Proceedings of the Cross-Language Search and Summarization of Text and Speech Workshop, pages 58–67
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

An Investigative Study of Multi-Modal Cross-Lingual Retrieval

Piyush Arora[*†], **Dimitar Shterionov, Yasufumi Moriya, Abhishek Kaushik, Daria Dzendzik, Gareth J. F. Jones**
[*]American Express AI Labs, Bangalore, India
ADAPT Centre Dublin City University, Dublin, Ireland
{firstname.lastname1}@aexp.com[*], {firstname.lastname}@adaptcentre.ie

Abstract

We describe work from our investigations of the novel area of multi-modal cross-lingual retrieval (MMCLIR) under low-resource conditions. We study the challenges associated with MMCLIR relating to: (i) data conversion between different modalities, for example speech and text, (ii) overcoming the language barrier between source and target languages; (iii) effectively scoring and ranking documents to suit the retrieval task; and (iv) handling low resource constraints that prohibit development of heavily tuned machine translation (MT) and automatic speech recognition (ASR) systems. We focus on the use case of retrieving text and speech documents in Swahili, using English queries, which was the main focus of the OpenCLIR shared task. Our work is developed within the scope of this task. In this paper we devote special attention to the automatic translation (AT) component which is crucial for the overall quality of the MMCLIR system. We exploit a combination of dictionaries and phrase-based statistical machine translation (SMT) systems to tackle effectively the subtask of query translation. We address each MMCLIR challenge individually, and develop separate components for automatic translation (AT), speech processing (SP) and information retrieval (IR). We find that results with respect to cross-lingual text retrieval are quite good relative to the task of cross-lingual speech retrieval. Overall we find that the task of MMCLIR and specifically cross-lingual speech retrieval is quite complex. Further we highlight open issues related to handling cross-lingual audio and text retrieval for low resource languages that need to be addressed in future research.

Keywords: Multimodal Retrieval, Cross Language Text Retrieval, Cross Language Speech Retrieval, Low resource language

1 Introduction

Cross-lingual information retrieval (CLIR) is an extension of the information retrieval (IR) task where query and documents are in different languages (Oard and Dorr, 1996). The goal of CLIR is to retrieve documents matching a user's query to satisfy their information need. In general, a user would pose a query in their own language (L1), retrieve a document in a foreign language (L2) that is translated into the user's language L1. Machine translation (MT) of some form is thus one of the fundamental components in enabling CLIR (Oard and Dorr, 1996). CLIR has been the focus of much research since its definition in the 1990s. Since this time significant progress have been made in CLIR, and in the associated research areas of automatic speech recognition (ASR) and machine translation (MT). However not much work has been done in the area of multi-modal cross-lingual retrieval (MMCLIR), apart from notable examples such as (Yarmohammadi et al., 2019; Zbib et al., 2019; Boschee et al., 2019), which bring these topics together.

With the increasing interest in information access for diverse multimodal content, there is a need to learn and provide better retrieval tools and technologies to support users, in their desire to satisfy their information needs and quest for new knowledge. The expanding volume and diversity of data made electronically available every day pushes the limits of IR research and development further to facilitate retrieval over different modalities, i.e., multi-modal IR (Chang et al., 2019). This work is a step in this direction to investigate and study the challenges and the performance of MMCLIR while combining individual component solutions of MT+IR+SP for MMCLIR, and in particular the situation where limited training resources for the technologies are available.

The MMCLIR task rests around four main pillars which need to be addressed adequately both independently and in combination:

1. **Cross-lingualism**: input queries and documents to be retrieved are in different languages;
2. **Document and query modalities**: documents to be retrieved can be in different modalities than the query, but also differ among themselves e.g. text documents and audio recordings;
3. **Information retrieval**: in which the IR mechanism depends on document indexing, query processing, ranking and retrieval,
4. **Low resource constraints**: how to build effective models without having access to the resources typical for MT and speech systems is a major challenge for MMCLIR tasks.

OpenCLIR challenge campaign: This benchmark challenge[1] focused on cross-lingual text and speech retrieval, under a low-resource data setting. In this challenge there was insufficient parallel data available to train state-of-the-art MT and ASR systems. In this task, queries are written (text) keywords in English and the documents are text or audio in Swahili. The work presented in this paper was conducted within the scope of this challenge. We outline the data provided by the OpenCLIR task organizers later (see Section 4) and report our results and findings from the

[†]This work was done when the author was a Postdoctoral Researcher at the ADAPT Centre, Dublin City University

[1]https://www.nist.gov/itl/iad/mig/openclir-evaluation

OpenCLIR evaluation.

We investigate a general mechanism for MMCLIR, which can be applied for other such similar low resource languages for which there is insufficient data to train effective MT, ASR and IR systems. Having this use case in mind we present our analysis of the challenges and alternative solutions guided by the aforementioned four pillars.

The main contributions of our work are as follows:

1. We explore not only the strengths and weaknesses of various paradigms for automatic translation: dictionaries, phrase-based statistical machine translation (PBSMT), but we also combine these into a hybrid system for query translation that optimises the performance of the MMCLIR pipeline.
2. We investigate how different components (MT, ASR and IR) perform in a MMCLIR pipeline, and whether decent scores and performance are obtainable while combining SOTA components for addressing the MMCLIR problem.
3. We assess challenges and provide solutions related to the different pillars of MMCLIR that could serve as baselines for future research on this task.

In this study we pose the following research questions:

1. **RQ1**: Can we exploit alternative automatic translation (AT) approaches for effective query translation in the context of low-resource limitations?
2. **RQ2**: Can we use the most effective state-of-the-art MT, ASR and IR models under the conditions set by our use case to develop a reasonable model for MMCLIR?

This paper is organised as follows. In Section 2 we discuss related work. Section 3 presents our pipeline framework. Our use case and data are discussed in Section 4. We address the main pillars in Section 5, Section 6 and Section 7. Our results and analysis are presented in Section 8. In Section 9 we conclude and present future research directions.

2 Related Work

The strategy for crossing the language barrier between queries and documents in CLIR can be either query translation, document translation or both. Document translation is the preferred method when users need to both search and access documents in their own language (L1) (Croft et al., 1991; Buckley et al., 1997). In query translation, the query is translated into the target language (L2), and then used to retrieve indexed documents in the original language L2 (Oard et al., 2008; Narasimha Raju et al., 2014).

Multiple approaches have been explored to address query translation (sub)task over the years. These can be divided into several categories: dictionary-based, MT-based, corpus-based and ontology-based (Monti et al., 2013). Dictionary-based methods were predominant in early work on query translation (Hull and Grefenstette, 1996; Pirkola et al., 2001; Levow et al., 2005). However, out-of-vocabulary (OOV) issues may easily arise as these dictionaries are limited and require exact matches, thus the whole IR performance may be negatively impacted.

In corpus-based approaches translations of keywords in L1 are extracted from parallel or comparable corpora in L2 based on statistical methods (Picchi and Peters, 1998; Littman et al., 1998). Improvements in MT systems mean that most recent work on CLIR has focused on the use of MT for query translation (Leuski et al., 2003; Madankar et al., 2016). State-of-the-art MT now uses neural approaches (NMT) (Bahdanau et al., 2015; Vaswani et al., 2017). However, it is challenging to train effective NMT systems using limited amounts of parallel data, thus our work focuses on statistical MT approaches.

Recent work on MMCLIR has focused on document translation (Yarmohammadi et al., 2019) using SMT and NMT, learning a shared embeddings space for both queries and documents (Boschee et al., 2019), and learning better word translation probabilities using neural network (Zbib et al., 2019). The authors found that retrieval results using SMT for document translation are relatively better than NMT, possibly due to the limited nature of data (Yarmohammadi et al., 2019). A neural network based approach has also been explored to estimate word translation probabilities for CLIR (Zbib et al., 2019). The authors found that the neural network model estimates better probabilities for word translations than automatic word alignments alone, since using neural network they can encode the character sequences of input source words to generate translations of out-of-vocabulary words.

Following on from our overview, the approach adopted in this paper aims to address our task using a combination of dictionary and statistical MT due to the very limited amounts of bilingual training data available and initial findings that retrieval performance is relatively better when using SMT rather than NMT for CLIR (Yarmohammadi et al., 2019).

3 Approach

For our experiments, we used an MT system to translate the input queries from English to Swahili (described later in Section 5). As the resources available for building translation models were very limited, we focused on translation of input queries rather than attempting to translate the target documents.

We divided our investigation of the MMCLIR task into two phases:

1. **text-based retrieval**: performing retrieval on the Swahili text documents as monolingual retrieval using queries translated from English to Swahili (details on retrieval approach described later in Section 7).
2. **speech-based retrieval**: performing retrieval on the Swahili speech documents using translated queries. In this approach we explored three different alternative approaches for speech-based retrieval: i) generating ASR transcripts, ii) keyword search and iii) phoneme search (all three approaches to speech-based retrieval are described later in Section 6).

For performing document retrieval, we explored data fusion and combination techniques for ranking documents, details are provided later in Section 7.

The system architectures for our text- and speech-based retrieval methods are shown in Figure 1 and Figure 2, respectively. The main components of our system are: i) an MT system (described later in Section 5), ii) an IR system (described later in Section 7), iii) speech processing systems (described later in Section 6)

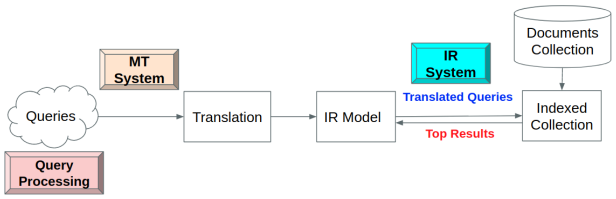

Figure 1: System architecture for text-based retrieval

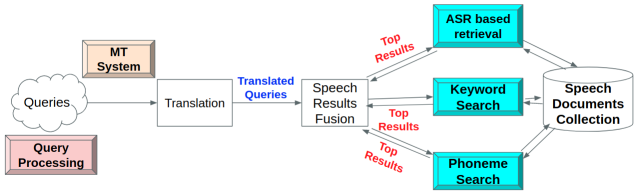

Figure 2: System architecture for speech-based retrieval

Next, we describe the dataset provided for the CLIR task.

4 Dataset & Evaluation mechanism

In this section we describe the dataset used to train MT and ASR systems, and to tune the retrieval system to address the task of text and speech retrieval. We also describe the evaluation mechanism used for this task.

4.1. BUILD corpus

We were provided a BUILD corpus to train the MT and ASR systems by NIST as a part of the OpenCLIR benchmark evaluation. The corpus is described below:

Machine Translation: The data provided by the organisers comprised of $800,000$ words of bitext for MT training. It contained $24,900$ Swahili sentences and their corresponding English translations. In Section 5.1. we provide more details about the external data we used for our translation system.

Speech corpus: We were provided with 50 hours of audio for training the ASR system, the OpenCLIR organizers recommended using 40 and 10 hours of the audio data for training and development purposed respectively.

4.2. Information Retrieval data set

The gold dataset for developing and tuning the IR model was of three types: i) analysis, ii) development, and iii) evaluation. The composition of documents (both speech and text) within all three phases varied as per shown in Table 1.

Phase	No. of text docs	No. of speech docs
Analysis	112	88
Development	101	76
Evaluation	5269	1217

Table 1: Number of documents (docs) for text and speech categories across three phases.

We had a common set of 350 queries for the analysis and development phase and another set of 350 queries for the evaluation phase. We were provided with relevance judgements for the $ANALYSIS$ corpus to gauge the performance of our models and perform error analysis to develop effective models for retrieving text and speech documents. We were also provided with human translated and transcribed data for the text and speech documents for the analysis set. We had to submit our results to the OpenCLIR evaluation portal to get system scores on the development and evaluation sets. More details on alternative approaches explored in this work and the corresponding performance is provided in Section 8.

4.3. Task evaluation

The evaluation mechanism adopted in this task sought to capture the effectiveness of the system by retrieving more relevant documents (thus minimizing false negatives, i.e. documents which are relevant, but marked as non-relevant by the system) and minimizing the errors made by the system (reducing false positives, i.e. documents which are non-relevant but marked as relevant by the system). A combined measure, shown in Equation 1, is used as an official evaluation measure[2], qv represents score for a query, for a system the qv scores for different queries are averaged and reported as $aqwv$ scores.

$$QV(Q,\theta) = 1 - [P_{Miss}(Q,\theta) + \beta * P_{FA}(Q,\theta)] \quad (1)$$

where θ is an IR threshold to tune the system to maximize QV scores, P_Miss is the false negative probability, P_FA is the false positive probability and β is a penalty factor which was set to 20 for this task.
While seeking the best models and exploring different combinations of MT, Speech and IR components on the $Analysis$ set, we measured the system performance by calculating the number of relevant documents retrieved (Recall), the number of retrieved relevant documents (Precision) and the number of queries for which relevant documents are retrieved.

5 Automatic Translation

Automatic Translation (AT)[3] between English and Swahili is a challenging task due to the lack of parallel data available for training high-quality systems. Furthermore, the specifics of the CLIR task, e.g. queries can be single words or phrases with specific constraints, impose additional constraints on how to approach the AT task.

5.1. Training Data for MT

First, we acquired the parallel data for the OpenCLIR 2019 shared task, i.e. the BUILD corpus. We also acquired and experimented with extra resources aiming to give better translation coverage. We first extended the BUILD data with the Tanzil dataset (`http://tanzil.`

[2] `https://www.nist.gov/sites/default/files/documents/2019/06/12/openclir19_evalplan_v1.21.pdf`

[3] Terms automatic translation and machine translation are used interchangeably, however AT captures dictionary as well as MT systems trained using parallel corpus in this work.

	Lang.	Tokens	ASL.	< 5	< 10
BUILD	EN	26,788	33	79	979
(22,900)	SW	44,672	30	182	1,469
+ Tanzil	EN	39,648	22	2,232	27,852
(159,853)	SW	54,715	17	5,615	45,784
+ Tanzil	EN	61,162	23	6,702	83,562
+ sing./plur.	SW	56,557	17	16,857	137,382
(483,459)					

Table 2: Statistics of the parallel data used for MT. The unique word count is after preprocessing and without the 2,000 sentences taken aside as the dev and test sets. The total number of parallel sentences is indicated between parenthesis in the first column. ASL is the average sentence length. The number of sentences with length smaller than 5 and 10 tokens is given in the last two columns.

net/) (Tiedemann, 2012) which resulted in a total of 163,153 parallel sentences. Due to the small amount of data, we opted for phrase-based statistical MT (PB-SMT) (Zens et al., 2002; Koehn et al., 2003) which handles translations at a phrase level, and typically requires much less data than building an NMT system. Details on our MT systems are presented in Section 5.2.

After conducting initial CLIR experiments with the aforementioned system we noticed that many plural/singular words are not translated, while their singular/plural forms are. Since this will impede overall IR performance more than if a translation is not correct with respect to only the number, we decided to implement a mechanism to deal with nouns in both forms (plurals or singulars). We extended the training data (BUILD + Tanzil) with additional singular and plural versions, where in each sentence all nouns had been converted to their singular and plural forms leading to a triple increase of the translation data. In order to balance the data and not put extra emphasis on part of the data (the part that contains nouns), we made three copies of all sentence pairs, even if they do not contain nouns.

From the original (BUILD) data we randomly extracted 1,000 sentence pairs as a development set and another 1,000 as a test set, (leaving 22,900 sentences in the training set). All data were tokenised and lowercased.

5.2. Core algorithmic approach

Once we had analysed the available data we decided to handle the problem of translation between English and Swahili through word- and phrase-based approaches, i.e. a dictionary and PB-SMT systems.

Dictionary: We used the resources provided by http://swahili.vickio.net/dictionary/, containing 25,000 words collected via the Kamusi Project (https://kamusi.org/). We used dictionaries of Swahili words with English translation which were obtained from 1000 Most Common Words platform (Swahili[4]), 101languages (SWAHILI 101[5]), and The Swahili-English Dictionary[6] which is based on Swahili-Kiswahili to English Translation Program by Morris Fried. [7]

We combined these dictionaries and formed a single unified dictionary providing a list of possible Swahili words for the corresponding English query words. Few examples from this combined dictionary where an English word is mapped to multiple possible Swahili words are shown in Table 3.

English word	Swahili words mapped in the dictionary
road	barabara, ndia, njia
congress	bunge, kongamano
refugees	mhamiaji, mkimbizi, mtoro

Table 3: Examples from the combined English-Swahili dictionary.

PB-SMT: Using the data described in Section 5.1. we trained three PB-SMT systems: one for each data set listed in Table 2. Our PB-SMT systems were trained using the MOSES toolkit (Koehn et al., 2007) with default settings and a 5-gram language model. Each system was further tuned with MERT (Och and Ney, 2003) until convergence or for a maximum of 25 iterations. To assess the performance of our MT systems, we used the BLEU evaluation metric (Papineni et al., 2002). Our BLEU scores on the test set are presented in Table 4.

System name	MT type	Training data	BLEU ↑
PB-SMT-B	PB-SMT	BUILD	44.40
PB-SMT-BT	PB-SMT	+Tanzil	41.73
PB-SMT-BTP	PB-SMT	+Tanzil +sing-plur	41.76

Table 4: BLEU scores for our EN→SW PB-SMT systems (higher blue value is the better).

A note on BLEU: The BLEU scores shown in Table 4 indicate that the PB-SMT-B system trained only on the BUILD corpus, performs better than the other PB-SMT systems trained with more data. The main reason is the domain-specific test set that we used – this test set is very similar to the BUILD data – which leads to the higher BLEU scores for systems trained on less data. Furthermore, due to the similarity between the data in the BUILD corpus and the documents to be retrieved, we expect that using MT systems with higher BLEU scores will lead to higher IR performance. However, we are more interested in the overall impact that these systems can have on performance when they are used in combination.

That is, we assess the quality of the alternative translation systems by measuring retrieval performance on the Analysis set (IR results using different translation systems are described later in Section 8). Our retrieval pipeline uses all three MT systems as shown in Table 4 for query translation (described later in Section 7).

[4] http://1000mostcommonwords.com/1000-most-common-swahili-words/

[5] https://www.101languages.net/swahili/swahili-word-list/

[6] https://www.mimuw.edu.pl/~jsbien/BW/Swa-Eng-xFried/Swa-eng.txt

[7] www.dict.org/links.html

The original query text is passed to each of the MT engines and a translation is generated. For broader coverage of the possible translation we consider the top 3 hypotheses returned by a MT system, under the hypothesis that this can improve IR effectiveness.

6 Speech Processing

Searching for a textual query in speech documents is often performed on speech transcripts of the documents created using an ASR system. In low-resource scenarios, it is difficult to build a high quality ASR system for the target language due to the shortage of labelled speech corpora. To alleviate the quality of our ASR system, we combined output of a keyword spotting system and a phoneme search system with ASR output.

6.1. Core algorithmic approach

The following three subsections overview conventional ASR, a keyword spotting system and a phoneme search algorithm.

6.1.1. ASR

The goal of ASR is to transcribe an audio file into speech transcripts. A conventional ASR system consists of an acoustic model, a language model and a pronunciation lexicon. While the acoustic model and language model are often developed with a machine learning approach, the pronunciation lexicon is a list of hand-crafted mapping between words and pronunciations. An acoustic model can be trained on transcribed speech data, and typically consists of a deep neural network (DNN) incorporated into an hidden Markov model (HMM) to compute posterior probabilities of phones (Hinton et al., 2012). The language model is trained on raw text of the target language, and it enforces grammatical constraints on output of an ASR system. For the CLIR task, approximately 50 hours of transcribed speech and corresponding text and a Swahili pronunciation lexicon were provided for the OpenCLIR task (see Section 6.2.). When an ASR system decodes input audio into word strings, it often employs a finite state transducer to represent phone posterior probabilities and word probabilities from which n pre-defined paths of output strings can be recovered (Mohri et al., 2002).

6.1.2. Keyword search

It is difficult to train a high quality acoustic model and a language model when only small amounts of audio and text data are available. An alternative to ASR for searching spoken documents is a keyword search system. A keyword search system takes as input a query word and a speech document and decides whether the query word is uttered in the document. One of the approaches to keyword search is transforming finite state transducers to a single generalized factor transducer, where each word token of the transducers is stored with its associated scores (Trmal et al., 2017). Given a factor transducer of a speech document and a query word, the keyword search system returns a binary decision whether the query word is in the document.

6.1.3. Phoneme search

A phoneme search is based only on a sequence of intermediate phoneme representations and a pronunciation lexicon. Given a pronunciation of a query word found in a pronunciation lexicon and a sequence of phonemes corresponding to a speech document, the system searches for an exact match of the phoneme sequence of the query word with a phoneme level transcription of the document. While this approach is likely to induce more false alarms particularly for short query words, by combining this system with ASR and keyword search, it can enrich the content of the search index.

6.2. Resources and Tools used

We used the Kaldi speech recognition toolkit to build an ASR system (Povey et al., 2011). The acoustic model consists of 6 linear layers with size 1,024 and one output layer to 1,552 context-dependent phones. The input is standard 13 dimensional MFCC speech vectors. The model has a time-delayed architecture (Peddinti et al., 2015). A language model was built using the SRI LM toolkit (Stolcke, 2002). The language model is a 3-gram built using Kneser-Ney interpolation (Chen and Goodman, 1995). For generation of pronunciations of out-of-vocabulary (OOV) words, a squiter G2P model was trained using the provided Swahili pronunciation lexicon.

A Keyword search system was also built using the Kaldi toolkit (Trmal et al., 2017). The toolkit converts decoding lattices generated using our ASR system to a generalized factor transducer of word tokens. The system then decides whether a query word exists in the given collection of speech documents.

Phone strings of utterances for phoneme search were generated using a decoded lattice of ASR. Based on translated queries of Swahili and given strings, queries are matched with entries in the pronunciation lexicon. When corresponding entries are missing, the G2P model was applied to the queries to obtain pronunciation of the queries. Then, exact matching of the pronunciation of queries with phoneme strings was performed based on a regular expression.

6.3. Data processing

Since the Kaldi speech recognition toolkit randomly selects a portion of speech data for a validation set on the fly, all of the provided speech data belonging to the "BUILD" partition was used for training of an acoustic model. Speech data was segmented into shorter speech utterances based on time-stamps of transcripts of phone conversation, because excessively long speech data leads to inefficient decoding. The "Evaluation" set was, however, not provided with time-stamped transcripts. Therefore, we decoded the "Eval" set once without segmenting it, and then created shorter utterances of the "Eval" set based on silence points in speech, in order to keep the maximum duration of speech utterance to 30 seconds. For training our language model, we used the provided speech transcripts of the training set and the external Tanzil dataset mentioned in Section 5.1.

	Example-1	Example-2
English query word	kick	messenger
PB-SMT-B	kiki	messenger
PB-SMT-BT	kiki	mtume
PB-SMT-BTP	kiki	mtume
PB-SMT-B top K words	piga; kiki	messenger
PB-SMT-BT top K words	kiki; kumpiga piga	mtume;mjumbe; mitume
PB-SMT-BTP top K words	kiki; kumpiga piga	mtume;mjumbe; mitume
Dictionary mapping	teke; kiki	mjumbe;mshenga; mtume;rasuli; tarishi;tume

Table 5: Examples of input query translation for an MT system. The translated hypotheses are sorted in decreasing order of translation scores.

7 Information Retrieval

In this section we describe the different components of our IR system, and present the tools and resources that are used for the development of the IR components.

7.1. Resources used and Data pre-processing

We used whoosh version 2.7.4, a python based library of classes and functions for performing IR operations such as indexing of documents and searching over the indexed collection.

Stopwords: We obtained top 10 words based on the term frequency in the document collection. We experimented with indexing and searching with and without stopwords, we found that retrieval results using stopwords are relatively better for our CLIR task. Our Swahili stopword list comprised of the following 10 words: ”ya”, ”wa”, ”na”, ”kwa”, ”katika”, ”la”, ”za”, ”ni”, ”le”, ”cha”.

7.2. Document indexing

For text-based retrieval and speech-based retrieval using ASR transcripts, we indexed the documents using the whoosh indexer. We removed stopwords and all non-alphanumeric keys from the data before indexing the raw documents. We maintained two separate indexes for text- and speech-based retrieval. We used these indexed collections to retrieve documents matching a given input query.
After query processing, the input queries are translated using the MT component described in Section 5. Thus for each of the input queries we have multiple translated candidates as shown in Table 5.

7.3. Document retrieval and ranking

To retrieve and rank documents effectively for a given query over the indexed collection we use the BM25 model (Robertson et al., 1995). BM25 is a probabilistic model that assigns a probability score to each document indicating its relevance to a given query. The investigations of document retrieval and ranking focused on two main aspects:

1. **Query translation selection (QTS)**: As shown above in Table 5, we have multiple possible translation hypothesis for a given query. We explored alternative methods to select and retrieve results corresponding to different translation hypotheses.
2. **Optimum threshold detection (OTD)**: The focus of the task is to maximize the number of relevant documents and minimize the number of non-relevant documents retrieved by the developed model. Thus we focus on selecting different cut-off rank to prune the retrieved ranked list to maximise measured retrieval effectiveness.

To find effective QTS and OTD techniques to boost retrieval performance we experimented with the $Analysis$ set. As described earlier in Table 1, we have gold relevance judgments ($qrels$) for the Analysis set, where for a set of queries we have corresponding relevant text and speech documents which can be used to develop and tune the text and speech retrieval models for optimal performance.
As shown in Table 6, the distribution of relevant documents across queries varies considerably for both the textual and speech collections. In the $analysis$ set about 43% of the queries have no relevant documents. About 2% of the queries have 10 or more relevant documents, with the maximum number of relevant documents being 18. The $analysis$ set is just 112 and 88 documents for text- and speech data respectively which is much less than the $evaluation$ set which has 5269 and 1217 documents, for text- and speech data respectively. Varying the size of the collections poses challenges for effective tuning of the system, such as determining the best cut-off rank for pruning the retrieved ranked list. We explored different cut-off ranks $[10, 15, 20]$ for the analysis set and $[50, 100, 200]$ for the evaluation set in our experiments.

	Total Queries	Queries with RR	No. of rel docs
Complete dataset	350	198	491
Text documents only	198	166	339
Speech documents only	198	99	152

Table 6: Distribution of relevance judgements for the analysis set. RR indicates relevant results, rel docs indicates the number of relevant documents

8 Results and Analysis

In this section we present our results on the analysis set for text- and speech-based retrieval before moving on to present our results on the evaluation set.[8]

8.1. Results on the Analysis set

Table 7 shows our results using different translation methods for the input queries for both text- and speech-based retrieval using the ASR approach. Due to the absence of other comparative models, we present retrieval results using queries translated using the Google translation engine[9] as a comparison for the behaviour of different MT systems explored in our work. We found that all alternative translation

[8] Due to space limitations we avoid results on development set as the composition of the development set is similar to analysis set, and instead present results with respect to the analysis and evaluation sets, which have quite different document collection sizes.

[9] https://translate.google.com/

		Text Retrieval				Speech Retrieval using only ASR			
Model	System	Recall ↑	Precision↑	AQWV↑	Rel Q ↑	Recall ↑	Precision↑	AQWV↑	Rel Q ↑
Google Translation		0.377	0.367	0.242	76	0.118	0.130	0.045	14
PB-SMT-B	MT-1	0.307	0.380	**0.202**	59	0.066	0.108	**0.015**	7
PB-SMT-BTP	MT-2	0.301	0.301	0.155	66	0.072	0.094	0.008	11
PB-SMT-BT	MT-3	0.295	0.376	0.193	59	0.066	0.104	0.013	9
PB-SMT-B top 3 hypotheses	MT-4	0.339	0.311	0.182	69	0.125	0.074	-0.020	14
PB-SMT-BTP top 3 hypotheses	MT-5	0.345	0.185	0.028	78	0.131	0.048	-0.112	18
PB-SMT-BT top 3 hypotheses	MT-6	0.319	0.268	0.137	67	0.092	0.051	-0.068	13
Dictionary	MT-7	0.407	0.191	0.047	76	0.105	0.050	-0.082	13

Table 7: Results on the Analysis set for text- and speech-based retrieval using only the ASR approach, where Rel Q indicates the number of relevant queries found by the system having atleast one relevant document, best scores are in bold face

models appear to find complementary relevant documents for different types of query as shown in Table 7.

Next, we explored combination approaches where we exploited the query translation results from different translation models. We explored an interpolation mechanism where we used a combination of MT systems (list of MT systems) for query translation, we perform query translation using the first MT system, and if we find no results using this first translation system for an input query, we perform query translation using the second translation system from the list of MT systems. For example for a translation system using a combination of $MT-1$, $MT-2$ and $MT-3$, first we perform search using the query translated through system $MT-1$, and if we retrieve zero results, we perform search using the query translated through system $MT-2$, and repeat until documents are retrieved or all MT systems have been tried. We find in our investigation that combining MT systems in this linear interpolated manner leads to less false positives (non-relevant results identified as relevant). The results of different interpolation approaches investigated in our work for both text- and speech-based retrieval are presented in Table 8. The best scoring MT systems are selected for carrying experiments on the evaluation set.

Table 7 and Table 8 show the best results for the analysis set which correspond to a cut-off rank of 20, where for each query we just retrieve and return the top 20 results. Table 9 shows the variations in the results for text- and speech-based retrieval while varying the number of top k documents retrieved using the MT-1 translation system. Table 10 presents results of the alternative speech retrieval approaches explored in our work. In the combined model for speech retrieval we combine the output of alternative retrieval approaches (ASR, Keyword search, and Phoneme search) to formulate a single ranked list for a given query. For Keyword and Phoneme search we used queries translated using PB-SMT-B (MT-1) system. There is a considerable difference in the speech retrieval results on the human transcribed data and the ASR results as shown in Table 10, indicating the need to improve the quality of ASR outputs.

8.2. Results on the Evaluation set

The main variations that we explore for the evaluation set correspond to: i) exploring the top MT systems and their combinations, and ii) varying the document cut-off ranks to $[50, 100, 200]$ for pruning the relevant results. There is a considerable difference in the cut-off rank for the analysis and evaluation sets, since the size of evaluation document collection is relatively bigger than the analysis document collection. Table 11 presents the results of our models on the evaluation set for text- and speech-based retrieval.

Main Findings and Challenges: In our work the best retrieval scores are attained by MT model combining the output of Dictionary + PBMT. We find it is better to combine the output of multiple MT systems rather than to rely on one best MT system for cross-lingual retrieval. The speech retrieval scores are relatively poor, reflecting that cross-lingual speech retrieval is quite a complex problem. Based on our analysis we can conclude that we need better methods and models for leveraging information from the different MT systems and the speech processing models to boost retrieval performance.

We investigated alternative methods for text- and speech-based retrieval. These are not the best results as we focus on the combination of different modules in a greedy manner rather than exploring the optimal best combination of the whole pipeline. We were interested in finding the individual best MT, Speech and IR systems and combining these to address the task of MMCLIR. Using the limited relevance judgments that were available for the analysis set, and the limited feedback provided on the evaluation set, we combined and investigated alternative approaches and explored different cut-off ranks for retrieving documents. We anticipate that given a larger relevance dataset ($qrels\ judgement$), we would be able to combine these different components more effectively to boost the retrieval performance. We learnt that unlike traditional MT modules and Speech modules, a combination of diverse MT systems, which capture diverse information, performs better overall for the MMLCIR task as indicated in Table 8.

9 Conclusion and Future work

In this work we investigated a MMCLIR task focusing on English-Swahili search carried out within the OpenCLIR challenge. We examine solutions to several challenges for MMCLIR in the context of low-resource availability. We investigated two research questions and examined alternative AT approaches for effective query translation. We build

	Text Retrieval				Speech Retrieval using only ASR			
MT Systems	Recall ↑	Precision↑	AQWV↑	Rel Q ↑	Recall ↑	Precision↑	AQWV↑	Rel Q ↑
MT-1+MT-2	0.354	0.397	0.242	73	0.092	0.122	0.030	10
MT-1+MT-2+MT-3	0.363	0.386	0.242	74	0.112	0.130	**0.042**	13
MT-1+MT-2+MT-4	0.372	0.341	0.222	79	0.131	0.079	-0.011	15
MT-1+MT-2+MT-7	0.466	0.280	0.216	99	0.118	0.061	-0.050	14
MT-1+MT-2+MT-3+MT-4	0.378	0.332	0.219	79	0.145	0.083	-0.004	17
MT-1+MT-2+MT-3+MT-7	0.460	0.273	0.205	97	0.125	0.063	-0.048	15
MT-1+MT-2+MT-4+MT-7	0.475	0.308	**0.253**	102	0.151	0.067	-0.043	18
MT-1+MT-2+MT-3+MT-4-MT-7	0.469	0.301	0.242	100	0.158	0.069	-0.041	19

Table 8: Interpolation model exploration, where Rel Q indicates the number of relevant queries found by the system having atleast one relevant document, best scores are in bold face, results on the Analysis set

	Text Retrieval			Speech Retrieval using only ASR		
Rank K	Recall↑	Precision ↑	AQWV scores↑	Recall↑	Precision ↑	AQWV scores↑
10	0.3038	0.4345	0.2214	0.0590	0.1058	0.0128
15	0.3067	0.3950	0.2090	0.0660	0.1075	0.0150
20	0.3067	0.3795	0.2022	0.0660	0.1075	0.0150
All results	0.3067	0.3795	0.2022	0.0660	0.1075	0.0150

Table 9: Optimum threshold selection, using MT-1 translation system, results on the Analysis set

Speech System	Recall↑	Rel Q↑	Precision ↑	AQWV scores↑
Using Human Transcriptions	0.2500	30	0.3064	0.1974
Using ASR (Single best MT)	0.0660	7	0.1075	0.0150
Phoneme Search	0.1052	12	0.0443	-0.1056
Keyword Spotting	0.0789	9	0.1237	**0.0269**
Combined	0.0197	22	0.0667	-0.0593

Table 10: Speech results on the Analysis set, where Rel Q indicates the number of relevant queries found by the system, best scores are in bold face

	Text Retrieval				Speech Retrieval		
System Settings	P_MISS_REL ↓	P_FA ↓	AQWV↑	System Settings	P_MISS_REL ↓	P_FA ↓	AQWV↑
Google, k=50	0.6933	0.0013	0.2804	**ASR Google**, k=50	0.8691	0.0047	**0.0362**
Google, k=100	0.6710	0.0020	0.2896	ASR Google, k=100	0.8603	0.0058	0.0240
Google, k=200	0.6590	0.0027	0.2864	ASR Google, k=200	0.8584	0.0066	0.0103
Sys-1, k=50	0.8005	0.0007	0.1853	ASR Sys-1, k=50	0.9213	0.0027	0.0255
Sys-1, k=100	0.7836	0.0011	0.1947	ASR Sys-1, k=100	0.9174	0.0034	0.0137
Sys-1, k=200	0.7756	0.0016	0.1934	ASR Sys-1, k=200	0.9162	0.0038	0.0074
Sys-2, k=50	0.6730	0.0011	0.3047	ASR Sys-2, k=50	0.9044	0.0037	0.0214
Sys-2, k=100	0.6535	0.0016	**0.3140**	ASR Sys-2, k=100	0.8980	0.0048	0.0058
Sys-2, k=200	0.6444	0.0022	0.3116	ASR Sys-2, k=200	0.8967	0.0054	-0.0050
–	–	–	–	Phoneme Search	0.9511	0.0011	0.0260
–	–	–	–	Keyword Search	0.9879	0.0043	-0.0736
–	–	–	–	Combined	0.9213	0.0027	0.0255

Table 11: Results on the evaluation set, Google and ASR Google indicates using Google translation, Sys-1 and ASR Sys-1 corresponds to the single best MT-1 translation system, Sys-2 corresponds to the MT combination system representing: <MT-1, MT-2, MT-4, MT-7> system, and ASR Sys-2 corresponds to the MT combination system representing: <MT-1, MT-2, MT-3> systems

an end to end system for MMCLIR using state-of-the-art MT, ASR and IR models. The retrieval scores are quite low, specifically for cross-lingual speech-based retrieval, indicating that there is likely to be quite some scope for improvement. There is a need to explore diverse mechanisms such as effective combination of multiple outputs to address the complex problem of MMCLIR involving multiple modalities and multiple languages.

We anticipate that work on MMCLIR will open new avenues and increase the scope of future research and promote interesting new research collaborations and pathways as the amount of multi-modal content is expected to rise very significantly as we consume and interact with more applications and content (Chang et al., 2019). In the future we would like to be able to explore new language pairs, and already plan to work on building better MT and Speech models to boost retrieval effectiveness.

Acknowledgement

This research is supported by Science Foundation Ireland (SFI) as a part of the ADAPT Centre at Dublin City University (Grant No: 12/CE/I2267).

10 Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of International Conference on Learning Representations (ICLR2015)*, San Diego, USA, May.

Boschee, E., Barry, J., Billa, J., Freedman, M., Gowda, T., Lignos, C., Palen-Michel, C., Pust, M., Khonglah, B. K., Madikeri, S., et al. (2019). Saral: A low-resource cross-lingual domain-focused information retrieval system for effective rapid document triage. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 19–24.

Buckley, C., Mitra, M., Walz, J. A., and Cardie, C. (1997). Using clustering and superconcepts within SMART: TREC 6. In *Proceedings of The Sixth Text REtrieval Conference, TREC 1997, Gaithersburg, Maryland, USA, November 19-21, 1997*, pages 107–124.

Chang, S.-F., Hauptmann, A., and Morency, L.-P. (2019). Key challenges for multimedia research in the next ten years.

Chen, S. F. and Goodman, J. (1995). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, pages 310–318.

Croft, W. B., Turtle, H. R., and Lewis, D. D. (1991). The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Chicago, Illinois, USA, October 13-16, 1991 (Special Issue of the SIGIR Forum)*, pages 32–45.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Hull, D. A. and Grefenstette, G. (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 49–57, New York, NY, USA. ACM.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54, Edmonton, Cannada.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open-Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions (ACL2007)*, pages 177–180, Prague, Czech Republic.

Leuski, A., Lin, C., Zhou, L., Germann, U., Och, F. J., and Hovy, E. H. (2003). Cross-lingual c*st*rd: English access to hindi information. *ACM Trans. Asian Lang. Inf. Process.*, 2(3):245–269.

Levow, G.-A., Oard, D. W., and Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Inf. Process. Manage.*, 41(3):523–547, May.

Littman, M. L., Dumais, S. T., and Landauer, T. K., (1998). *Automatic Cross-Language Information Retrieval Using Latent Semantic Indexing*, pages 51–62. Springer US, Boston, MA.

Madankar, M., Chandak, M., and Chavhan, N. (2016). Information retrieval system and machine translation: A review. *Procedia Computer Science*, 78:845 – 850. 1st International Conference on Information Security Privacy 2015.

Mohri, M., Pereira, F., and Riley, M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88.

Monti, J., Monteleone, M., di Buono, M. P., and Marano, F. (2013). Cross-lingual information retrieval and semantic interoperability for cultural heritage repositories. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 483–490, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Narasimha Raju, B. N. V., Bhadri Raju, M. S. V. S., and Satyanarayana, K. V. V. (2014). Translation approaches in cross language information retrieval. In *International Conference on Computing and Communication Technologies*, pages 1–4, Dec.

Oard, D. W. and Dorr, B. J. (1996). A survey of multilingual text retrieval. College Park, MD, USA. University of Maryland at College Park.

Oard, D. W., He, D., and Wang, J. (2008). User-assisted query translation for interactive cross-language information retrieval. *Inf. Process. Manage.*, 44(1):181–211.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics, Volume 29:1*, pages 19–51. MIT Press, Cambridge, Massachusetts, USA.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318, Philadephia, PA, USA, July.

Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Interspeech*, pages 2–6.

Picchi, E. and Peters, C., (1998). *Cross-Language Information Retrieval: A System for Comparable Corpus Querying*, pages 81–92. Springer US, Boston, MA.

Pirkola, A., Hedlund, T., Keskustalo, H., and Järvelin, K. (2001). Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval*, 4(3):209–230, Sep.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In *Pro-*

ceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pages 1–4.

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1995). Okapi at trec-3. *NIST special publication*, (500225):109–123.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *International Conference on Spoken Language Processing (ICSLP)*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2214–2218.

Trmal, J., Wiesner, M., Peddinti, V., Zhang, X., Ghahremani, P., Wang, Y., Manohar, V., Xu, H., Povey, D., and Khudanpur, S. (2017). The kaldi openkws system: Improving low resource keyword search. In *Proc. Interspeech 2017*, pages 3597–3601.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Yarmohammadi, M., Ma, X., Hisamoto, S., Rahman, M., Wang, Y., Xu, H., Povey, D., Koehn, P., and Duh, K. (2019). Robust document representations for cross-lingual information retrieval in low-resource settings. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track, MTSummit 2019, Dublin, Ireland, August 19-23, 2019*, pages 12–20.

Zbib, R., Zhao, L., Karakos, D., Hartmann, W., DeYoung, J., Huang, Z., Jiang, Z., Rivkin, N., Zhang, L., Schwartz, R., et al. (2019). Neural-network lexical translation for cross-lingual ir from text and speech. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 645–654.

Zens, R., Och, F. J., and Ney, H. (2002). Phrase-based statistical machine translation. In *KI 2002: Advances in Artificial Intelligence, 25th Annual German Conference on AI, KI 2002, Aachen, Germany, September 16-20, 2002, Proceedings*, pages 18–32.