Neural-Network Lexical Translation for Cross-lingual IR from Text and Speech

Rabih Zbib
Raytheon BBN Technologies
Cambridge, MA
rabih.zbib@raytheon.com

Lingjun Zhao
Raytheon BBN Technologies
Cambridge, MA
lingjun.zhao@raytheon.com

Damianos Karakos
Raytheon BBN Technologies
Cambridge, MA
damianos.karakos@raytheon.com

William Hartmann
Raytheon BBN Technologies
Cambridge, MA
william.hartmann@raytheon.com

Jay DeYoung*
Northeastern University
Boston, MA
deyoung.j@husky.neu.edu

Zhongqiang Huang*
Alibaba Technologies
Hangzhou, China
z.huang@alibaba-inc.com

Zhuolin Jiang
Raytheon BBN Technologies
Cambridge, MA
zhuolin.jiang@raytheon.com

Noah Rivkin*
Franklin W. Olin College of Engineering
Newton, MA
nrivkin@olin.edu

Le Zhang
Raytheon BBN Technologies
Cambridge, MA
le.zhang@raytheon.com

Richard Schwartz
Raytheon BBN Technologies
Cambridge, MA
rich.schwartz@raytheon.com

John Makhoul
Raytheon BBN Technologies
Cambridge, MA
john.makhoul@raytheon.com

ABSTRACT

We propose a neural network model to estimate word translation probabilities for Cross-Lingual Information Retrieval (CLIR). The model estimates better probabilities for word translations than automatic word alignments alone, and generalizes to unseen source-target word pairs. We further improve the lexical neural translation model (and subsequently CLIR), by incorporating source word context, and by encoding the character sequences of input source words to generate translations of out-of-vocabulary words. To be effective, neural network models typically need training on large amounts of data labeled directly on the final task, in this case relevance to queries. In contrast, our approach only requires parallel data to train the translation model, and uses an unsupervised model to compute CLIR relevance scores.

We report results on the retrieval of text and speech documents from three morphologically complex languages with limited training data resources (Swahili, Tagalog, and Somali) and short English queries. Despite training on only about 2M words of parallel training data for each language, we obtain neural network translation models that are very effective for this task. We also obtain further improvements using (i) a modified relevance model, which uses the probability of occurrence of a translation of each query term in the source document, and (ii) confusion networks (instead of 1-best output) that encode multiple transcription alternatives in the output of an Automatic Speech Recognition (ASR) system.

We achieve overall MAP relative improvements of up to 24% on Swahili, 50% on Tagalog, and 39% on Somali over the baseline probabilistic model, and larger improvements over monolingual retrieval from machine translation output.

CCS CONCEPTS

• **Information systems → Information retrieval**; **Probabilistic retrieval models**; • **Computing methodologies → Neural networks**.

KEYWORDS

Cross-lingual information retrieval; speech recognition; machine translation; probabilistic modeling; neural networks

ACM Reference Format:
Rabih Zbib, Lingjun Zhao, Damianos Karakos, William Hartmann, Jay DeYoung*, Zhongqiang Huang*, Zhuolin Jiang, Noah Rivkin*, Le Zhang, Richard Schwartz, and John Makhoul. 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 (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3331184.3331222

*Work was done while the author was at Raytheon BBN Technologies.

SIGIR'19, July 2019, Paris, France
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331222

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Air Force Research Laboratory contract number FA8650-17-C-9118.

1 INTRODUCTION

Cross-lingual Information Retrieval (CLIR) has the challenge of contending with translation ambiguity, in addition to the general challenges of information retrieval. Whether the documents are translated to the query language or vice versa, it is important to consider alternative translations during retrieval for good CLIR performance [23], since alternatives provide more possibilities for matching query words in relevant documents compared to using a single translation. Using probabilities associated with the translations to compute retrieval scores is also beneficial, as these probabilities serve as an indicator of the system's confidence in the translations. An effective yet relatively simple method for CLIR is to use multiple translations of individual words, either of the queries or the documents to compute a relevance score [33], where the probabilistic word translations are usually generated from automatic word alignments of a parallel training corpus.

The coverage and quality of translations are a major factor in the performance of the CLIR system. In this paper, we propose a Neural Network Lexical Translation Model (NNLTM) to improve the estimation of source word translations, which are then used in CLIR. By virtue of mapping the input to continuous embeddings, the neural network generalizes to source-target words not directly observed in the parallel data alignments. We also condition the translations on the context surrounding the source word, thus making the translations context-dependent. The neural network can model the large input space of context-dependent word translations reliably, which would not be possible with a discrete model because of data sparsity. We use the word translations estimated by the NNLTM in the same unsupervised CLIR relevance model that uses alignment-based translations. Our proposed method only requires parallel training data to train the NNLTM, but no supervised relevance data, which is significantly harder to obtain. By translating the document rather than the query, the model takes advantage of the sentence context. We describe how indexing and retrieval of documents can still be done efficiently with the proposed model.

Estimating translations reliably is especially important when the amount of bilingual training data is limited. We report experiments on retrieval of Swahili, Tagalog and Somali text and speech documents against short English queries. We use queries and retrieval corpora from the IARPA MATERIAL [3] (Machine Translation for English Retrieval of Information in Any Language) program, and limited training data: roughly 2M words of parallel text data for each language, and between 48 and 128 hours of transcribed speech to train the Automatic Speech Recognition (ASR) system. We report large CLIR improvements on the three language. We further improve the neural model by encoding the character sequences of the source words using Convolutional Neural Networks (CNNs), thus allowing the model to estimate translations for out-of-vocabulary (OOV) source words, which are more common in low-resource and morphologically complex languages like the ones we report about in this work.

We compare probabilistic CLIR with monolingual retrieval performed on translated documents and show that alignment-based and NNLTM-based methods outperform retrieval on Machine Translation (MT) output, and also retrieval on human-produced reference translations.

We also propose an improved retrieval model based on the probability that each query term occurs *at least once* in the source document. This model is independent of the method in which the translations are estimated. We also propose a method to improve retrieval of speech documents.

To deal with speech recognition errors, which are especially prevalent with limited ASR training data, we propose a method for using multiple probabilistic outputs of the ASR system in the form of consensus networks (c-nets), and show improvements over using the 1-best output of ASR. We show further improvements resulting from the integration of the c-nets with NNLTM.

To summarize, the contributions of this paper are:

- A Neural Network Lexical Translation model for Cross-Lingual IR that uses source context and character-level encodings of the input. To our best knowledge, this is the first work using a neural lexical translation model for CLIR.
- An improved model for computing CLIR relevance based on the probability of occurrence of each query word at least once in the document.
- Improved retrieval of speech documents using a probabilistic ASR output instead of 1-best.
- A comparison of probabilistic CLIR to monolingual IR based on MT output and on human translations.

We achieve overall MAP relative improvements of up to 24% on Swahili, 50% on Tagalog and 39% on Swahili over the baseline probabilistic CLIR, and larger improvements over monolingual retrieval from machine translation output.

Although we describe our work in terms of retrieval of non-English documents based on English queries, the techniques we propose are language independent, and can be readily applied to any language pair.

2 PREVIOUS WORK

The approaches to CLIR that use a translation model can be divided into two broad categories: [23]: (i) translate the queries to the language of the retrieval corpus, or (ii) translate the retrieval corpus to the same language as the queries. In both cases, the CLIR problem is then reduced to monolingual information retrieval [22, 33]. Translation of the query is more efficient than translating the documents, but translation of the document words can take advantage of the larger context to disambiguate the translations.

An alternative approach is to directly model the CLIR problem without a separate translation step. For example, [17] proposed a cross-lingual relevance model to estimate the joint probability of query words and document words from a parallel corpus or a bilingual lexicon, and then they apply it to rank the documents. Although this model has the advantage of not requiring that we train a machine translation system on the parallel corpus, its drawback is that the coverage is limited to only those query and document word combinations found in the corpus. Sasaki et al. [28] designed a neural ranking model by adopting a ranking model from monolingual information retrieval. It estimates the joint probability of a query and document, and minimizes pairwise ranking loss. The limitation of this model is that it is hard to take advantage of existing parallel data for low resource languages, and the coverage is limited for various combinations of query and document words in the testing data.

Other approaches rely heavily on training bilingual embeddings. For example, Vulic et al. [32] trained bilingual embeddings on bilingual comparable document data, used them to construct document and query embeddings, and then ranked document relevance by computing cosine similarity. Litschko et al.[20] proposed to train bilingual embeddings using monolingual data. However, the approach to generate document and query embeddings is not done in a way that directly maximizes performance on the end task of CLIR relevance ranking. On the other hand, Bai et al. [4] developed a method for mapping documents and queries to a space that optimizes ranking performance.

Neural approaches to monolingual IR have received a lot of interest in recent years. [36] is a comprehensive review of neural IR. [7] proposed to train neural ranking models with weak supervision by using the output of a BM25 unsupervised ranking model. They represented the input document and query with weighted embeddings, and tried several ranking architectures. Our approach is complementary; in principle any of the CLIR models we describe in this paper can be used to generate the weak supervision labels, which could then be used to train the ranking model. In practice, it is an open question whether this approach would generalize effectively to the cross-lingual setting.

The neural network we propose for estimating word translations is similar to the model of [8], where they use the word translations as a feature in a phrase-based MT system. The translation probabilities in [8] are conditioned on the previous hypothesized target words and the source context. In our work, we condition on the source context only. The target context is not available since our model only requires single word translation probabilities.

Information retrieval from speech has traditionally been referred to as spoken document or spoken content retrieval. The standard approach is to cascade the ASR output with a text retrieval system [19].

For a recent overview of techniques in monolingual IR from speech, see [16]. Similar approaches have been applied to CLIR. Sheridan et al. [30] presented one of the first cross-lingual speech retrieval systems, though results were far from monolingual systems. Performance quickly improved with the yearly evaluations produced by the Cross-Language Evaluation Forum (CLEF) [9, 26].

In addition to the retrieval literature, work in keyword spotting (KWS) and spoken term detection (STD) [10] can also be considered a form of information retrieval. The IARPA Babel [1] program further advanced the state-of-the-art in KWS. Improvements came not only from general improvements in ASR systems [29] and features [14], but also in search strategies [5] and normalization [15]. Even when word error rate (WER) is high, KWS can still work remarkably well [12]. A similar conclusion was noted in the IR literature: that once WER reaches a point of about 35%, performance is not much worse than using reference transcripts [13].

3 CROSS-LINGUAL INFORMATION RETRIEVAL

3.1 Probabilistic Cross-Lingual IR Model

Following [22] and [33], we model the IR problem using a generative probabilistic model. We compute the probability that a document *Doc* in the retrieval corpus is relevant (denoted by *Doc* is *R*) given a user-issued query Q as:

$$P\big(Doc \text{ is } R \mid Q\big) = \frac{P\big(Q \mid Doc \text{ is } R\big) \times P\big(Doc \text{ is } R\big)}{P(Q)} \tag{1}$$

Ignoring $P(Q)$, which is independent of the documents, and assuming a uniform prior on document relevance, the relevance probability is then proportional to the probability of the query being "generated" from a relevant document (Eq 2a). This in turn is the probability of Q being generated from the document. Following [22], we also include a component that corresponds to the probability of the query being generated from a general language model in the query language (Eq 2b).

$$\begin{aligned} P\big(Doc \text{ is } R \mid Q\big) &\propto P\big(Q \mid Doc \text{ is } R\big) && \text{(2a)} \\ &= \prod_{q \in Q} \Big(\alpha P(q \mid Doc) + \big(1-\alpha\big) P_{LM}(q)\Big) && \text{(2b)} \\ &= \prod_{q \in Q} \Big(\alpha \sum_{f \in Doc} \frac{P(q \mid f)}{\mid Doc \mid} + \big(1-\alpha\big) P_{LM}(q)\Big) && \text{(2c)} \end{aligned}$$

The cross-lingual component of the model, that is the probability that the query is generated from the foreign document is the probability that each query term q is the translation of any foreign term f in the document (Eq 2c). This model requires a *probabilistic translation dictionary*, which we estimate from a parallel training corpus. In our baseline experiments, the translation dictionary is generated from the word alignments. We later show that a dictionary generated from a neural network lexical translation model improves CLIR. The language model component ($P_{LM}(q)$) is a backoff model that avoids zero scores for query terms that are not in the document. In our experiments we use a unigram language model and a weight of 1-α = 0.1. The probability produced by equation (2c) is used to rank the retrieval corpus documents with respect to an input query.

Note that this formulation assumes that the words of a multi-term query are independent. This assumption simplifies the model, since as it allows us to use single word translations. It is useful in the low-resource setting. The alternative would be to estimate translation probabilities for the whole query phrases, but with limited training data, the phrase translations and their probabilities will be noisy.

A key advantage of this model is that it considers alternative translations for the document words. As we show in the experimental results, using alternative translations for IR results in better performance than IR on a single translation of the documents, whether the translation is automatic or performed by humans.

We chose to translate the document terms into English rather than translate the query words to the foreign language; this allows us to extend the model to using the sentence translations that depend on the sentence context, as we describe in equation (4).

The retrieval is still done efficiently, by indexing the documents over the English terms that the document words can translate to. For each English term and each document, we store the expected count of that term in the document obtained by multiplying the foreign term count with its probability of translating to the English term. At retrieval time, to compute the probability of relevance for a given query we look up the expected counts for its terms in the index and combine them with the unigram LM probabilities.

3.2 Probability of Occurrence Model

The cross-lingual component of (2c) is the product of expected counts of query terms with respect to the lexical translation probability distribution. The normalization of the translation probabilities $\sum_{f \in Doc} P(q \mid f)$ by the size of the document $|Doc|$ has the effect of penalizing longer documents compared to shorter ones. We propose a model that mitigates this effect by computing the relevance score as the probability that each query term occurs *at least once* in the document. If $\mathcal{T}(Doc)$ is the set of valid translations of all words and phrases in document Doc, the relevance score is computed as:

$$\begin{aligned} P\big(Doc \text{ is } R \mid Q\big) &= P\big(Q \text{ occurs at least once in } \mathcal{T}(Doc)\big) && \text{(3a)} \\ &= \prod_{q \in Q} P\big(q \text{ occurs at least once in } \mathcal{T}(Doc)\big) && \text{(3b)} \\ &= \prod_{q \in Q} \Big[1 - P\big(q \notin \mathcal{T}(Doc)\big)\Big] && \text{(3c)} \\ &= \prod_{q \in Q} \Big[1 - \prod_{f \in Doc} P\big(q \notin \mathcal{T}(f)\big)\Big] && \text{(3d)} \\ &= \prod_{q \in Q} \Big[1 - \prod_{f \in Doc} \big(1 - p\,(q \mid f)\big)\Big] && \text{(3e)} \end{aligned}$$

In equation (3b), we make the same assumption of the model of Section 3.1, which is that the query terms are independent of each other. Retrieval can be done efficiently by pre-computing the probability term of equation (3e) for each English term in the translation dictionary and then indexing the corpus documents by those terms ahead of time.

This model will still assign a higher score to documents that have a higher expected count of the query terms, but does not penalize large documents proportionally to their length as the model of Section 3.1 does. The goal is similar to lower-bounding TF normalization in BM25+ [21] for example.

3.3 CLIR for Speech

The default procedure for performing CLIR on speech documents is to transcribe the document using an Automatic Speech Recognition (ASR) system, and then perform the retrieval on the 1-best output in the same way that retrieval on text is done. The disadvantage of this approach is that the retrieval is performed on the erroneous output of the ASR system. The CLIR system has no way of recovering from speech recognition errors. The effect of recognition errors is especially severe in low resource settings, where the Word Error Rate (WER) is high - on the order of 30% to 50%.

In fact, the ASR system produces multiple outputs with confidence scores, which we propose to use for CLIR in order to mitigate the effect of ASR errors. We modify the relevance model of equation 3a by multiplying the term corresponding to each foreign word f with the probability that f is recognized by ASR. Denoting the latter by $p\,(f \mid Doc)$, the relevance equation becomes:

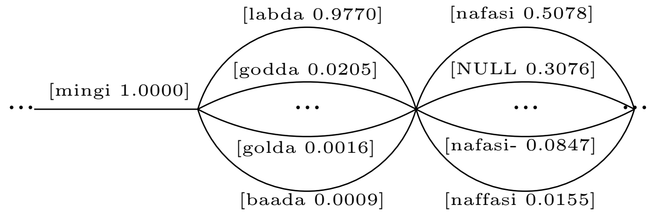

Figure 1: Example consensus network for a Swahili utterance.

$$P\big(Doc \text{ is } R \mid Q\big) = \prod_{q \in Q} \Big[1 - \prod_{f \in Doc} \big(1 - p\,(f \mid Doc)\,p\,(q \mid f)\big)\Big] \quad (4)$$

We compute $p\,(f \mid Doc)$ from the *consensus network* (c-net) output of the ASR system. A consensus network is obtained from a speech lattice by aligning the lattice arcs to form a "sausage". Figure 1 shows an example of a consensus network network of a Swahili utterance. Note that the ASR system might hypothesize that a certain slot in the network has no speech (the arc tagged NULL in the example network). The CLIR system ignores this arc, and uses the other alternatives.

We compute the probability of each unique term f in the c-net of a foreign document Doc as the probability that f appears at any location i in the document:

$$\begin{aligned} p\,(f \mid Doc) &= p\,(\text{f occurs at least in one position } i \text{ in } Doc) && \text{(5a)} \\ &= 1 - \prod_{i \in 1 \ldots |Doc|} \Big[1 - p(f|i)\Big] && \text{(5b)} \end{aligned}$$

$p(f|i)$ is the probability that term f is produced in location i obtained from the c-net.

Note that the terms $p(f|Doc)$ can be pre-computed for the whole corpus ahead of time, and therefore the terms in equation (5a) can also be pre-computed for each English word in the translation dictionary and used to index the documents for efficient retrieval.

4 NEURAL NETWORK MODEL FOR LEXICAL TRANSLATIONS

The models of Sections 3.1 and 3.2 both rely on the word alignments for the probabilistic lexical translations that they use to compute the relevance scores. The translations suffer from two limitations:

1. The translation probabilities are a *context-independent* probabilistic dictionary: the same translation probability distribution $p\,(q \mid f)$ is used for a given foreign term f across the whole corpus.
2. The translations do not generalize to unseen word pairs: Unless a source-target word pair (f, q) occurs in a parallel training sentence, and is aligned, its translation probability will be zero.

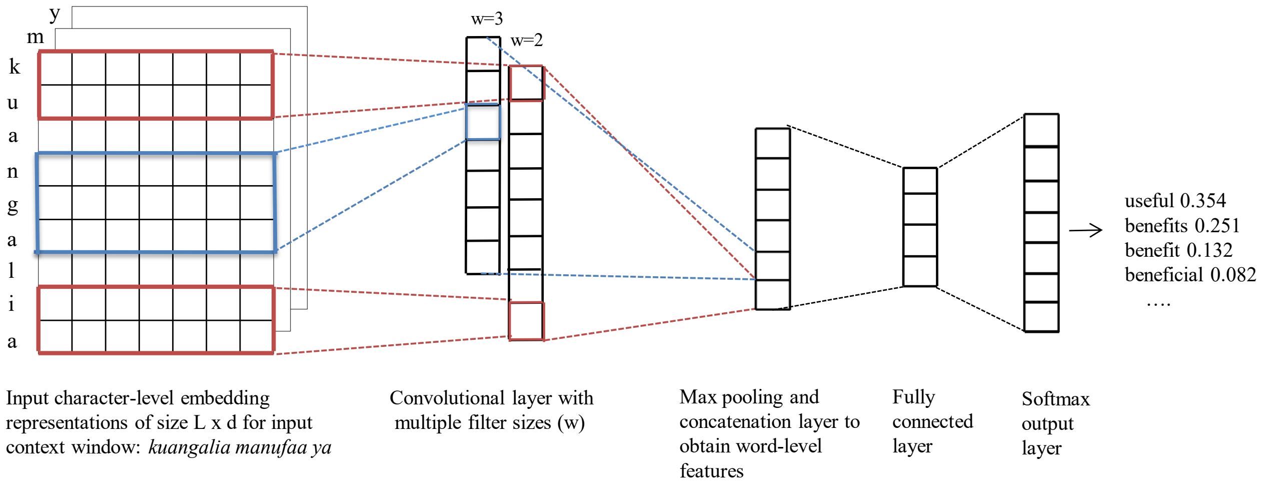
y
m
k
u
a
n
g
a
l
i
a
w=3
w=2
useful 0.354
benefits 0.251
benefit 0.132
beneficial 0.082
….
Input character-level embedding representations of size L x d for input context window: *kuangalia manufaa ya*
Convolutional layer with multiple filter sizes (w)
Max pooling and concatenation layer to obtain word-level features
Fully connected layer
Softmax output layer

Figure 2: Character level NNLTM architecture. Each word from the input context window is decomposed into characters, then fed into a CNN and max-pooling layer to obtain word-level features, which are finally used for lexical translation.

The meaning of a word, and therefore its translations depend on the sentence context in which it occurs. So one would expect the use of context-dependent translation distributions to improve CLIR accuracy. The second limitation is especially disadvantageous in low training-resource settings, where many valid translation pairs would typically not be observed, which then results in relevant documents being missed by the CLIR model.

To address these limitations, we propose using a context-dependent *Neural Network Lexical Translation Model (NNLTM)* in CLIR. The model computes the probability of translation of a given source word in its sentence context into the query word. The translation probabilities are then used in the baseline CLIR model or the Probability of Occurrence model in the same way that alignment-based translations are used. We then propose an variant of the NNLTM that encodes the character sequences of source words instead of whole words, to make the model more robust to morphological and spelling variations, and make it generalize better to out-of-vocabulary source word translations.

4.1 Word-Level NNLTM

Formally, given a source word f_i and a context window of k words on each side:

$$C_k(f_i) = f_{i-k}, f_{i-k+1}, ..., f_i, ..., f_{i+k-1}, f_{i+k} \tag{6}$$

the NNLTM estimates $P\big(q \mid C_k(f)\big)$, the context-dependent probability of translation of f_i into every word q in the target language vocabulary. The model is implemented as a feed-forward neural network. Each of the $2k + 1$ input words is mapped to a separate d-dimensional embedding vector. The separate embeddings allow the model to estimate separate parameters for each position in the context window. We use one h-dimensional hidden layer with *tanh* activation function. Additional hidden layers did not yield further improvements in our experiments. We limit the source vocabulary to the most frequent n words in the parallel training data, and map other words to *unknown*. The output layer is a *softmax* over the entire output vocabulary, which is the target vocabulary of training data. The neural network is trained on samples extracted from parallel sentences. We align the corpus automatically using the same procedure that generates the probabilistic dictionary of the baseline models, then for each aligned source-target word pair (f_i, q_j) we create a training sample $(C_k(f_i), q_j)$.

Neural networks are used almost exclusively lately for modeling a large context. Modeling context-dependent translations using discrete probabilities would suffer from data sparsity as the training data would not contain enough samples to estimate the probabilities reliably. Another advantage of the neural network over the discrete probabilities is that the neural network, by mapping the input to an embedding space, can generalize to other words that might have similar meanings as the aligned word pairs, even though they were not observed in the parallel data. The discrete model would assign a translation probability of zero to such pairs. Both of these advantageous are especially important in low-resource settings.

4.2 Character-Level NNLTM

The word-level model generalizes to unseen source-target word pairs, but it still cannot estimate translations for a source word that is unseen in the training data. We generalize the model to encode the character sequence of source words rather than embed the whole source words. Encoding character representations of the input rather that whole words has been shown to improve other NLP tasks such as neural machine translation ([34], [18]). Modeling the input as character sequences mitigates the problem of out-of-vocabulary, which is especially severe for low-resource languages, and also allows the model to be robust to morphological variations, a challenge particularly for the languages we report on in this paper since they all have a rich morphology.

Figure 2 shows a diagram of the character-level model. In this case, we still use a word context window; we decompose each word f_j in $C_k(f)$ into a sequence of characters $[x_1, x_2, ..., x_l]_{f_j}$, and pad the characters to a maximum length L. We then map this sequence to d-dimensional character embeddings:

$$X_{f_j} = [g(x_1), g(x_2), ..., g(x_L)]_{f_j} \tag{7}$$

where g is the character embedding look up table: $g \in \mathbb{R}^{d \times |x|}$. The resulting representation for f_j is then a $L \times d$ dimension matrix X_{f_j}.

We use a one-dimensional narrow convolution(CNN) filter $H \in \mathbb{R}^{w \times d}$ of width w on each character embedding matrix X_{f_j} ($j = i - k, i - k + 1, ..., i + k$), and add a bias with $ReLu$ non-linearity to obtain a feature map $M^{f_j} \in \mathbb{R}^{L-w+1}$, such that the l-th element of M^{f_j} is given by:

$$M^{f_j}[l] = ReLu(\langle X_{f_j}[*, l : l + w - 1], H \rangle + b) \tag{8}$$

where $X_{f_j}[*, l : l + w - 1]$ is a slice of X_{f_j} that contains all columns between l and $l + w - 1$, and $\langle A, B \rangle = \mathrm{Tr}(AB^T)$ is the sum of the element-wise product of matrices A and B (Frobenius inner product). Then we apply max-pooling on the feature to obtain y_{f_j}:

$$y_{f_j} = max_l M^{f_j}[l] \tag{9}$$

For m convolutional filters, we would obtain features for f_j:

$$Y_{f_j} = y^1_{f_j}, y^2_{f_j}, \ldots, y^m_{f_j} \tag{10}$$

The character-level representation of the whole input $C_k(f)$ is:

$$F(C_k(f)) = Y_{f_{i-k}}, Y_{f_{i-k+1}}, \ldots, Y_{f_{i+k}} \tag{11}$$

Similar to the word-level model, the input is then fed to an h-dimensional hidden layer with $tanh$ activation function.

In Section 5.6 we report significant CLIR improvements from the NNLTM, and further improvements from Character-Level NNLTM. In addition to allowing sentence context to be used, our choice of translating words in the source documents as opposed to translating the query allows us to encode the character sequence of the input on the source language side, but still decode English words. Source-language input tends to be noisier (more spelling variations) than the target since the former is naturally occurring, while the latter is produced by translators. Also, the languages we deal with have significantly more complex morphologies than English, and therefore benefit more from character-level modeling.

4.3 Training and Decoding

For both models, we use Adam optimization to minimize the cross-entropy loss of predictions between the output layer and reference target translations. We apply dropout with probability $p_{dropout}$ on all trainable parameters except for the input embeddings to reduce over-fitting. The detailed experiment setup and results are given in Section 5.6.

During decoding, we output the target words with K top probabilities $P(q \mid C_k(f))$. We still do efficient retrieval by decoding each source word in the retrieval corpus and creating an index from the target words to the documents, which we use to look up the probabilities to compute the relevance score. The number of decodings needed is at worst equal to the number of words in the retrieval corpus. The decodings can be performed efficiently using GPUs.

Lang	Data Set	Text	Speech	Queries	#
Swahili	Tune	(547, 237k)	(266, 236k)	Q_{Tune}	400
	Test	(449, 196k)	(217, 77k)	Q_{Test1}	300
	$Test_L$	(10435, 4835k)	(4309, 1445k)	$Q_{Test1+2}$	900
Tagalog	Tune	(291, 169k)	(315, 134k)	Q_{Tune}	400
	Test	(460, 232k)	(244, 100k)	Q_{Test}	300
Somali	Tune	(480, 218k)	(279, 145k)	Q_{Tune}	400
	Test	(482, 157k)	(213, 103k)	Q_{Test}	300

Table 1: Retrieval data statistics. The number of documents and the number of tokens (thousands) is shown for each retrieval corpus.

5 EXPERIMENTAL RESULTS

5.1 Query Sets and Retrieval Corpora

We next discuss the CLIR results. We report experimental results on two conditions for each of the three languages: A *Tune* condition, which we use to tune high-level parameters, such as interpolation weights or context size, and a *Test* condition, that we evaluate blindly. We use different query sets for the different conditions. For Swahili, we also have available a larger corpus[1] (*Test*$_L$), on which we present additional results. Note that the CLIR models themselves are unsupervised, and we only use the relevance labels of the *Tune* sets to estimate the high-level parameters. Table 1 shows the statistics of the retrieval corpora. The average number of relevant documents per query for the different retrieval conditions ranges between 0.09% and 0.25%.[2]

5.2 Data Resources

5.2.1 *Parallel data.*

The same parallel training data was used to train the MT systems and to estimate the probabilistic dictionaries (for both the alignment-based dictionary and the neural lexical translation models). The data consists mostly of parallel sentences released under the MATERIAL and the LORELEI [2] programs. We also include a parallel lexicon downloaded automatically from Panlex (https://panlex.org/). Table 3 shows the amount of parallel data (sentences and words) for each language.

5.2.2 *Speech.*

The amount of transcribed speech used to train the ASR system varies for each language: 68 hours for Swahili, 128 hours for Tagalog, and 48 hours for Somali. In addition to the MATERIAL data, Swahili and Tagalog also include training data from the IARPA Babel program [1]. It should be noted that the transcribed training data contains only conversational telephone speech, while the evaluation corpora consists mostly of broadcast data.

[1] The corpus used for the official MATERIAL program evaluation

[2] Note that, in this paper, we have chosen to ignore two aspects of the MATERIAL data collections, since they do not affect our conclusions about the relative differences of the various models: (i) the "domain" constraints on the queries; and (ii) the fact that some documents are spoken/written in a different (distractor) language. Taking these aspects into account requires training separate models for topic and language identification (which get combined with the CLIR models described in this paper); due to lack of space, we cannot describe these models in sufficient detail here, but we plan to do this in a future publication.

MAP	Tune/$\mathbf{Q_{Tune}}$								Test/$\mathbf{Q_{Test1}}$					
	Text				Speech				Text			Speech		
	prob.	occ.	MT	Ref Transl.	prob.	occ.	MT	Ref Transl.	prob.	occ.	MT	prob.	occ.	MT
Swahili	40.2	43.7	27.4	30.1	49.6	50.0	47.0	44.8	39.7	45.4	34.4	43.7	47.3	36.2
Tagalog	57.1	57.8	42.1	45.6	59.1	57.8	48.6	50.1	49.8	57.2	42.4	55.8	58.7	60.2
Somali	32.2	40.0	25.7	25.0	29.4	31.0	18.1	22.7	29.3	36.9	27.3	23.1	25.4	19.8
									Test$_\mathbf{L}$/$\mathbf{Q_{Test1+2}}$					
Swahili	–	–	–	–	–	–	–	–	23.3	24.1	13.6	21.1	22.2	13.7

Table 2: MAP scores comparing between the probabilistic model (prob.) and the probability of occurrence model (occ.) , CLIR from MT 1-best output (MT) and CLIR from reference translations. The first three models use 1-best ASR output. The reference translations were done over the speech reference transcripts. The occurrence model achieves the best results overall.

	Parallel Data		**Tune/$\mathbf{Q_{Tune}}$ Text**		
Lang	**Parallel text (sents, words)**	**Parallel lexicon (words)**	**OOV**	**BLEU**	**MAP**
Swahili	(72k, 1738k)	190k	4.99%	36.0	49.0
Tagalog	(98k, 1950k)	65k	4.25%	43.0	63.9
Somali	(98k, 2278k)	8k	13.7%	21.5	41.4

Table 3: Parallel training resources for each language. The OOV rate, BLEU score of the MT output, and MAP score of the probability of occurrence model are also shown for the Tunes sets.

We report CLIR performance using Mean Average Precision (MAP). Other metrics, such as gmap (Geometric Mean Average Precision) and P@5 correlate very well with MAP. We don't show scores in other metrics for brevity, since those metrics don't change the conclusions we draw.

5.3 Comparison of Baseline and Probability of Occurrence

We first compare the performance of the baseline Probabilistic Model (Section 3.1) to that of the Probability of Occurrence Model (Section 3.2). In both cases, we use a probabilistic lexical dictionary obtained from automatic word alignments of the parallel training data. We concatenate the output of two aligners: GIZA++ [24], and the Berkeley Aligner [11], and then estimate the forward (target given source) translation probabilities by normalizing the alignment counts. Table 2 shows the MAP scores for both models. On the Test set, the Probability of Occurrence model improves the scores for text and speech documents for all data sets in the three languages. We see a larger increase on text (5 to 7 points) than on speech (2-3 points). The text documents are on average longer than the speech documents, and therefore they benefit more from the Probability of Occurrence model since, unlike the baseline model, it does not penalize longer documents by normalizing the expected count of matched query words. We run CLIR on the 1-best ASR output for all conditions. It is worth noting that the MAP scores for Swahili Test$_L$are significantly lower than those for the smaller Test set. The Test$_L$corpus is much larger (15,000 docs. vs. 700 docs), which results in a lower average precision over the queries. The Probability of Occurrence model yields 1 MAP point gain on Test$_L$.

5.4 Comparison of MT+IR vs. Probabilistic CLIR

We next compare the results of both probabilistic CLIR models to monolingual retrieval using the 1-best MT of the retrieval documents. We train a state-of-the-art Neural MT system using the Transformer model [31], using the same parallel training data used to train word alignments and estimate the translation probabilities we used in the probabilistic CLIR models. We train a multi-lingual model by combining the parallel data from the three languages, then further fine-tune the model for each language separately using the corresponding parallel data.

It is known that alternative translations are crucial for good CLIR performance [23], since the alternative provide more possibilities for matching query words in relevant documents than a single translation. Table 2 shows that MAP scores for both the probabilistic baseline and the Probability of Occurrence models are significantly higher than those corresponding to CLIR based on 1-best MT on all languages and sets, expect for the speech subset of Tagalog Test. We particularly note that the gain is up to 10 points on the more challenging large Swahili test set (Test$_L$).

One could argue that the CLIR from 1-best MT is at a disadvantage in this comparison, and that a more fair comparison would use the N-best MT output. Using N-best MT output has two disadvantages: (i) Producing N-best MT output (and 1-best output for that matter) requires full sentence decoding, which is significantly more complex and more computationally expensive than producing probabilistic word translations, and (ii) It is not clear how translation probabilities can be obtained for the query terms from the MT output; system scores for the overall hypotheses from the N-best could be used, but those might not reflect the probabilities of the specific query terms accurately.

Reference translations produced by professional translators are available for the Tune sets, which allows us to make the comparison between the probabilistic CLIR methods and monolingual IR independent of MT errors. The results in Table 2 show that the probabilistic CLIR methods still perform significantly better than monolingual IR on the reference translations, and that the latter improve over MT-output IR only slightly (up to 3 MAP points). This

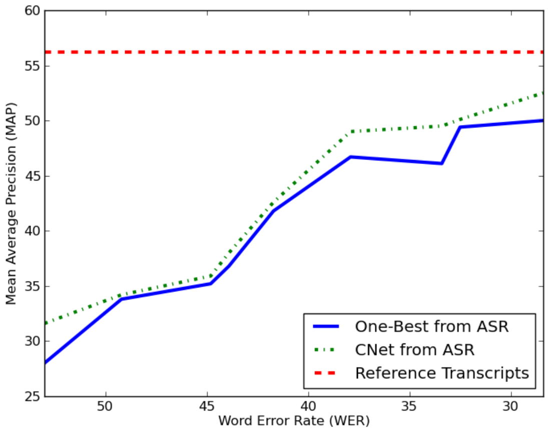

Figure 3: MAP vs. WER on Swahili. Retrieval from ASR output using nine different acoustic model and language model configurations are shown demonstrating the relationship between WER and MAP.

is further indication of the importance of alternatives, since even when the human translator translates a given query term correctly, that particular translation might not match the query term.

5.5 CLIR for Speech

For ASR we use a speech processing platform [anonymous], which integrates multiple machine learning toolkits, and uses Kaldi [27] for acoustic model training. Our acoustic models are pre-trained on 1500 hours of data from 11 languages [anonymous] and then fine-tuned to the target language. We use a CNN-LSTM acoustic model, which is similar to the recently proposed TDNN-LSTM [6], but with eight additional convolutional layers prepended to the network. Performance is further improved through semi-supervised training. After an initial decoding of the evaluation data, the acoustic model is retrained with the hypothesized transcripts. During decoding we use standard trigram language models.

Overall performance on spoken documents can be seen in Table 4. Due to recognition errors, the use of the one-best output from the ASR system is inherently sub-optimal. In all cases we see a significant performance improvement from using the consensus network (c-net) output from the ASR system. For Swahili and Tagalog, while the WER is relatively low, there is still a gap in performance compared to the reference transcripts. Somali performs significantly worse than the other two languages. This is likely caused by a variety of issues including the reduced training data and inconsistencies in orthography. While Somali does perform worse overall, it is interesting to note that the ASR actually outperforms the reference transcripts. This could also be related to the orthography. Since the ASR inherently produces spelling variants as alternatives, this could partially alleviate the difficulties of translating words not present in the machine translation model due to spelling variations.

Figure 3 more fully demonstrates the relationship between MAP and WER on Swahili. The results are generated by mixing different acoustic and language models. The models range from simple feed forward acoustic models trained only on the Swahili training transcripts, to our best system described above. Reducing WER below 40% required augmenting the language model and lexicon with additional data collected from the web [35]. As the WER decreases, we see a steady increase in MAP. In addition, the gap between the ASR performance and the reference transcripts steadily decreases. We also note that in all cases the consensus nets provide better performance than the one-best transcript.

5.6 Neural Network Lexical Translation CLIR Results

Next, we describe the CLIR results using the Neural Network Lexical Translation Model (NNLTM). For all NNLTM-related experiments, we use the NNLTM translation probabilities in the Occurrence model of Section 3.2.

5.6.1 NNLTM Experimental Settings. Table 5 describes the architectures of the word-level and character-level NNLTMs. For the character-level model, we limit the maximum word length L to be 30, and use convolutional widths of [1, 2, 3, 4, 5, 6, 7], and 500 filters for each size.

We use the same word alignments of the baseline models (GIZA++ [24], and Berkeley [11]) and extend the source word with its context to extract the training samples. Both models are trained using Adam for 20 epochs with a batch size of 512. We use a dropout probability $p_{dropout}$ of 0.8 for the word-level model and 0.7 for the character-level model. The learning rate is 0.001 for the word-level model and 0.0005 for the character-level model. We use the 10 translations with the highest probabilities from NNLTM output in CLIR.

5.6.2 Neural Lexical Translation Model Results. Table 6 shows the MAP scores for the word-level and character-level NNLTMs for the three languages. *Occ.* in the table corresponds to the baseline using the alignment-based dictionary, *NN(wd)* is the word-level NNLTM, and *NN(ch)* is the character-level NNLTM. We highlight and compare the results of each model bellow:

1. **Word-level NNLTM vs. Baseline.** On the Tune set, the Word-level NNLTM improves over the alignment-based baseline across all conditions. With the exception of Swahili Speech and Tagalog Text, we see an improvement on the Test set. We note the 3 point improvement on the more challenging Test$_L$set. The results show that the translation probabilities produced by the NNLTM result in better CLIR.
2. **Character-level vs. Word-level NNLTM.** The character-level model improves even further on most conditions, showing the benefit of the model's ability to deal with OOVs and spelling variations. For speech documents, we integrate the NNLTM with the consensus network output of ASR (Section 3.3), by decoding the output of the consensus network with the NNLTM and multiplying the two probabilities. The right-most column of Table 6 shows further improvement on Swahili and Somali Test sets.

Comparing the final condition to the alignment-based baseline, we see an improvement on all sets and all languages ranging from around 2 to 7 MAP points.

			$\text{Tune/Q}_{\text{Tune}}$			$\text{Test/Q}_{\text{Test1}}$		$\text{Test}_{\text{L}}\text{/Q}_{\text{Test1+2}}$	
	ASR Training	WER	MAP			MAP		MAP	
			1-best	c-net	transcr.	1-best	c-net	1-best	c-net
Swahili	68 hrs	28.4	50.0	52.5	56.2	47.3	51.7	22.2	25.0
Tagalog	128 hrs	29.1	57.8	61.5	69.4	58.7	66.4	—	—
Somali	48 hrs	45.7	31.0	32.2	30.9	25.4	28.2	—	—

Table 4: MAP results showing the gain from using ASR confusion networks instead of the 1-best transcription. For the Tune set, MAP scores for running CLIR on the reference transcripts are shown. The number of speech hours used for training the ASR system and the corresponding Word Error Rate (WER) are also shown for each language.

	Word Level	Character Level
Source vocab size	30,000	427
Target vocab size	54,042	54,042
Embedding size *d*	256	15
Hidden layer size *h*	128	256

Table 5: Word level and character level NNLTM architectures.

5.6.3 Effect of NNLTM Context Size. To study how context size affects word-level and character-level NNLTMs, we test different context sizes for the Swahili Tune set. Table 7 shows the MAP scores on the text and speech parts of the corpus. The optimal context size is different for different conditions, but two points are worth noting. The first is that context size greater than 0 performs better than context 0 on all conditions, and the second is that the 0-context model still perform better than the alignment-based baseline which indicates that some of the benefits of the NNLTM come from the generalization capability of the neural network and from encoding the input character sequence.

5.7 Comparison between languages

The various techniques we have introduced in this paper show consistent gains across the three languages. But we note that the range of scores for Tagalog is significantly higher than that for Swahili, which is in turn higher than Somali. We further investigate the difference between the languages by measuring the out of vocabulary rate of the Tune set with respect to the parallel data, and the BLEU [25] scores on the MT of the Tune sets in Table 3. Despite all three languages having a comparable amount of training data, we see a large difference in BLEU scores, correlating with the MAP scores, and a difference in OOV rates. We verified that the difference in OOV and BLEU is not due to the smaller size of the lexicon used for Somali, by excluding the lexicons from the MT training data. In a sense, Somali is more challenging that Swahili, which in turn is more challenging than Tagalog. This observation is consistent with anecdotal knowledge that Somali orthography is not strongly standardized, and that it reflects dialectal variation, and with the fact that Swahili is spoken across a much wider geographic area than Tagalog.

6 CONCLUSIONS AND FUTURE WORK

In this paper we propose using a neural network lexical translation model for CLIR. We show that through the neural network's ability to generalize to unobserved source-target word pairs, by using source context in the translation probability estimations, and through encoding the character sequences of the words, the model estimates better translation probabilities. Using these translation probabilities in an unsupervised CLIR model we obtain significant improvements over a baseline that uses the translation probabilities estimated directly from word alignments, on a retrieval task from three languages with limited training resources. We also propose an improved CLIR model based on the probability of the query words occurring at least once.

In the future, we plan to develop models that estimate the translation probabilities of multi-word queries directly, rather than treat the query words independently. We also plan to pursue models that detect the occurrence of the English query in a foreign sentence directly, and drop reliance on word alignments, which are usually noisy, especially when the amount of parallel training data is limited.

REFERENCES

[1] 2011. IARPA Babel Program - Broad Agency Announcement (BAA). https://www.iarpa.gov/index.php/research-programs/babel.

[2] 2015. DARPA LORELEI Program - Broad Agency Announcement (BAA). https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents.

[3] 2017. IARPA MATERIAL Program - Broad Agency Announcement (BAA). https://www.iarpa.gov/index.php/research-programs/material.

[4] Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Weinberger. 2010. Learning to rank with (a lot of) word features. *Information Retrieval* 13, 3 (2010), 291–314.

[5] Guoguo Chen, Oguz Yilmaz, Jan Trmal, Daniel Povey, and Sanjeev Khudanpur. 2013. Using proxies for OOV keywords in the keyword search task. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 416–421.

[6] Gaofeng Cheng, Vijayaditya Peddinti, Daniel Povey, Vimal Manohar, Sanjeev Khudanpur, and Yonghong Yan. 2017. An exploration of dropout with LSTMs. In *Proc. Interspeech*.

[7] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural Ranking Models with Weak Supervision. In *SIGIR*.

[8] Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard M. Schwartz, and John Makhoul. 2014. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*. 1370–1380. http://aclweb.org/anthology/P/P14/P14-1129.pdf

[9] Marcello Federico, Nicola Bertoldi, Gina-Anne Levow, and Gareth JF Jones. 2004. CLEF 2004 cross-language spoken document retrieval track. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 816–820.

[10] Jonathan G. Fiscus, Jerome Ajot, and John S. Garofolo. 2007. Results of the 2006 Spoken Term Detection Evaluation.

[11] Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. Better Word Alignments with Supervised ITG Models. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2 (ACL '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 923–931. http://dl.acm.org/citation.cfm?id=1690219.1690276

MAP	Tune/Q_{Tune}							Test/Q_{Test1}						
	Text			Speech				Text			Speech			
	occ.	NN (wd)	NN (ch)	occ.	NN (wd)	NN (ch)	NN (ch) +c-net	occ.	NN (wd)	NN (ch)	occ.	NN (wd)	NN (ch)	NN (ch) +c-net
Swahili	43.7	47.1	49.0	50.0	52.5	51.7	52.1	45.4	49.2	46.2	47.3	46.1	48.3	50.8
Tagalog	57.8	60.1	63.9	57.8	62.6	62.0	64.5	57.2	56.5	63.7	58.7	63.9	62.3	61.2
Somali	40.0	40.0	41.4	31.0	31.9	32.5	32.6	36.9	38.6	40.8	25.4	28.1	29.0	29.6
								Test_{L}/$Q_{\text{Test1+2}}$						
Swahili	—	—	—	—	—	—	—	23.3	26.3	26.4	21.1	24.5	24.6	26.1

Table 6: CLIR results comparing the Probability of Occurrence model using alignment-derived translations (occ.) to the same model using the Word-Level NNLTM (NN(wd)) and the Character-level NNLTM (NN(ch)). NN(ch)+c-net shows results of integrating the Character-Level NNLTM with ASR consensus networks.

	Word-level NNLTM		Character-level NNLTM	
MAP	**Text**	**Speech**	**Text**	**Speech**
Occ. (baseline)	43.7	48.6	43.7	48.6
Context=0	46.3	51.6	47.4	49.5
Context=1	47.1	52.5	47.9	55.0
Context=2	46.6	53.0	49.0	51.7
Context=3	47.0	53.9	47.4	53.4

Table 7: MAP scores for Swahili NNLTMs with different context sizes on Tune/Q_{Tune}.

[12] William Hartmann, Damianos Karakos, Roger Hsiao, Le Zhang, Tanel Alumäe, Stavros Tsakalidis, and Richard Schwartz. 2017. Analysis of keyword spotting performance across IARPA babel languages. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5765–5769.

[13] Alexander G Hauptmann, Rong Jin, and Tobun Dorbin Ng. 2002. Multi-modal information retrieval from broadcast video using ocr and speech recognition. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*. ACM, 160–161.

[14] Martin Karafiát, Frantisek Grezl, Mirko Hannemann, and Jan Honza Cernocky. 2014. BUT neural network features for spontaneous vietnamese in BABEL. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5622–5626.

[15] Damianos Karakos, Richard Schwartz, Stavros Tsakalidis, Le Zhang, Shivesh Ranjan, Tim Tim Ng, Roger Hsiao, Guruprasad Saikumar, Ivan Bulyko, Long Nguyen, et al. 2013. Score normalization and system combination for improved keyword spotting. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 210–215.

[16] Martha Larson, Gareth JF Jones, et al. 2012. Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends® in Information Retrieval* 5, 4–5 (2012), 235–422.

[17] Victor Lavrenko, Martin Choquette, and W. Bruce Croft. 2002. Cross-lingual Relevance Models. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)*. ACM, New York, NY, USA, 175–182. https://doi.org/10.1145/564376.564408

[18] Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully Character-Level Neural Machine Translation without Explicit Segmentation. *TACL* 5 (2017), 365–378.

[19] Lin-shan Lee, James Glass, Hung-yi Lee, and Chun-an Chan. 2015. Spoken content retrieval—beyond cascading speech recognition with text retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 9 (2015), 1389–1420.

[20] Robert Litschko, Goran Glavas, Simone Paolo Ponzetto, and Ivan Vulic. 2018. Unsupervised Cross-Lingual Information Retrieval Using Monolingual Data Only. In *SIGIR*.

[21] Yuanhua Lv and ChengXiang Zhai. 2011. Lower-bounding Term Frequency Normalization. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*. ACM, New York, NY, USA, 7–16. https://doi.org/10.1145/2063576.2063584

[22] David R. H. Miller, Tim Leek, and Richard M. Schwartz. 1999. A Hidden Markov Model Information Retrieval System. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*. ACM, New York, NY, USA, 214–221. https://doi.org/10.1145/312624.312680

[23] Jian-Yun Nie. 2010. *Cross-Language Information Retrieval*. Morgan and Claypool Publishers.

[24] Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29, 1 (2003), 19–51.

[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. 311–318.

[26] Pavel Pecina, Petra Hoffmannová, Gareth JF Jones, Ying Zhang, and Douglas W Oard. 2007. Overview of the CLEF-2007 cross-language speech retrieval track. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 674–686.

[27] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

[28] Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-Lingual Learning-to-Rank with Shared Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 458–463. http://aclweb.org/anthology/N18-2073

[29] Tom Sercu, Christian Puhrsch, Brian Kingsbury, and Yann LeCun. 2016. Very deep multilingual convolutional neural networks for LVCSR. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4955–4959.

[30] Paraic Sheridan, Martin Wechsler, and Peter Schäuble. 1997. Cross-language speech retrieval: Establishing a baseline performance. In *ACM SIGIR Forum*, Vol. 31. ACM, 99–108.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[32] Ivan Vulic and Marie-Francine Moens. 2015. Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings. In *SIGIR*.

[33] Jinxi Xu and Ralph Weischedel. 2000. Cross-lingual Information Retrieval Using Hidden Markov Models. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13 (EMNLP '00)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 95–103. https://doi.org/10.3115/1117794.1117806

[34] Yoon, Yacine Kim, David Jernite, Alexander Sontag, and Rush. 2016. Character-Aware Neural Language Models. In *2016 Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*. 2741–2749.

[35] Le Zhang, Damianos Karakos, William Hartmann, Roger Hsiao, Richard Schwartz, and Stavros Tsakalidis. 2015. Enhancing Low Resource Keyword Spotting with Automatically Retrieved Web Documents. In *Interspeech*. 839–843.

[36] Yingjie Zhang, Md. Mustafizur Rahman, Alex Braylan, Brandon Dang, Heng-Lu Chang, Henna Kim, Quinten McNamara, Aaron Angert, Edward Banner, Vivek Khetan, Tyler McDonnell, An Thanh Nguyen, Dan Xu, Byron C. Wallace, and Matthew Lease. 2016. Neural Information Retrieval: A Literature Review. *CoRR* abs/1611.06792 (2016).