# Robust Document Representations for Cross-Lingual Information Retrieval in Low-Resource Settings

**Mahsa Yarmohammadi[1], Xutai Ma[2], Sorami Hisamoto[3], Muhammad Rahman[4], Yiming Wang[5], Hainan Xu[6], Daniel Povey[7], Philipp Koehn[8] and Kevin Duh[9]**

Center for Language and Speech Processing,
Johns Hopkins University, Baltimore, MD, USA
{mahsa[1],xutai_ma[2],sorami[3],mahbubur[4],yiming.wang[5],phi[8]}@jhu.edu
{hainan.xv[6],dpovey[7]}@gmail.com kevinduh@cs.jhu.edu[9]

## Abstract

The goal of cross-lingual information retrieval (CLIR) is to find relevant documents written in languages different from that of the query. Robustness to translation errors is one of the main challenges for CLIR, especially in low-resource settings where there is limited training data for building machine translation (MT) systems or bilingual dictionaries. If the test collection contains speech documents, additional errors from automatic speech recognition (ASR) makes translation even more difficult. We propose a robust document representation that combines N-best translations and a novel bag-of-phrases output from various ASR/MT systems. We perform a comprehensive empirical analysis on three challenging collections; they consist of Somali, Swahili, and Tagalog speech/text documents to be retrieved by English queries. By comparing various ASR/MT systems with different error profiles, our results demonstrate that a richer document representation can consistently overcome issues in low translation accuracy for CLIR in low-resource settings.

## 1 Introduction

Cross-lingual Information Retrieval (CLIR) is a search task where the user's query is written in a language different from that of the documents in the collection. There are some important niche applications, for example, a local news reporter searching foreign-language news-feeds to obtain different perspectives for her story, or a patent writer exploring the patents in another country to understand prior art before submitting her application, or an aid worker monitoring the social media of a disaster-affected area, looking for unmet needs and new emergencies. In all these scenarios, CLIR increases the user base by enabling users who are not proficient in the foreign language to productively participate as knowledge workers. Even if the user requires manual translations of the retrieved documents to complete her task, CLIR can at least provide a triage/filtering step.

CLIR performance depends critically on the accuracy of its underlying machine translation or bilingual dictionary component. Recent advances in MT suggest that it is now ever more possible to build CLIR for practical use. In particular, the availability of large amounts of parallel text in some language-pairs (e.g. English sentences and their aligned German translations from European Parliamentary proceedings) had led to dramatic improvements in MT quality. However, there are many language-pairs–what we term "low-resource" settings–where parallel text is limited and the challenge is to make CLIR robust to translation errors. Missing words in translation may lead to degradation in recall, while extraneous words may lead to degradation in precision.

In this work, we focus on the document translation approach to CLIR, where all foreign documents in the collection are translated into the language of the user query prior to indexing and search. While the use of N-best translations in CLIR is not a new idea, the contribution of the paper is a comprehensive analysis of how different kinds of document representations perform under low-resource settings. We compare whether in-

dexing the N-best translations from MT leads to better CLIR than indexing only the 1-best (most-likely) translation. We also propose a novel bag-of-phrases document representation and show that it can be effectively combined with the N-best document representations. The idea behind the bag-of-phrases translation is the fact that less strict syntax is required in a CLIR system, which is often based on keyword search. The bag-of-phrases method relaxes the strict language grammar in the target language when producing translations, and instead, emphasizes the selection of translation words.

We perform comprehensive experiments on three low-resource test collections from the IARPA MATERIAL project (OpenCLIR Evaluation, 2018), where the documents are in Somali, Swahili, and Tagalog and the queries are in English. The inclusion of speech documents (audio files) in this collection means that automatic speech recognition (ASR) has to be run before MT, leading to further challenges in translation accuracy. Our results demonstrate that a rich document representation containing many translation hypotheses consistently improves CLIR performance in these low-resource settings.

## 2 Related Work

The key component in CLIR is translation, to resolve language gap between documents and queries. An appropriate approach is query translation (Oard et al., 2008), where the query is translated into the desired language based on a dictionary (Pirkola et al., 2001), or parallel corpora (Dumais et al., 1996). Query translation often suffers from translation ambiguity due to the limited amount of context in short queries. Another approach is document translation (Croft et al., 1991), which can produce more precise translation due to having more context. Several studies have compared the query translation and document translation approaches (Nie, 2010; Dwivedi and Chandra, 2016).

In recent years, deep neural networks have shown significant results on NLP tasks such as machine translation (Bahdanau et al., 2014), however, applying such models to information retrieval tasks has had relatively less positive results (Craswell et al., 2016). The reason is that, first, IR tasks are fundamentally different from NLP tasks, and second, the application of neural networks to IR problems has been under-explored. Recently some work on CLIR adopt word embedding approaches to use unlabeled text to learn the representations in unsupervised manner, and use them for document search (Vulić and Moens, 2015; Litschko et al., 2018; Josifoski et al., 2019). Such methods allow to learn representations from comparable data or independent monolingual data and alleviate the need for full-fledged machine translation. However, these methods are mostly useful when operating at Web scale, such as searching in Wikipedia articles, is considered. In this study, we focus on searching on a limited set of given documents in foreign low-resource language.

## 3 Task

The goal of the task we focus on this paper is to develop ASR, MT, and IR methods to most efficiently respond to queries against multilingual speech and text data in low-resource languages. The system will take English queries as input, and returns retrieved documents relevant to the queries as output. To resolve language differences in documents and queries, we focus on the document translation approach: all source documents in the foreign low-resource language are translated into English before search. Since some of the source documents are speech documents (audio files), we first run our ASR system on those to convert them to text before translation.

For each input query, the translated speech and text documents are searched via standard monolingual information retrieval approaches (e.g., BM25), which match words between query and document. Translation errors will naturally make this retrieval step more difficult. The retrieved documents are sorted according to their match scores, and we evaluate performance by comparing with the true (human-labeled) relevance ranking using standard metrics like Mean Averaged Precision (MAP).

## 4 Methods

### 4.1 Index and Search

Our CLIR engine is based on the document translation approach, where all foreign documents are translated beforehand and the English is what is indexed. We use a pre-existing search engine implementation Elasticsearch [1] to index, search, and

[1] https://www.elastic.co/products/elasticsearch

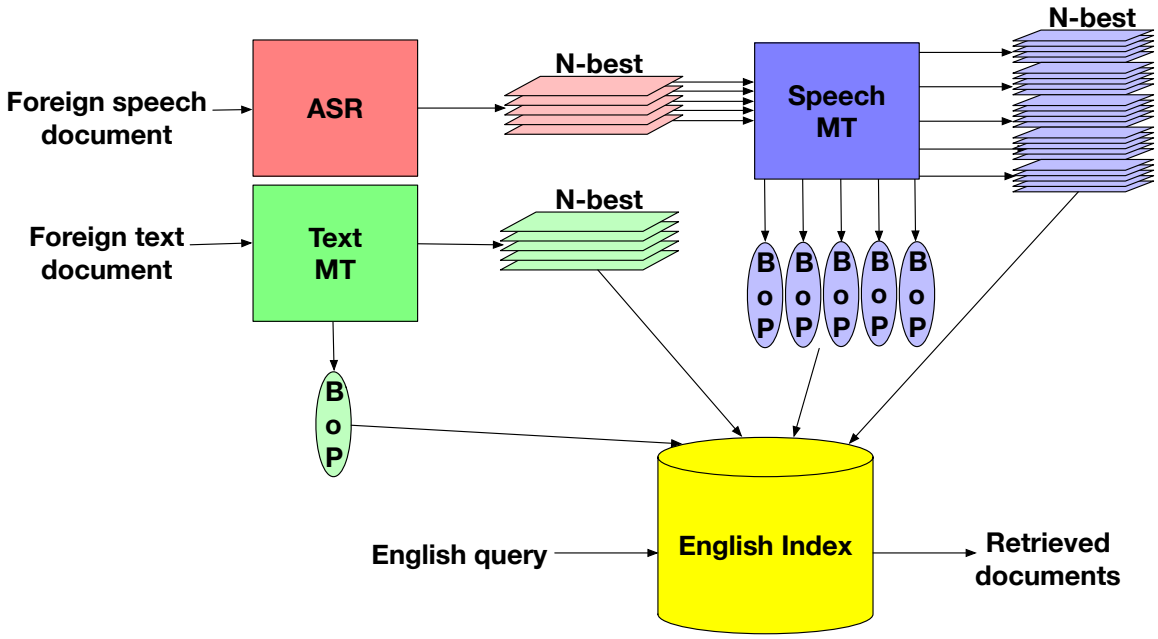


**Figure 1:** N-best+Bag-of-Phrases document representations for CLIR.

rank our translated documents. We use a standard built-in English analyzer to pre-process the document and query text. The analyzer conducts tokenization, word stemming, and stop word removal. We parse input query strings and convert them into Elasticsearch executable JSON format, then use those to retrieve search results from the Elasticsearch engine. We use Okapi BM25 (Robertson et al., 2009) algorithm to score the documents. BM25 is a popular algorithm to rank documents based on the relevance to a given query. We tuned the BM25 hyper parameters (for term frequency normalization and document length normalization), for each language to get the best CLIR performance. Finally the document ranking scores for each query are passed to the evaluation. CLIR performance is evaluated using the standard Mean Average Precision (MAP) measure.

### 4.2 Document Representations

To increase recall of documents and prevent error propagation from potential ASR or MT errors, we added multiple hypotheses capability to our CLIR pipeline. We implemented three types of pipelines, N-best decoding, bag-of-phrases, and combination of the two representations.

**N-best decoding** For speech documents, first, ASR generates N-best list for each input segment. Then MT decodes each of the ASR segment transcripts, generating M-best translations. The result is an N×M list, which is indexed into the IR system with equal weighting. We explored two variations of N-best decoding, first where the full N×M matrix is included in the document to be indexed. The second variation is where we sub-sample the full N×M matrix to its diagonal elements, that is the best translation of the best ASR output, the second best translation of the second best ASR output, and so on and so forth. We did not notice gains in the CLIR performance from including the full matrix in the document as opposed to including only its diagonals. This shows that the redundancy of hypotheses in the full matrix is not necessary for CLIR. For simplicity, we only present results where N=M. For text documents, MT generates N-best translations of each sentence.

**Bag-of-Phrases** For speech documents, first, ASR generates N-best list for each input segment. Then we use the phrase-based MT system to generate all possible phrases whose source side matches the ASR transcripts. In other words, we output all the translation options but do not perform a full decoding search with language models. For each input segment, all of the output phrases are concatenated together to form the bag-of-phrases for that segment. For $N > 1$, bag-of-phrases of all of N-best lists are considered. These bag-of-phrases are then indexed into the IR system. The same procedure is applied to each sentence in text documents.

**Combination of N-best decoding and Bag-of-**

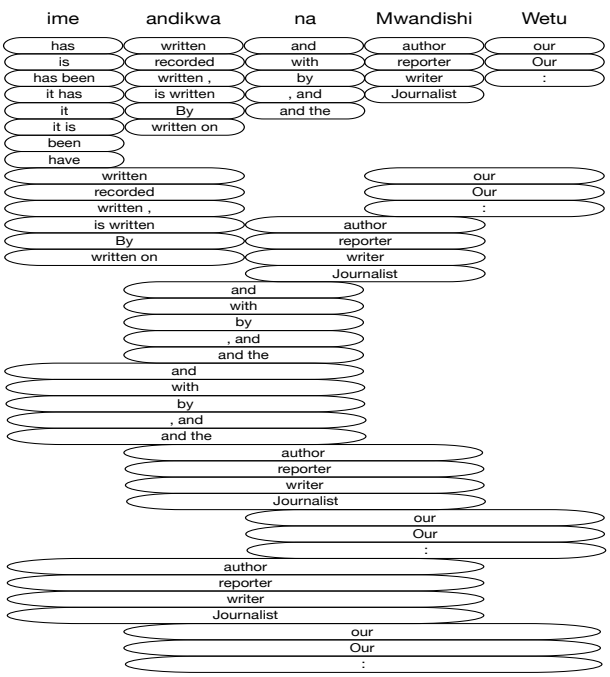


**Figure 2:** Bag-of-phrases (BoP) representation of the Swahili sentence "Imeandikwa na Mwandishi Wetu". The phrases in the boxes are all possible phrases that can be extracted from the phrase-based decoder.

**Phrases** Our IR system allows multiple "views" of the same document. We can index on both N-best decoding and bag-of-phrases (BoP). The search function will score documents based on how well the query matches either of the views. As shown in Figure 1, foreign text documents are run through a text MT system to produce N-best and BoP outputs. Foreign speech documents are first run through the ASR system to be transcribed to N-best hypotheses. The hypotheses are then run through the speech MT system, which is the same as the text MT system but adapted to interface better with ASR, to produce N-best and BoP outputs. Finally, N-best and BoP outputs from foreign speech and text documents are indexed and searched in response to English queries and relevant documents are retrieved.

By indexing and searching both N-best and BoP representations we not only consider the most accurate translations achieved via N-best but also take advantage of additional lexical variety provided by BoP. Figure 2 shows all possible phrases from the phrase-based MT decoder for the example input sentence "Imeandikwa na Mwandishi Wetu" in Swahili. These phrases, form the BoP representation of that sentence, and as can be seen, a variety of translations for different input spans are produced (e.g., the translations "author", "reporter", "writer", and "Journalist" for the Swahili word "Mwandishi"). The N-best (N=5) translations of the sentence are all the same sentence "it has been written by our writer" with different probabilities. Although the N-best output is a descent translation of the input in this example, it does not have as much word diversity as we could get from the BoP translation, thus hurting the retrieval of documents relevant to the query. For example, the word "Journalist" that is present in the BoP representation does not appear in the top 100 translations of the N-best representation. Thus, if a query includes that specific word, the chance of retrieving the document decreases if only the N-best representation is searched.

## 5 Data

### 5.1 CLIR Data

Given a query the system should detect which documents out of a set of documents are responsive to the query. Queries are English word strings that may contain words from any part of speech. There are different types of queries such as a lexical query consisting of a single word (e.g., "ocean"), a lexical query consisting of a multiple words (e.g., "bicycle race"), or conceptual queries that are subject to semantic expansion (e.g., "expiration+"). The set of documents includes speech and text documents from different genres. Table 1 shows the number of queries and documents we used for testing our CLIR system. Number of text documents is almost as twice as number of speech documents in each language.

### 5.2 ASR and MT Data

To train our ASR systems, we used "train" and "tune" data, which are transcribed conversational audio, as training and development sets. In addition, we used a large amount of untranscribed audio, the "unlabaled" set, for semi-supervised training of the acoustic model, as described in Section 6.1.

| | # queries | # documents (Foreign) | | |
|---|---|---|---|---|
| | (English) | speech | text | total |
| Somali | 442 | 279 | 559 | 838 |
| Swahili | 547 | 266 | 547 | 813 |
| Tagalog | 537 | 315 | 529 | 844 |

**Table 1:** CLIR test collection statistics.

| ASR | Length (hours) |
|---|---|
| train | ∼40 |
| tune | ∼10 |
| unlabeled | ∼250 |
| test | ∼20 |

| MT | train (# Eng tokens) | | test |
|---|---|---|---|
| | baseline | crawled | (#sent) |
| Somali | 800k | 1.7M | 9.5k |
| Swahili | 808k | 5.2M | 11.7k |
| Tagalog | 759k | 12.3M | 11.4k |

**Table 2:** ASR and MT data statistics.

We used parallel corpora (bitext) of around 800k English words to train our MT systems for translating from Somali, Swahili, or Tagalog to English. This data is provided in the BUILD package of the MATERIAL project and contains news, topical, and blog texts with provided source URLs. In addition, we harvested and filtered bitext from Web to augment this baseline bitext. We made this data publicly available[2]. It is important to filter web bitext to reduce noise. We filtered the web bitext using Zipporah (Xu and Koehn, 2017) and chose filter thresholds optimized on tune sets. The crawled data improved the MT system by 1 point BLEU or more for these languages. We also added monolingual WMT news and LDC Gigaword data, which include 8.2 billion English tokens in total to train the language models of our MT systems.

The IR system indexes and searches "test" documents that are either speech or text. There are around 20 hours of test speech data and 10k foreign sentences of test text data for each language. We have the reference transcripts and translations of "test", hence, we can measure the performance of our ASR and MT systems on the test set in terms of WER and BLEU scores, and also investigate how ASR/MT systems with different WER/BLEU scores impact CLIR. Table 2 shows the statistics of the ASR and MT data. For information about the number of test speech and text documents in each language see Table 1.

## 6 Experimental Setup

### 6.1 ASR system

Our ASR system follows normal pattern for Kaldi-based (Povey et al., 2011) system build. Our recipe is publicly available at GitHub[3].

**Acoustic and language model**. We use GMM training to create alignments and lattice-free MMI-trained neural network (Povey et al., 2016) with factorized TDNN (Povey et al., 2018). We generate lattices with n-gram ARPA-style language model and re-score them with an n-best RNN language model (Xu et al., 2018a; Xu et al., 2018b). Source-side bitext and crawled monolingual data are used in building the n-gram LM, RNNLM re-scoring, as well as extending the baseline lexicon.

In addition to supervised training, we ran semi-supervised training of acoustic models using the extension of lattice-free MMI to semi-supervised scenarios (Manohar et al., 2018). We added unlabeled audio to the labeled audio in the training set to train the acoustic model. Table 3 shows the WER improvements from supervised to semi-supervised setup for Somali, Swahili, and Tagalog. To study the effect of ASR errors on CLIR, we tried both supervised and semi-supervised ASR systems in our experiments.

**ASR input and output**. Test data come in long unsegmented files of over a minute. To deal with this, we split the input into equal-size (15 second) slightly overlapping segments and stitch together the ASR outputs. For consistency, we lower-case all text resources that are used in training the ASR system, which include transcripts and external resources for language modeling (source-side bitext, web crawled monolingual text). As a result, the ASR output would be all lower-case. However, the machine translation system expects inputs that have been tokenized and true-cased. Thus, we post-process ASR output to normalize punctuation, tokenize, and true-case using the models and scripts that are used in MT training and decoding. This post-processing helps passing names through the MT system, and improves the IR performance.

### 6.2 MT System

We tried phrase-based machine translation (PBMT) as well as neural machine translation (NMT) for Somali-English, Swahili-English, and Tagalog-English language pairs. The PBMT systems were developed using the Moses SMT toolkit (Koehn et al., 2007). We trained our systems with the following settings: a maximum sentence

[2] http://www.paracrawl.eu/

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

| ASR | | ASR1 | ASR2 |
|---|---|---|---|
| Somali | tune | 57.8 | 57.7 |
| | test | 56.7 | 48.4 |
| Swahili | tune | 38.9 | 36.7 |
| | test | 39.7 | 32.9 |
| Tagalog | tune | 47.5 | 46.6 |
| | test | 51.4 | 40.3 |

| MT | BLEU PBMT | BLEU NMT |
|---|---|---|
| Somali | 18.31 | 18.83 |
| Swahili | 28.66 | 30.18 |
| Tagalog | 33.05 | 29.95 |

**Table 3:** %WER for supervised (ASR1) and semi-supervised (ASR2) systems, BLEU scores for PBMT and NMT systems.

length of 80, grow-diag-final-and symmetrization of GIZA++ alignments, an interpolated Kneser-Ney smoothed 5-gram language model with KenLM (Heafield, 2011) used at runtime, hierarchical lexicalized reordering (Galley and Manning, 2008), a lexically-driven 5-gram operation sequence model (OSM) (Durrani et al., 2013) with count bin features (Chiang et al., 2009), a distortion limit of 6, maximum phrase-length of 5, 200-best translation options, compact phrase table (Junczys-Dowmunt, 2012) minimum Bayes risk decoding (Kumar and Byrne, 2004), cube pruning (Huang and Chiang, 2007), with a stack-size of 1000 during tuning and 5000 during test. We optimize feature function weights with k-best MIRA (Cherry and Foster, 2012).

The NMT systems are LSTM sequence-to-sequence models (Luong et al., 2015). The layer size is 512, and the number of layers is 4 for Swahili and Tagalog, 2 for Somali. The models were developed using the Fairseq[4] toolkit. For NMT, we applied Byte Pair Encoding (BPE) (Sennrich et al., 2016) to split word into subword segments for both source and target languages. The number of BPE operations is 3000 for all three languages. We observed improvements in BLEU scores under small BPE settings for all three language pairs.

We filtered noisy crawled bitext using Zipporah (Xu and Koehn, 2017) and applied the unsupervised morphology induction tool Morfessor (Virpioja et al., 2013) to split words up into putative morphemes, with keeping numbers and names unchanged. We noticed that splitting the words to morphemes improves BLEU scores for Somali and Swahili, but does not help for Tagalog.

To better translate speech documents, we built systems that are adapted to interface better with ASR, which we refer to as speech MT systems. For building speech MT systems, we removed punctuation and spelled out the numbers in the bitext before training the MT systems, which both improved BLEU scores.

[4]We used a PyTorch implementation : `https://github.com/pytorch/fairseq`

## 7 Results

We run our CLIR system using document representations based on a combination of N-best transcriptions/translations and the novel bag-of-phrases output from ASR/MT. A simple baseline for comparison is the query translation approach, where each word in English query is translated into its most likely foreign word using dictionary extracted from bitext. This baseline achieves the MAP scores of 0.0967, 0.1204, 0.2293 for Somali, Swahili, and Tagalog respectively, which all are inferior to the results we present in this section.

Table 4 shows MAP scores for different MT/ASR and document representation combinations for the three languages. For N-best and BoP representations, the results for $N = 5$ are shown in the table. For text, top 5 translations for each sentence are combined and indexed as the N-best document. For speech, 5 translations of the diagonal of the ASR $\times$ MT matrix for each speech segment are combined and indexed as the N-best document. For speech, BoP is the aggregation of bag-of-phrases translations of top 5 ASR outputs.

We observe that N-best+BoP achieves the best MAP scores across all settings. For example in the Somali ASR1+PBMT / PBMT pipeline, N-best+BoP achieves 0.2444, outperforming the 1-best baseline (0.1894), and isolated N-best (0.1902) and BoP (0.1999). This result even outperforms the 1-best reference translation (0.1956), indicating that a richer document representation based on multiple ASR/MT hypotheses, even if potentially error-prone, is better than a single professional translator's result in the context of CLIR. This is likely due to the challenge of finding exact

Somali

| Speech/Text | 1-best | N-best | BoP | N-best+BoP |
|---|---|---|---|---|
| ASR1+PBMT/PBMT | 0.1894 | 0.1902 | 0.1999 | **0.2444** |
| ASR2+PBMT/PBMT | 0.1970 | 0.2182 | 0.2080 | **0.2526** |
| ASR1+NMT/NMT | 0.1322 | **0.1623** | n/a | n/a |
| ASR2+NMT/NMT | 0.1321 | **0.1630** | n/a | n/a |
| ASR2+PBMT+NMT/PBMT+NMT | 0.1999 | 0.2231 | 0.2080 | **0.2521** |
| Ref transcript+PBMT/PBMT | 0.1965 | 0.2169 | 0.2268 | **0.2633** |
| Ref transcript+NMT/NMT | 0.1509 | **0.1788** | n/a | n/a |
| Ref translation/Ref translation | 0.1956 | n/a | n/a | n/a |

Swahili

| Speech/Text | 1-best | N-best | BoP | N-best+BoP |
|---|---|---|---|---|
| ASR1+PBMT/PBMT | 0.2234 | 0.2398 | 0.2072 | **0.2582** |
| ASR2+PBMT/PBMT | 0.2306 | 0.2474 | 0.2135 | **0.2634** |
| ASR1+NMT/NMT | 0.1897 | **0.2061** | n/a | n/a |
| ASR2+NMT/NMT | 0.1896 | **0.2104** | n/a | n/a |
| ASR2+PBMT+NMT/PBMT+NMT | 0.2299 | 0.2516 | 0.2135 | **0.2632** |
| Ref transcript+PBMT/PBMT | 0.2437 | 0.2600 | 0.2170 | **0.2768** |
| Ref transcript+NMT/NMT | 0.1902 | **0.2099** | n/a | n/a |
| Ref translation/Ref translation | 0.2408 | n/a | n/a | n/a |

Tagalog

| Speech/Text | 1-best | N-best | BoP | N-best+BoP |
|---|---|---|---|---|
| ASR1+PBMT/PBMT | 0.2947 | 0.3162 | 0.3114 | **0.3355** |
| ASR2+PBMT/PBMT | 0.2945 | 0.3159 | 0.3392 | **0.3617** |
| ASR1+NMT/NMT | 0.2226 | **0.2437** | n/a | n/a |
| ASR2+NMT/NMT | 0.2470 | **0.2683** | n/a | n/a |
| ASR2+PBMT+NMT/PBMT+NMT | 0.3150 | 0.3380 | 0.3392 | **0.3623** |
| Ref transcript+PBMT/PBMT | 0.3660 | 0.3906 | 0.3884 | **0.4187** |
| Ref transcript+NMT/NMT | 0.2803 | **0.3039** | n/a | n/a |
| Ref translation/Ref translation | 0.3847 | n/a | n/a | n/a |

**Table 4:** MAP scores for various ASR/MT systems and document representations (N=5) on Somali, Swahili, and Tagalog test sets.

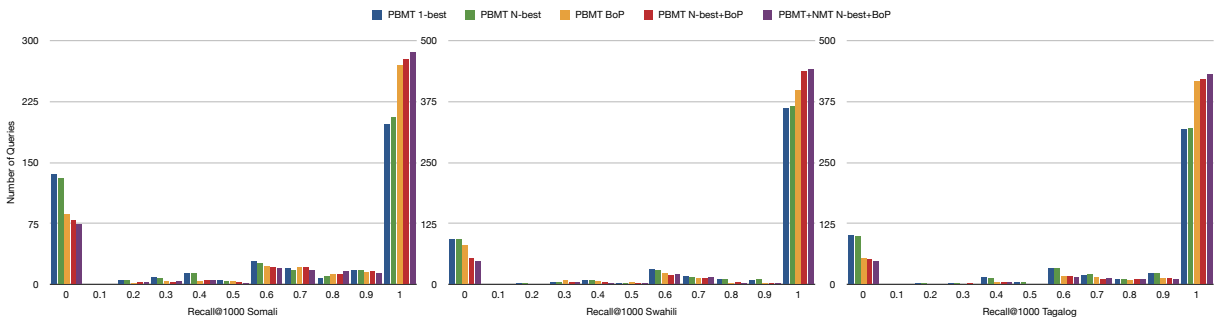


**Figure 3:** Per query recall@1000 for different systems.

match between the query and the document.[5] We also observe that the MAP scores from the ASR systems with lower word error rate (ASR2) are in general better than those from the ASR systems with higher word error rate (ASR1). This observation underscores the impact of a high quality ASR system on improving the performance of CLIR.

We noticed that NMT has much higher missed detection rate compared to PBMT, which turns into a low MAP score. Although the NMT model has comparable BLEU score, high missed detection indicate that NMT somehow fails to produce the tokens that IR system is interested in. More investigation of the reason is future work. We also use NMT translations as an additional field to PBMT translations (ASR2+PBMT+NMT / PBMT+NMT). We can observe that there is a small improvement over PBMT N-best+BoP method for Tagalog. We plotted number of queries versus the recall after 1000 documents are retrieved for different systems. As Figure 3 shows, when using N-best, BoP, N-best+BoP, and NMT as an additional feature, the number of queries with 0 recall decreases consistently in all three languages. This indicates that a richer document representation is indeed helping in retrieving relevant documents.

## 8 Conclusion and Future Work

The key component in CLIR is translation. The objective of translation in CLIR is different from Machine Translation tasks, as in information retrieval settings the goal is to retrieve relevant documents rather than having a high quality translation per se. In this study, we augmented high quality translation through N-best lists with the lexical variety of translation required for IR through BoP translations. We explored combinations of ASR and MT systems with different error profiles, and showed that our proposed N-best+BoP representation consistently performs well for CLIR on all three low-resource languages we studied. We plan to conduct various error analyses in future work to categorize the error types in our end-to-end CLIR system, as well as comparing PBMT and NMT systems. Another interesting future direction is to re-investigate these representations in the context of high-resource languages and stronger component systems, to contrast with the low-resource setting.

[5]Note that these results are not necessary our best results, since we have not tuned for scoring function and various other hyper-parameters. This exercise is meant to compare multiple systems in a simple setting that varies only the document representation.


## Acknowledgments

This work is supported in part by the Office of the Director of National Intelligence, Intelligence Advanced Research Projects Activity (IARPA), via contract #FA8650-17-C-9115. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.


## References


Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Cherry, Colin and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the NAACL-HLT*, pages 427–436. Association for Computational Linguistics.

Chiang, David, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *Proceedings of the 2009 Conference of the NAACL-HLT*, pages 218–226. Association for Computational Linguistics.

Craswell, Nick, W. Bruce Croft, Jiafeng Guo, Maarten de Rke, and Bhaskar Mitra. 2016. Report on the SIGIR 2016 workshop on Neural Information Retrieval (Neu-IR). *SIGIR Forum*, 50(2):96–103, December.

Croft, W Bruce, Howard R Turtle, and David D Lewis. 1991. The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th ACM SIGIR conference*, pages 32–45.

Dumais, Susan T., Thomas K. Landauer, and Michael L. Littman. 1996. Automatic cross-linguistic information retrieval using latent semantic indexing. In *Proceedings of SIGIR Workshop on cross-Linguistic information retrieval*, pages 16–23.

Durrani, Nadir, Alexander Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn. 2013. Can markov models over minimal translation units help phrase-based smt? In *Proceedings of the 51st Annual Meeting of the ACL*, volume 2, pages 399–405.

Dwivedi, Sanjay and Ganesh Chandra. 2016. A survey on cross language information retrieval. *Int'l Journal on Cybernetics & Informatics*, 5:127–142, 02.

Galley, Michel and Christopher D Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on*

*EMNLP*, pages 848–856. Association for Computational Linguistics.

Heafield, Kenneth. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197. Association for Computational Linguistics.

Huang, Liang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the 45th annual meeting of the ACL*, pages 144–151.

Josifoski, Martin, Ivan S. Paskov, Hristo S. Paskov, Martin Jaggi, and Robert West. 2019. Crosslingual document embedding as reduced-rank ridge regression. In *Proceedings of the 12 ACM International Conference on WSDM*, pages 744–752.

Junczys-Dowmunt, Marcin. 2012. A phrase table without phrases: Rank encoding for better phrase table compression. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180.

Kumar, Shankar and William Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the HLT-NAACL 2004*.

Litschko, Robert, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised cross-lingual information retrieval using monolingual data only. In *The 41st International ACM SIGIR Conference*, pages 1253–1256.

Luong, Thang, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on EMNLP*, pages 1412–1421.

Manohar, Vimal, Hossein Hadian, Daniel Povey, and Sanjeev Khudanpur. 2018. Semi-supervised training of acoustic models using lattice-free mmi. In *2018 IEEE ICASSP Conference*, pages 4844–4848.

Nie, Jian-Yun. 2010. *Cross-Language Information Retrieval*. Morgan and Claypool.

Oard, Douglas W, Daqing He, and Jianqiang Wang. 2008. User-assisted query translation for interactive cross-language information retrieval. *Information Processing & Management*, 44(1):181–211.

OpenCLIR Evaluation. 2018. `https://www.nist.gov/itl/iad/mig/openclir-evaluation`.

Pirkola, Ari, Turid Hedlund, Heikki Keskustalo, and Kalervo Järvelin. 2001. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information retrieval*, 4(3-4):209–230.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on ASRU*.

Povey, Daniel, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech 2016*, pages 2751–2755.

Povey, Daniel, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. 2018. Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proc. Interspeech 2018*, pages 3743–3747.

Robertson, Stephen, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 1715–1725.

Virpioja, Sami, Peter Smit, Stig-Arne Gronroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline. Technical report.

Vulić, Ivan and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference*, pages 363–372.

Xu, Hainan and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on EMNLP*, pages 2945–2950.

Xu, Hainan, Tongfei Chen, Dongji Gao, Yiming Wang, Ke Li, Nagendra Goel, Yishay Carmiel, Daniel Povey, and Sanjeev Khudanpur. 2018a. A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition. In *2018 IEEE ICASSP Conference*, pages 5929–5933.

Xu, Hainan, Ke Li, Yiming Wang, Jian Wang, Shiyin Kang, Xie Chen, Daniel Povey, and Sanjeev Khudanpur. 2018b. Neural network language modeling with letter-based features and importance sampling. In *2018 IEEE ICASSP Conference*, pages 6109–6113.