Experiments with Cross-Language Speech Retrieval for Lower-Resource Languages

Suraj Nair[1], Anton Ragni[2], Ondrej Klejch[3], Petra Galuščáková[1], and Douglas Oard[1]

[1] University of Maryland, College Park MD 20742, USA
{srnair,petra,doug}@umd.edu
[2] University of Cambridge, Cambridge CB2 1TN, UK
ar527@cam.ac.uk
[3] University of Edinburgh, Edinburgh EH8 9YL, UK
ondrej.klejch@gmail.com

Abstract. Cross-language speech retrieval systems face a cascade of errors due to transcription and translation ambiguity. Using 1-best speech recognition and 1-best translation in such a scenario could adversely affect recall if those 1-best system guesses are not correct. Accurately representing transcription and translation probabilities could therefore improve recall, although possibly at some cost in precision. The difficulty of the task is exacerbated when working with languages for which limited resources are available, since both recognition and translation probabilities may be less accurate in such cases. This paper explores the combination of expected term counts from recognition with expected term counts from translation to perform cross-language speech retrieval in which the queries are in English and the spoken content to be retrieved is in Tagalog or Swahili. Experiments were conducted using two query types, one focused on term presence and the other focused on topical retrieval. Overall, the results show that significant improvements in ranking quality result from modeling transcription and recognition ambiguity, even in lower-resource settings, and that adapting the ranking model to specific query types can yield further improvements.

1 Introduction

The problem of Cross-Language Information Retrieval (CLIR) involves finding relevant documents in one language for a given query in different language. For example, one straightforward approach to CLIR is to use a Machine Translation (MT) system for translating queries into the document language and performing the retrieval in the document language. In the case of cross-language speech retrieval, the system must also determine which words were spoken in each speech "document." This might, for example, be done by transcribing the speech using Automatic Speech Recognition (ASR). If ASR and MT were perfect, we would expect this approach to yield retrieval results that are about as good as could be achieved by monolingual text retrieval. Neither process is perfect, however, and

moreover fairly good ASR and MT systems are available for only a few dozen of the world's thousands of languages. ASR and MT errors arise from two causes: (1) Out-Of-Vocabulary (OOV) errors in which words that are not known to the system aren't correctly handled, and (2) selection errors, in which the system could have selected the correct transcription or translation, but failed to do so. Modern ASR and MT systems learn to minimize both types of errors by training on large corpora. When large training corpora are not available, as is still the case for the vast majority of the world's languages, both types of errors increase. Those errors can adversely affect cross-language speech retrieval results.

These challenges are well understood, and methods have been developed for mitigating the effects of OOV and selection errors in CLIR and in monolingual speech retrieval. In early dictionary-based CLIR research, pre-translation query expansion helped to mitigate the effect of translation OOV errors by augmenting the query with related terms that may be translatable [13]. Today, dictionaries are often used together with translation lexicons learned from parallel (i.e., translation-equivalent) corpora [8]. Such an approach allows alternative translations to be weighted using translation probability weights learned from corpus statistics [25]. In speech retrieval, similar approaches can be used to mitigate the effect of transcription errors, again with reliance on transcription probabilities to limit the effect of selection errors [1].

In this paper, we leverage two new test collections to study error mitigation techniques for cross-language speech retrieval with English queries and spoken content in either Tagalog (a Philippine language) or Swahili (an African language). Section 2 introduces related work on CLIR, speech retrieval and cross-language speech retrieval. Section 3 describes our test collections, and Section 4 then introduces our CLIR and speech retrieval techniques and how those techniques are used together for cross-language speech retrieval. Section 5 presents our experiment design, results, and discussion of those results. Finally, we conclude the paper with some remarks on future work.

2 Related Work

Our approach to cross-language speech retrieval draws on three lines of research that we summarize in this section.

2.1 Cross-Language Information Retrieval

Much early work on CLIR involved replacing query terms with dictionary translations and then searching with that set of translations as a document-language query. Using multiple dictionary translations can help to avoid selection errors, but at the risk of considerably lower precision than a correct single selection could have produced. Pirkola [17] is generally credited with having been the first to introduce the idea of a "structured query" method for CLIR, although the idea has antecedents in Hull's work a year earlier [6]. In this method, term frequency statistics for alternative translations of the same query term are used

differently from term frequency statistics for different languages in order to limit selection error effects. This idea was subsequently extended to incorporate translation probabilities by Darwish in an approach that is now known as Probabilistic Structured Queries (PSQ) [4], although this idea too had an antecedents in the work of Xu [29] three years earlier. Subsequently the PSQ method was extended by Wang [25] to leverage evidence for meaning equivalence constructed from bidirectional translation probabilities. We use Wang's method in our experiments.

2.2 Speech Retrieval

A good deal of the early work on retrieval of spoken content involved cascading the output of an ASR system to a text retrieval system [2, 9–11, 24, 27]. This works fairly well when ASR transcription accuracy is high, but at word error rates above about 30% the adverse effect on recall becomes severe. This happens because ASR systems typically do best on common words, but the Inverse Document Frequency statistic used by many query-document matching methods (including the BM25 scores used in our experiments) gives the most weight to the least common query words. In early work, OOV terms that were not known to the ASR system explained some of the failures to correctly transcribe less common terms, but as the vocabularies of ASR systems have grown, selection errors have clearly emerged as the most common cause for the transcription errors. In other words, ASR systems trained for lower-resource languages often make mistakes on less common words, but when they do it is because they guessed wrong, not because they didn't know the word. ASR systems typically do, however, generate internal representations of the alternative words that might plausibly have been spoken (e.g., as a word lattice), and probabilities for these alternative hypotheses can be estimated from acoustic model, language model and pronunciation probabilities. When word error rates are relatively high, as is the case for ASR systems that are trained to recognize speech in lower-resource languages, these transcription probabilities can be used to compute expected term counts in a manner that is strikingly similar to the term frequency estimation in PSQ [1, 22]. We use expected term counts in our experiments.

There has also been considerable work on speech recognition and keyword spotting for lower-resource languages, including Tagalog [3, 7, 12, 19, 26]. With one recent exception [30], all of this work has focused on monolingual applications. The one exception, Zbib's SIGIR 2019 paper, also uses MATERIAL test collections, but with different retrieval models than we use.

2.3 Evaluation of Cross-Language Speech Retrieval

There has been far less work on cross-language speech retrieval than on CLIR for text, in part because ASR for speech is far more expensive computationally than is the corresponding process (tokenization) for text, and in part because cross-language speech retrieval test collections have until now been rather rare.

The first widely available cross-language speech retrieval test collections were produced for the Topic Detection and Tracking (TDT) evaluations between 1999

and 2004. In TDT, the closest task to retrieval was a filtering task known as "topic tracking" in which between 1 and 4 news stories formed an example-based query, and the system's goal was to find all future stories addressing the same event (or its closely connected consequences). TDT included English text and speech from the outset, with Chinese text and speech added in 1999, and Arabic text and speech added in 2002 [5]. It thus became possible to use TDT collections for cross-language speech retrieval, but only using by-example queries in which one or more examples implicitly specify the content that is sought.

The first cross-language speech retrieval test collections to use more traditional Web-like ("ad hoc") queries were created for the Cross Language Evaluation Forum (CLEF) Cross-Language Speech Retrieval (CL-SR) evaluations between 2005 and 2007 [14, 16, 28]. Two document collections were built, one in English (with 96 topics) and one in Czech (with 113 topics). For the English test collection, queries were available in six languages (Czech, Dutch, French, German, Spanish, and, for comparison, English). For the Czech test collection, only English and Czech queries were available. One limitation of the CLEF CL-SR test collections is that all of the experimentation with that test collection at CLEF was based on ASR transcripts (and associated metadata) that were provided by the organizers; at the time this precluded experimentation with techniques based on indexing alternative plausible transcriptions.

After a decade-long hiatus in CL-SR research, the Intelligence Advanced Research Projects Activity (IARPA) began the MATERIAL program in 2017[4] with the goal of accelerating work on cross-language retrieval of text and speech in lower-resource languages. To date, MATERIAL has produced test collections for Bulgarian, Lithuanian, Somali, Swahili and Tagalog; test collections for several more languages are expected over the next few years. Like the CLEF CL-SR test collections, these MATERIAL test collections include relevance judgments for ad hoc queries (all of which are available only in English), but in MATERIAL there are two broad types of ad hoc queries: (1) topical queries, which like the queries in most information retrieval test collections ask for content on a topic, and (2) lexical queries, which ask for content in which some translation of a specific query term was spoken.[5] Some MATERIAL queries are also formed as a Boolean conjunction of two lexical queries, or of one conceptual and one lexical query. As we show below, these different query types can benefit from different ranking functions. Unlike the CLEF CL-SR test collections, speech processing for the MATERIAL collections is done directly on the audio rather than on automatic transcriptions provided by the organizers. In this paper we report on experiments using the spoken content in two of these collections, for Swahili as

[4] Material is an acronym for Machine Translation for English Retrieval of Information in Any Language [21]

[5] In the MATERIAL program these are referred to as *conceptual* and *simple queries*, but we prefer to refer to them as topical and lexical in keeping with the way those terms are used in information retrieval and natural language processing, respectively. Some topical and lexical queries also contain additional clues (e.g., synonyms or hypernyms) to guide the interpretation of query terms, but we do not make use of these additional clues in our experiments.

a development setting, and for Tagalog as a second application of the retrieval approach that we first developed on the larger Swahili test collection.

3 Test Collections

Table 1 lists the details of the Swahili and Tagalog speech collections. For Swahili, the Validation (Val) collection is the union of the MATERIAL DEV, ANALYSIS1 and ANALYSIS2 sets. The Evaluation (Eval) collection is the union of the MATERIAL EVAL1, EVAL2 and EVAL3 sets. For Tagalog, the single test collection is the the union of the MATERIAL ANALYSIS1 and ANALYSIS2 sets. The larger EVAL sets for Tagalog have not yet been released by IARPA, and restricting ourselved to the ANALYSIS sets in the case of Tagalog allows us to additionally report results on manual transcriptions and manual translations (both of which are available only for the ANALYSIS sets). As Table 1 summarizes, these audio files were obtained from three types of sources: news broadcasts, topical broadcasts (e.g., podcasts), and conversational telephone speech.

	Swahili		Tagalog
	Val	Eval	
News Broadcast	173 (5 hours)	1,327 (48 hours)	131 (4 hours)
Topical Broadcast	157 (12 hours)	1,343 (115 hours)	130 (11 hours)
Conversational	153 (3 hours)	597 (29 hours)	54 (0.5 hours)

Table 1: Document counts (and duration) for MATERIAL speech collections.

	Swahili		Tagalog
	Val	Eval	
Lexical Queries	71	352	199
Topical Queries	17	29	67
Conjunctive Queries	38	319	121

Table 2: MATERIAL query statistics.

We use only queries that have at least one relevant document in the collection being searched. For Val we use MATERIAL Swahili query set Q1 and for Eval we use query sets Q2 and Q3. Val and Eval thus have disjoint queries and documents. For Tagalog we use MATERIAL Tagalog query sets Q1, Q2 and Q3. As Table 2 shows, there are many more lexical queries than topical queries.

4 Methods

This section introduces the specific methods that we use in our experiments.

4.1 Keyword Spotting

In the technique we refer to as Keyword Spotting (KWS), we use the posterior probability of each term to compute the expected counts [22]. For word k and document d, the expected count is:

$$\mathbb{E}(k|d) = \sum_{u \in d} \sum_{a:l(a)=k} P\left(a|O^{(u)}\right) \tag{1}$$

where a is a lattice arc, $l(a)$ is a term label associated with a, u is a segment of document audio and $O(u)$ are associated observations used by an ASR system to yield posterior probabilities.

4.2 Translation probabilities

We follow the approach of [25] for estimating the probability of meaning equivalence $p(s \leftrightarrow t)$ for a word s in one language and a word t in another language from the bidirectional translation probabilities $p(s|t)$ and $p(t|s)$. Translation probabilities in each direction are generated using Giza++ [15]. These probabilities are multiplied and then normalized to sum to one, an approach that has the effect of suppressing translations that are not well attested in one of the two directions.

4.3 Probabilistic structured queries

Following Darwish, we compute the expected term frequency in the query language (English) based on document language statistics as follows:

$$TF_j(Q_i) = \sum_{\{k|D_k \in T(Q_i)\}} [TF_j(D_k) \times P(D_k \leftrightarrow Q_i)] \tag{2}$$

where Q_i is a query term, D_k is a document term (in our case, a transcribed term in the spoken content), $TF_j(Q_i)$ is the term frequency of Q_i in document j, $T(Q_i)$ is the set of translation-equivalent English terms for document-language term Q_i, $TF_j(D_i)$ counts the number of times term i occurs in document j, and $P(D_k \leftrightarrow Q_i)$ is the translation probability of D_k given the query term Q_i [4]. Darwish also estimated the Inverse Document frequency (IDF) that is used to model query term specificity in the BM25 formula in a similar way, but the combination of small collections sizes and the presence of transcription ambiguity would make that a questionable choice for our speech retrieval task. Instead, we estimated IDF directly from a side collection in English, the New York Times corpus[6]. In practice we can get the IDF from any sufficiently large and sufficiently representative collection. So although the New York Times is perhaps less representative of language use in our task than the actual collections being searched would be, its much larger size makes it a reasonable choice in this case.

[6] https://catalog.ldc.upenn.edu/LDC2008T19

4.4 Combined approach

To glue the two pieces together, all we need to do is to modify the term frequency given in Equation (2) by using the posterior probability in Equation (1). The updated term frequency computation is:

$$TF_j(Q_i) = \sum_{\{k|D_k \in T(Q_i)\}} [\mathbb{E}(k|j) \times P(D_k \leftrightarrow Q_i)] \quad (3)$$

where $\mathbb{E}(k|j)$ is the expected count of the term computed by using the posterior probabilities as denoted in Equation (1).

5 Experiments

In this section we present our experiment design and our results.

5.1 KWS

Along with news, topical broadcast and conversational telephone speech (CTS) audio released for CLIR evaluation, the National Institute of Standards and Technology (NIST) released limited quantities of manually transcribed CTS data exclusively for training and evaluating ASR systems. We augmented this CTS training data with additional content read from scripts that had been previously created for the IARPA Babel program.[7] As Table 3 shows, a total of 96 hours of training data were available for Swahili, and a total of 194 hours for Tagalog. The use of limited amounts of CTS and read speech for training ASR systems results in relatively high word error rates on the news and topical broadcasts in the test collection, both of which were recorded at a higher sampling rate [18]. As Table 3 shows, word error rates are generally above 30%.

Training Data (hours)

	Swahili	Tagalog
Scripted (read)	14	33
Conversational	82	161

Word Error Rate

	Swahili	Tagalog
News Broadcast	28.6%	31.1%
Topical Broadcast	42.2%	38.9%
Conversational	32.8%	38.7%

Table 3: ASR Statistics.

Lattices produced by the ASR system for spoken content were indexed by converting them to a time-factored representation amenable to efficient indexing. Before that, acoustic and language model scores for each lattice arc were combined into a single arc score by scaling down the acoustic score to adjust the dynamic range mismatch. Arc scores were additionally scaled down, as that had

[7] https://www.iarpa.gov/index.php/research-programs/babel

been found beneficial for spoken term detection in the Babel program. Optimal scaling factors in both cases were determined on a portion of the Swahili Val set.

For the methods that use keyword search output, all output terms provided by the KWS system, even those with temporal overlap, were used. Low probability terms were not filtered, based upon the evidence using the Val set. However, terms longer than 20 characters were removed from the index.

5.2 Translation

The parallel text used for training Giza++ includes aligned sentences from MATERIAL "build pack" for each language, LORELEI,[8] GlobalVoices[9] and CommonCrawl[10]. Lexicons downloaded from Panlex[11] and Wiktionary[12] were additionally used as training data. The data was lowercased and cleaned to remove punctuations and diacritics. Table 4 details the statistics of the training data.

	Unit	Swahili	Tagalog
MATERIAL	Sentences	37.0k	65.9k
LORELEI	Sentences	-	32.9k
GlobalVoices	Sentences	30.3k	2.5k
CommonCrawl	Sentences	8.9k	18.2k
Panlex+Wiktionary	Words	190.1k	107.2k

Table 4: Giza++ Training Data.

5.3 CLIR

Our experiments were run using the Okapi BM25 ranking function, with the default parameter values of $b = 0.75$ and $k1 = 1.2$ [20]. We used Indri [23] to index each document collection, but the rankings are computed using PSQ with BM25 outside Indri. Our baseline ranking model uses PSQ with 1-best ASR output (PSQ+ASR). It is compared against a model that uses the KWS index in place of 1-best ASR (PSQ+KWS). For conjunctive queries, both parts of the query are scored separately and the scores are then combined using either an arithmetic, geometric or harmonic mean.

The MATERIAL test collections contain some content that is spoken entirely in a different language that were intended to measure the effect of spoken language identification, but that is outside the scope of our experiments for this paper. We therefore used test collection metadata to filter out those documents

[8] https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents
[9] https://globalvoices.org/
[10] http://commoncrawl.org/
[11] https://panlex.org/
[12] https://en.wiktionary.org/wiki/Wiktionary:Main_Page

from our result sets prior to evaluation. Additionally, in the MATERIAL test collections each query is labeled with a domain, and relevance judgments are available only for documents that are labeled with the same domain (e.g. Military, Sports, Government, Business, Law). The system used in our experiments has no domain-specific processing. This has the effect of scoring some otherwise-relevant documents as not relevant if the domain does not match. The effect is, however, consistent across systems, and thus comparisons made under this condition remain informative.

5.4 Swahili Results

This section details the effect of different retrieval methods for Swahili, using Mean Average Precision (MAP) as the evaluation measure. Note, however, that MAP is equivalent to Mean Reciprocal Rank when only one relevant document exists, as is often the case for the Val collection in which about half (64) of the 126 queries that have any relevant documents have just one.

The effect of the two retrieval methods, PSQ+ASR and PSQ+KWS, is shown by query type (for non-conjunctive queries) in Table 5. MAP for lexical queries increases significantly when using KWS over that of 1-best ASR. For topical queries, gains are apparent on the Val set, but MAP is essentially unchanged on the larger Eval set. We therefore conclude that PSQ+KWS is the preferred approach for both basic query types.

	Val		Eval	
	Lexical	Topical	Lexical	Topical
PSQ+ASR	0.367	0.146	0.157	0.145
PSQ+KWS	0.394	0.163	0.160	0.144

Table 5: MAP scores of different query types for Swahili test collection

Computing BM25 scores for each part of a query separately and then combining those scores using a geometric or a harmonic mean consistently results in higher MAP than simply treating the query as a flat bag of words (which is the arithmetic mean condition in Table 6) for both the Swahili Val and Eval sets. As that table shows, substantial improvements are observed in both the baseline PSQ+ASR condition and in the PSQ+KWS system. Moreover, the geometric mean seems to have a slight edge in the PSQ+KWS condition, which comports well with our intuition (since the geometric mean models an independence assumption between the two parts of the query). We therefore use the geometric mean for conjunctive queries in the remainder of our experiments.

Table 7 summarizes the overall improvements on the two Swahili test collections. We observe that switching from ASR to KWS yields a statistically significant improvement, and that then adding conjunction processing using the geometric mean yields a further statistically significant improvement. Moreover,

	Val			Eval		
	Arithmetic	Geometric	Harmonic	Arithmetic	Geometric	Harmonic
PSQ+ASR	0.202	0.329	0.329	0.177	0.181	0.181
PSQ+KWS	0.196	0.349	0.341	0.179	0.184	0.181

Table 6: MAP for conjunctive queries with three types of means, Swahili.

the net improvement from the combination of these two changes is substantial: 12% (relative) on the larger Eval collection, and 21% on the small Val collection.

	Swahili Val	Swahili Eval	Tagalog
PSQ+ASR	0.288	0.165	0.388
PSQ+KWS	0.303x	0.168	0.406x
PSQ+ASR+GeoMean	0.329x	0.181x	0.417x
PSQ+KWS+GeoMean	0.349xyz	0.184xyz	0.458xyz
PSQ+Manual Transcription			0.485
Manual Translation & Transcription			0.512
Manual Translation & Transcription+GeoMean			0.513

Table 7: MAP for all queries. x, y and z denote statistical significant improvements over ASR, KWS and ASR+GeoMean, respectively, using a two-tailed Wilcoxon signed rank test with $p < 0.05$

5.5 Tagalog Results

We do not have separate Val and Eval sets for Tagalog, but as Table 7 also shows, we can observe the same trends on the one relatively small Tagalog collection. Statistically significant improvements result from each change, and the net improvement from the two together is 18% (relative). We therefore conclude that the choices that we made on Swahili seem to be reasonable choices for Tagalog as well. About two-thirds (250) of the 387 Tagalog queries that have any relevant documents at all have only one relevant document. Our best Tagalog result (a MAP of 0.458) corresponds roughly to typically placing a single relevant document at rank 2 (since the Mean Reciprocal Rank for a system that always placed the first relevant document at rank 2 would be 0.5). This seems like a credible performance for a lower-resource language (noting that the comparable value for our similarly-sized Swahili Val set equates to roughly rank 3, which is also potentially good enough to be useful in practical applications).

For the Tagalog test collection we also have manual 1-best transcription and manual 1-best translation available. As Table 7 shows, using these 1-best manual processes yields a MAP of 0.513 for Tagalog, which is only 12% (relative) above our best present Tagalog result. While we note that 1-best transcription and translation are not an upper bound on what systems with good modeling

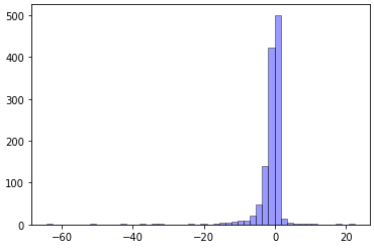

Fig. 1: Difference between expected and actual term count.

of translation ambiguity could achieve, we find this small gap to be further confirmation that our Tagalog system is yielding credible results. As Figure 1 shows, one possible source of this difference is that the expected term count more often underestimates than overestimates the correct term count (as measured on the one-best translation). Averaging over all terms in the collection, the mean absolute error of the expected counts is 1.727.

6 Conclusion and Future Work

We have presented an approach to mitigate some of the errors that arise from cascading of ASR and MT systems. First, we have shown that using word lattices from ASR to generate multiple hypothesis can be useful for cross-language speech retrieval, even in lower-resource languages. We have also shown that further substantial improvements can be obtained using specialized handing for conjunctive queries in the MATERIAL test collection. We have shown that these techniques are synergistic, each contributing to statistically significant improvements on two lower-resource languages.

A productive direction for future work would be to replicate the results for additional MATERIAL test collections. Another possible direction would be to explore the impact of using a retrieval model that is specifically designed for the term-presence condition that lexical queries seek to find, such as the approach described in [30]. A third possibility would be to explore whether query and document expansion techniques can yield further improvements when used together with the methods in this paper. While much remains to be done, it does seem that with these new test collections we can expect to see a renaissance of cross-language speech retrieval research.

References

1. Can, D., Saraclar, M.: Lattice indexing for spoken term detection. IEEE Transactions on Audio, Speech, and Language Processing **19**(8), 2338–2347 (2011)
2. Chelba, C., et al.: Retrieval and browsing of spoken content. IEEE Signal Processing Magazine **25**(3) (2008)

3. Chen, G., et al.: Using proxies for OOV keywords in the keyword search task. In: ASRU. pp. 416–421 (2013)
4. Darwish, K., Oard, D.: Probabilistic structured query methods. In: SIGIR. pp. 338–344 (2003)
5. Fiscus, J., Doddington, G.: Topic detection and tracking evaluation overview. In: Topic Detection and Tracking, pp. 17–31. Springer (2002)
6. Hull, D.: Using structured queries for disambiguation in cross-language information retrieval. In: AAAI Symp. on Cross-Language Text and Speech Retrieval (1997)
7. Karakos, D., et al.: Score normalization and system combination for improved keyword spotting. In: ASRU. pp. 210–215 (2013)
8. Kim, S., et al.: Combining lexical and statistical translation evidence for cross-language information retrieval. JASIST **66**(1), 23–39 (2015)
9. Lee, L.s., Chen, B.: Spoken document understanding and organization. IEEE Signal Processing Magazine **22**(5), 42–60 (2005)
10. Lee, L.s., Pan, Y.c.: Voice-based information retrievalhow far are we from the text-based information retrieval? In: ASRU. pp. 26–43 (2009)
11. Makhoul, J., et al.: Speech and language technologies for audio indexing and retrieval. Proceedings of the IEEE **88**(8), 1338–1353 (2000)
12. Mamou, J., et al.: Developing keyword search under the IARPA Babel program. In: Afeka Speech Processing Conference (2013)
13. McNamee, P., Mayfield, J.: Comparing cross-language query expansion techniques by degrading translation resources. In: SIGIR. pp. 159–166 (2002)
14. Oard, D., et al.: Overview of the CLEF-2006 cross-language speech retrieval track. In: CLEF. pp. 744–758 (2006)
15. Och, F., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29**(1), 19–51 (2003)
16. Pecina, P., et al.: Overview of the CLEF-2007 cross-language speech retrieval track. In: CLEF. pp. 674–686 (2007)
17. Pirkola, A.: The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: SIGIR. pp. 55–63 (1998)
18. Ragni, A., Gales, M.: Automatic speech recognition system development in the 'wild'. In: ICSA. pp. 2217–2221 (2018)
19. Riedhammer, K., et al.: A study on LVCSR and keyword search for tagalog. In: INTERSPEECH. pp. 2529–2533 (2013)
20. Robertson, S.: Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track. In: TREC (1998)
21. Rubino, C.: IARPA MATERIAL program (2016), `https://www.iarpa.gov/index.php/research-programs/material/material-baa`
22. Saraclar, M., Sproat, R.: Lattice-based search for spoken utterance retrieval. In: NAACL) (2004)
23. Strohman, T., et al.: Indri: A language model-based search engine for complex queries. In: International Conference on Intelligence Analysis (2005)
24. Tur, G., De Mori, R.: Spoken language understanding: Systems for extracting semantic information from speech. John Wiley & Sons (2011)
25. Wang, J., Oard, D.: Matching meaning for cross-language information retrieval. Information processing & management **48**(4), 631–653 (2012)
26. Wegmann, S., et al.: The TAO of ATWV: probing the mysteries of keyword search performance. In: ASRU. pp. 192–197 (2013)
27. Weintraub, M.: Keyword-spotting using SRI's DECIPHER large-vocabulary speech-recognition system. In: ICASSP. vol. 2, pp. 463–466 (1993)

28. White, R., et al.: Overview of the CLEF-2005 cross-language speech retrieval track. In: CLEF. pp. 744–759 (2005)
29. Xu, J., Weischedel, R.: Cross-lingual information retrieval using hidden Markov models. In: EMNLP. pp. 95–103 (2000)
30. Zbib, R., et al.: Neural-network lexical translation for cross-lingual IR from text and speech. In: SIGIR (2019)