is a set of documents (text-only, sentence-split), each corresponding to a web page.

2.7.2 Document Cleaning

Web pages contain boilerplate content such as navigational elements (*Home, About, News*, ...) and copyright notices. We developed a filtering program to remove this boilerplate.

The initial version collected counts for each line of text (a sentence or smaller text fragment) and discarded all sentences that occurred more than once --- with the addition of the special cases of lines that were next to such singleton lines that occurred more than once but still rarely. The motivation for this special case is that any text document may contain common sentences that occur more than once in news stories (e.g., *The bill was vetoed by the President.*)

A refined version was developed to handle the special case that on some web sites, the same document occurs under multiple Uniform Resource Locators (URL) (e.g., by having additional arguments in the URL such as the referring page). To address this problem, either a higher threshold than singleton can be manually specified, or a threshold can be automatically discovered by finding the minimum count across all lines in the document. So, for a duplicated web page, all lines occur at least two times, so any line occurring twice is preserved.

2.7.3 Document Filtering

At this point, we have a large collection of candidate documents. These may be filtered based on several criteria. One obvious filtering is document size. We did not perform such filtering since it is trivial to implement.

We carried out some filtering based on the URL. We examined the content of web pages for the news sources and checked which actually contain news stories and if this can be detected from the URL. This was often the case, e.g., for the domain civil.ge only URLs with the pattern old.civil.ge/geo/_print.php?id=[NUMBER] contain news stories. We carried a proof of concept for two webdomains each for each language and each content type (news, blog) but did not scale up this effort.

3.0 CONCLUSIONS

In conclusion, work has been accomplished in the areas of automatic speech recognition, MT, cross-lingual information retrieval, domain classification, and summarization under low-resource conditions, and an integrated system has been developed and evaluated.

In automatic speech recognition, we introduced TDNN, semi-supervised acoustic modelling, and RNNLM rescoring for language modelling. In MT, we explored several MT approaches and refined neural MT models, improved handling of textual data through morphological analysis, character-based embedding models, iterative back-translation, multi-language training, analysis and visualization of models, worked in the tighter integration with speech and information retrieval, and acquired large amounts of web-crawled data. In cross-lingual information retrieval, we developed a flexible system based on multi-view document representations and explored learning-to-rank methods. In domain classification, we used synthetic data from Wikipedia, collect human annotation data with EASL from crowd-sourced workers and experts. In summarization, we developed word cloud representations closely integrated with upstream tasks. We developed an experiment tracking system and deployment pipeline. We also assisted in the collection of language data in the targeted languages of the final phase of the project.

4.0 REFERENCES

- Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi and Sanjeev Khudanpur: *Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks*, Interspeech 2018
- H. Xu, Ke Li, Y. Wang, J. Wang, S. Kang, X. Chen, D. Povey and S. Khudanpur: *Neural network language modeling with letter-based features and importance sampling*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018
- H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey and S. Khudanpur: *A pruned RNNLM lattice-rescoring algorithm for automatic speech recognition*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018
- David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey and Sanjeev Khudanpur. *Speaker Recognition for Multi-Speaker Conversations Using X-Vectors*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019
- Adithya Renduchintala, Shuoyang Ding, Matthew Wiesner and Shinji Watanabe: *Multi-Modal Data Augmentation for End-to-end ASR*, Interspeech 2018
- Junyi Yang, Lucas Ondel, Vimal Manohar and Hynek Hermansky. *Towards Automatic Methods to Detect Errors in Transcriptions of Speech Recordings*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019
- Hainan Xu, Shuoyang Ding and Shinji Watanabe. *Improving End-to-end Speech Recognition with Pronunciation-assisted Sub-word Modeling*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019
- Shuoyang Ding, Adithya Renduchintala and Kevin Duh. *A Call for Prudent Choice of Subword Merge Operations*, MT Summit 2019
- Philipp Koehn, Kevin Duh and Brian Thompson. *The JHU MT Systems for WMT 2018*, Conference on MT (WMT) 2018
- Kelly Marchisio, Yash Kumar Lal and Philipp Koehn, *Johns Hopkins University Submission for WMT News Translation Task*, Conference on MT (WMT) 2019
- Adithya Renduchintala, Pamela Shapiro, Kevin Duh and Philipp Koehn, *Character-Aware Decoder for Translation into Morphologically Rich Languages*, MT Summit 2019
- Cong Duy Vu Hoang, Philipp Koehn, Gholamreza Haffari and Trevor Cohn: *Iterative Back-Translation for Neural MT*, Workshop on Neural MT 2018
- Huda Khayrallah, Brian Thompson, Kevin Duh and Philipp Koehn: *Regularized Training Objective for Continued Training for Domain Adaption in Neural MT*, Workshop on Neural MT 2018

- Brian Thompson, Huda Khayrallah, Jeremy Gwinnup, Kevin Duh, Philipp Koehn. *Overcoming Catastrophic Forgetting During Domain Adaptation of Neural MT*, NAACL 2019
- Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, Kevin Duh. *Curriculum Learning for Domain Adaptation in Neural MT*, NAACL 2019
- Sachith Sri Ram Kothur, Rebecca Knowles and Philipp Koehn: *Document-Level Adaptation for Neural MT*, Workshop on Neural MT 2018
- Kelly Marchisio, Jialiang Guo, Cheng-I Lai and Philipp Koehn, *Controlling the Reading Level of MT Output*, MT Summit 2019
- Brian Thompson, Huda Khayrallah, Antonios Anastasopoulos, Arya D. McCarthy, Kevin Duh, Rebecca Marvin, Paul McNamee, Jeremy Gwinnup, Tim Anderson and Philipp Koehn: *Freezing Subnetworks to Analyze Domain Adaptation in Neural MT*, WMT 2018
- Shuoyang Ding, Hainan Xu and Philipp Koehn, *Interpreting Word Alignment in Neural MT with Saliency*, Conference on MT (WMT) 2019
- Xutai Mai, Ke Li, and Philipp Koehn: *Source Context Dependency in Neural MT*, EAMT 2018
- Rebecca Marvin and Philipp Koehn: *Exploring Word Sense Disambiguation Abilities of Neural MT Systems*, AMTA 2018
- Rebecca Knowles and Philipp Koehn. *Context and Copying in Neural MT*, EMNLP 2018
- Huda Khayrallah and Philipp Koehn: *On the Impact of Various Types of Noise on Neural MT*, Workshop on Neural MT 2018 (outstanding paper award)
- Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk and Philipp Koehn, *Low-Resource Corpus Filtering using Multilingual Sentence Embeddings*, Conference on MT (WMT) 2019
- Huda Khayrallah, Hainan Xu and Philipp Koehn: *The JHU Parallel Corpus Filtering Systems for WMT 2018*, WMT 2018
- Francisco Guzman, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary and Marc'Aurelio Ranzato, *Two New Evaluation Datasets for Low-Resource MT: Nepali-English and Sinhala English*, EMNLP 2019
- Brian Thompson and Philipp Koehn, *Improved Sentence Alignment in Linear Time and Space*, EMNLP 2019
- Mahsa Yarmohammadi, Xutai Ma, Sorami Hisamoto, Muhammad Rahman, Yiming Wang, Hainan Xu, Daniel Povey, Philipp Koehn and Kevin Duh, *Robust Document Representations for Cross-Lingual Information Retrieval in Low-Resource Settings*, MT Summit 2019

- Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh and Kentaro Inui: *Cross-lingual Learning-to-Rank with Shared Representations*, NAACL 2018
- Seth Ebner, Felicity Wang and Benjamin Van Durme: *Bag-of-Words Transfer: Non-Contextual Techniques for Multi-Task Learning*, Deep Learning for Low-Resource NLP Workshop (DeepLo) 2019

5.0 LIST OF ABBREVIATIONS, ACRONYMS AND SYMBOLS

AMTA	Association for Machine Translation in the Americas
ASR	Automatic Speech Recognition
BLEU	Bilingual Evaluation Understudy
BLSTM	Bidirectional Long Short-Term Memory
BoP	Bag-of Phrase
CLIR	Cross-Lingual Information Retrieval
DAN	Deep Averaging Networks
DNN	Deep Neural Network
E2E	End-to-End
EAMT	European Association for Machine Translation
EASL	Efficient Annotation of Scalar Labels
EMNLP	Empirical Methods in Natural Language Processing
FST	Finite State Transducer
GLoVE	Global Vectors for Word Representation
GOV	Government
HCLG	Fully Expanded Encoding Graph
HMM	Hidden Markov Model
HTML	Hypertext Markup Language
IARPA	Intelligence Advanced Research Projects Activity
ICASSP	International Conference on Acoustics, Speech and Signal Processing
IEEE	Institute of Electrical and Electronics Engineers
IR	Information Retrieval
JHU	Johns Hopkins University
LIF	Lifestyle
LSTM	Long Short-Term Memory Network
MATERIAL	Machine Translation for English Retrieval of Information in any Language
MLP	Multilayer Perceptron
MMI	Maximum Mutual Information
MQWV	Maximum Query Weighted Value
MT	Machine Translation

MultiNLI	Multilingual Natural Language Inference
NAACL	North American Chapter of the Association for Computational Linguistics
NLP	Natural Language Processing
QuickSTIR	Quickly Adapted Speech and Translation Enabled Information Retrieval
RBMT	Rule-Based Machine Translation
ReLU	Rectified Linear Unit
RESTful	Python Representational State Transfer
RNNLM	Recurrent Neural Net Language Model
SMT	Statistical Machine Translation
SSD	Solid State Drives
SVD	Singular Value Decomposition
TDNN-F	Time Delay Neural Networks Factored
URL	Uniform Resource Locator
WER	Word Error Rate
WMF	Conference on Machine Translation