Proceedings of the Cross-Language Search and Summarization of Text and Speech Workshop, pages 38–43
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Reformulating Information Retrieval from Speech and Text as a Detection Problem

Damianos Karakos[†], Rabih Zbib[*], William Hartmann[†], Richard Schwartz[†], John Makhoul[†]
[†]Raytheon BBN Technologies, Cambridge, MA
[*]Avature, Spain
[†]e-mail: {damianos.karakos, william.hartmann, rich.schwartz, john.makhoul}@raytheon.com
[*]rabih.zbib@avature.net

Abstract

In the IARPA MATERIAL program, information retrieval (IR) is treated as a hard detection problem; the system has to output a single global ranking over *all* queries, and apply a hard threshold on this global list to come up with all the hypothesized relevant documents. This means that how queries are ranked *relative to each other* can have a dramatic impact on performance. In this paper, we study such a performance measure, the Average Query Weighted Value (AQWV), which is a combination of miss and false alarm rates. AQWV requires that the same detection threshold is applied to all queries. Hence, detection scores of different queries should be *comparable*, and, to do that, a *score normalization* technique (commonly used in keyword spotting from speech) should be used. We describe unsupervised methods for score normalization, which are borrowed from the speech field and adapted accordingly for IR, and demonstrate that they greatly improve AQWV on the task of cross-language information retrieval (CLIR), on three low-resource languages used in MATERIAL. We also present a novel supervised score normalization approach which gives additional gains.

1. Introduction

When an information retrieval system is used as a support tool in a decision-making process, the user is mainly interested in whether the data under consideration contains (or, is relevant to) any of the queries of interest. For example, consider the case of streaming audio where actions must be made based upon a query detection. As each document is processed, a binary decision must be made about relevance for each query[1]. Clearly, when dealing with a decision operation, the most appropriate way to measure system performance (from an operational viewpoint) is to incorporate the two error sources that affect a user's experience: misses and false alarms. Minimizing a linear combination of these two errors is a very reasonable optimization objective, and it was chosen by the IARPA MATERIAL program as the main performance measure. Specifically, the AQWV measure is defined as follows:

$$AQWV = 1 - \overline{\text{pMiss}} - \beta\, \overline{\text{pFA}}. \tag{1}$$

$\overline{\text{pMiss}}$ is the average per-query miss rate and is defined as follows

$$\overline{\text{pMiss}} = \frac{1}{|Q_r|} \sum_{q \in Q_r} \frac{\text{\# misses of } q}{\text{\# refs of } q}, \tag{2}$$

where Q_r is the set of queries with references in the data (i.e., each has at least one relevant document). The number of references and the number of misses of query q is computed based on the whole document collection $\mathscr{C}$ under consideration.

$\overline{\text{pFA}}$, the average per-query false alarm rate, is defined as follows

$$\overline{\text{pFA}} = \frac{1}{|Q|} \sum_{q \in Q} \frac{\text{\# FAs of } q}{|\mathscr{C}| \text{ - \# refs of } q}, \tag{3}$$

The constant β in Equation (1) changes the relative importance of the two types of error ($\beta = 40$ in MATERIAL). Note that this measure assumes a single decision threshold, which means that all detection scores, over all queries, have to be commensurate. In this paper, we present techniques for transforming the detection scores that are generated by an IR system so that they are comparable across queries.

The paper is organized as follows: Section 2 gives a short summary of previous work on score normalization. Section 3 presents a supervised method for score normalization, adapted to IR. Section 4 describes the experimental setup and presents results on three low-resource languages used in the IARPA MATERIAL program: Somali, Swahili and Tagalog. Finally, Section 5 contains concluding remarks.

2. Related Work

AQWV is very similar to the Average Term Weighted Value (ATWV) (Fiscus et al., 2007), which was first used in the NIST 2006 Spoken Term Detection evaluation and then in the IARPA BABEL program (Bab, 2011) for keyword spotting from speech. As was argued in (Karakos et al., 2013) and elsewhere, generating commensurate detection scores is important for optimizing this performance measure. The main difference between ATWV and AQWV is in the granularity of the detections: keyword spotting tries to find all occurrences of a keyword of interest, no matter how many times it is spoken in a speech document. By contrast, the IR task we consider here is about retrieving whole documents that contain the query of interest, but without the need to pinpoint its exact location in the document. In other words, the granularity of the keyword spotting task is at the second (or fraction of second) level, while the granularity of the information retrieval task is at the document level. So, when computing the denominators in pMiss and pFA, AQWV uses number of documents, not number of occurrences or number of seconds as in ATWV. For this reason, the range of AQWV is $[-\beta, 1]$ (as opposed to $(-\infty, 1]$ for ATWV). (Wegmann et al., 2013) contains a detailed discussion of ATWV; most of the salient points also apply to

[*]While at Raytheon BBN Technologies.

[1]We are using document-level granularity in this paper, although similar techniques can be used for different granularities as well.

AQWV.

A number of *unsupervised* score normalization approaches have been developed for keyword spotting. pFA normalization was introduced in (Zhang et al., 2012) and used again in (Karakos et al., 2013). Keyword-specific thresholds (KST) (Karakos et al., 2013) is the most principled approach, as it is derived from fundamental theorems of decision theory. Sum-to-one (STO) (Wu, 2012; Mamou et al., 2013) is yet another popular approach, which was initially applied to problems in IR and later to keyword spotting. An in-depth comparison of these last two techniques appears in (Wang and Metze, 2014), and, since we use them in our experiments, we give more details about them below (KST is renamed QST for obvious reasons). A version of QST was also used more recently in (Shing et al., 2019) for CLIR as well.

Query-Specific Thresholds (QST)

This method estimates a query-specific threshold $t(q)$, assuming the un-normalized scores are *posterior probabilities* or posterior-like numbers between 0 and 1. As mentioned in Section 1, the AQWV and ATWV metrics are similar, allowing us to use the same optimality reasoning to compute query-specific thresholds $t(q)$. Decision theory tells us that the optimal threshold is where the expected cost of a false alarm and miss are equal. With some algebra, it can be shown that the "optimal" decision thresholds are given by:

$$t^*(q) = \frac{\beta\, N_{\text{true}}(q)}{|\mathscr{C}| + (\beta - 1) N_{\text{true}}(q)} \tag{4}$$

where $N_{\text{true}}(q)$ is the number of documents that are truly relevant to query q. This number is unknown, but it can be approximated by the sum of posteriors over the whole collection, i.e.,

$$N_{\text{sum}}(q) = \sum_{d \in \mathscr{C}} \mathit{score}(q, d), \tag{5}$$

where $\mathit{score}(q, d)$ is the retrieval score returned by the core IR system for query q and document d. Then, the normalized scores can either be given by a linear shift, or by the non-linear transformation mentioned in (Karakos et al., 2013)

$$\mathit{score}_{\text{qst}} = \exp\left\{ -\frac{\log(\mathit{score})}{\log(t^*(q))} \right\}, \tag{6}$$

which makes the common decision threshold for all queries equal to $1/e \approx 0.3679$. This is the decision threshold we use for computing AQWV in the QST results of Section 4.

Sum-to-One Score (STO)

This method, mentioned in (Wu, 2012; Mamou et al., 2013), performs a per-query normalization so that the normalized detections of a query *over the whole document collection* sum to one. In other words,

$$\mathit{score}_{\text{sto}} = \frac{\mathit{score}}{N_{\text{sum}}(q)}, \tag{7}$$

where $N_{\text{sum}}(q)$ is given by (5). Unlike QST, this method does not produce a decision threshold. As mentioned in (Mamou et al., 2013), the decision threshold can be determined based on performance on a tuning set. In our experiments, we estimate the decision threshold on the training set and apply it on the two other datasets (Tune/Test).

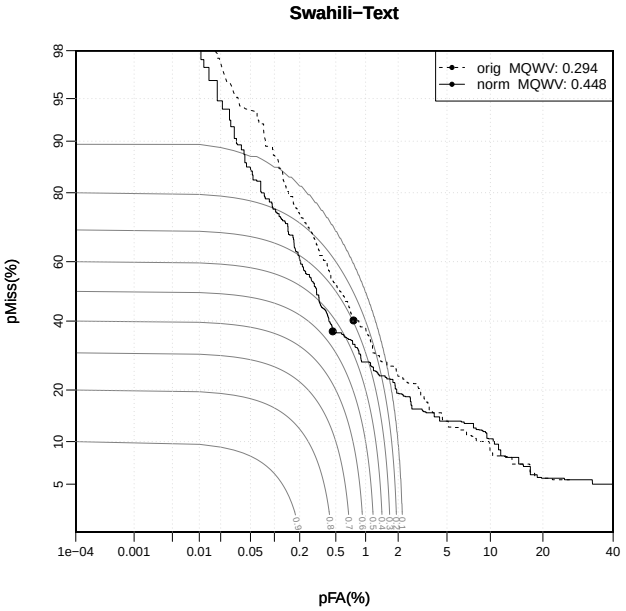

Figure 1: Comparison of the DET curves without/with score normalization. The gray lines are contours of equal AQWV.

Previous Supervised Techniques

Supervised (machine learning) techniques for score normalization focused on extracting a number of features and using them in a discriminative learning framework to directly compute the probability that a keyword is present in a specific location in the audio. For example, the authors in (Wang et al., 2009) used lattice-derived confidence scores as features in a MLP and SVM to come up with calibrated scores that significantly improved ATWV. In (Pham et al., 2014), they used features such as posterior probability, number of vowels, how many other competing arcs were present in the ASR lattice, etc., in a MLP to compute posterior-like scores, which were subsequently normalized with KST or STO. In (Lv et al., 2016), the features used were just the original posterior and KST-normalized score, but these were computed a few times, using different subword units. Finally, in (Soto et al., 2014), a large number of features (both related to posteriors in confusion networks and their transformations, as well features derived from acoustics, phonetic dictionary, etc.) was used in a SVM framework, which led to significant improvements over the unsupervised methods.

Many references related to keyword spotting and score normalization can also be found in (Tejedor et al., 2015).

Figure 1 shows a comparison of the DET curves for the un-normalized and normalized outputs of a CLIR system. There is a significant gain from normalization, especially around the range of values where the maximum AQWV

(i.e., MQWV) is attained.

3. Supervised Score Normalization

Our approach to supervised score normalization is to (i) use an optimization framework that directly optimizes the measure of interest (AQWV), and (ii) use features that are both functions of the query and the document, without making any assumptions about whether we deal with speech or text (our approach has to be able to work well with both, so, it cannot rely on the presence of speech lattices or confusion networks, in contrast to the aforementioned approaches). We generate several features—functions of the corpus, query, and the original retrieval score—and then weight them appropriately. We learn the feature weights so that, when thresholded, the combined score maximizes the performance metric.

We assume that each query-document pair (q, d) in the training data is labeled for relevance (0/1). We compute a number of features, such as the log of the following quantities:

- Original retrieval $score(q, d)$
- The QST-transformed score $score_{\text{qst}}(q, d)$
- The normalized sum $N_{\text{sum}}(q)/|\mathscr{C}|$
- The three features:

$$\min_{w \in q}\{score(w, d)\},\ \max_{w \in q}\{score(w, d)\},\ \operatorname{avg}_{w \in q}\{score(w, d)\},$$

  where avg_w is just the average over all words w in query q (esp. for multi-word queries).

- The three features:

$$\min_{w \in q}\{\text{count}(w)\},\ \max_{w \in q}\{\text{count}(w)\},\ \operatorname{avg}_{w \in q}\{\text{count}(w)\},$$

  where $\text{count}(w)$ is the count of w in the IR training data (e.g., parallel data used to train the bilingual dictionary for CLIR).

The above features $f_1, \ldots, f_F$, together with the binary labels, are fed into an optimizer that uses Powell's method (Press et al., 2007), with the goal to learn feature weights $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_F)$, as well as an optimal decision threshold t^* that maximize AQWV. At each optimization iteration, the weights are used to compute new retrieval scores

$$score_{\text{model}}(q, d) = \sum_{i=1}^{F} \alpha_i \cdot f_i$$

and new decisions

$$\text{decision}(q, d) = \mathbf{1}[score_{\text{model}}(q, d) \geq t^*].$$

During training, AQWV performance is also measured on a "tuning" set for early stopping. L2 regularization (which forces the trained weights to have small absolute values, to reduce the risk of overfitting) can also be used by changing the optimization criterion to

$$\text{AQWV}(\boldsymbol{\alpha}, t) - \lambda \cdot \text{L2}(\boldsymbol{\alpha}).$$

	Text			Audio		
	Train	Tune	Test	Train	Tune	Test
Somali	338	482	478	142	213	222
Swahili	316	449	493	155	217	207
Tagalog	291	460	-	171	244	-

Table 1: Size of various datasets (in terms of number of documents).

Note that some of the above features are dependent on various basic properties of the corpus (e.g., number of documents) and of the query set (e.g., OOV rate). In this paper, we do not study the effect of mismatched train/test conditions that may arise, for instance, when train and test corpora are significantly different. A test set that is an order of magnitude larger than the training set can cause significant mismatch in the train/test feature distributions, for the corpus-dependent features we described earlier (such as the QST-transformed score and the normalized sum). We plan to investigate such scenarios in future work.

Finally, note that, in lieu of Powell's method, we have also used a MLP framework. However, given that the data on which we train the learner is small, we did not manage to obtain results that generalized better.

4. Experimental Results

4.1. Query Sets and Retrieval Corpora

To show the benefit of normalization and thresholding to IR, we report experimental results on a Cross-language IR (CLIR) task from three different languages to English: Somali, Swahili and Tagalog. Using data from the IARPA MATERIAL program, we report on retrieval of Text and Speech documents. For each genre, we consider three data and query set conditions: (i) **Train:** A training data set D_{Train} and a training query set Q_{Train} are used for training the normalization model of Section 3 as well as decision thresholds. (ii) **Tune:** A tuning set D_{Tune} is used, together with Q_{Train}, for evaluating the stopping criterion. (iii) **Test:** Unseen data set D_{Test} and unseen query set Q_{Test} are used to assess blind performance. Statistics of these corpora appear in Table 1. As for the query set sizes, all languages have the same number of queries: Q_{Train} consists of 300 queries and Q_{Test} consists of 1000 queries.

4.2. The CLIR System

We give a brief description of the CLIR system that is used to generate the original retrieval scores. A more detailed description appears in (Zbib et al., 2019). It uses a probabilistic bilingual dictionary, trained on a set of parallel sentences and lexicons that were aligned with GIZA++ (Och and Ney, 2003). For each language pair (Somali-English, Swahili-English and Tagalog-English) the bilingual dictionary provides a translation probability $P(e|f)$ between a source word f and a target word e. Queries consist of one or more words in the target language (English), and a document is deemed relevant to a query if it contains at least one occurrence of each of the terms of the query.[2]

[2]For this program, each query consists of one or two English terms, each a word or short phrase. In some cases, there are fea-

In mathematical terms, for query q and document d, and assuming that $\mathscr{T}(d)$ is the set of all translations of all words and phrases in d, the CLIR system computes *score*(q, d) as follows:

$$\begin{aligned} & P(d \text{ is relevant to } q) \\ &= P(\text{each term } w \text{ of } q \text{ occurs at least once in } \mathscr{T}(d)) \\ &= \prod_{w \in q} P(w \text{ occurs at least once in } \mathscr{T}(d)) \\ &= \prod_{w \in q} \big(1 - P(w \text{ does not occur in } \mathscr{T}(d))\big) \\ &= \prod_{w \in q} \Big(1 - \prod_{f \in d} (1 - P(w|f))\Big) \end{aligned} \tag{8}$$

Note that (Zbib et al., 2019) performs lexical translation of source-language documents to English instead of translation of the (short) English queries to the source language; the longer context in the source documents gives a more accurate translation.

For speech documents, instead of using the translations of the 1-best output of the automatic speech recognition (ASR) system (which could be erroneous) we consider multiple ASR alternatives in the form of a *confusion network*. The latter allows us to have a probabilistic representation of the content of the foreign document, i.e., probability of occurrence $p(f|d)$ for source word f. This can be used seamlessly in (8), giving rise to a modified formula

$$P(d \text{ is relevant to } q) = \prod_{w \in q} \Big(1 - \prod_{f \in d} (1 - P(f|d) \cdot P(w|f))\Big) \tag{9}$$

Note that the occurrence probabilities of all English terms in the bilingual dictionary can be pre-computed, and accessed at retrieval time using an efficient indexing scheme.

4.3. Parallel Training Data

Parallel training data were used to estimate the probabilistic dictionaries. The data consist mostly of parallel sentences released under the IARPA MATERIAL and IARPA LORELEI (LOR, 2015) programs. A parallel lexicon downloaded automatically from Panlex (https://panlex.org/) was also included. Training data are completely disjoint from the data mentioned in Section 4.1.

4.4. ASR System Description

The amount of transcribed speech available for acoustic model training varied for each language: 48 hours for Somali, 68 hours for Swahili and 128 hours for Tagalog. For language modeling, automatically collected web data (using the techniques of (Zhang et al., 2015)) were also used. In addition to the MATERIAL data, Swahili and Tagalog also include training data from the IARPA Babel program (Bab, 2011).

tures associated with the term that constrain the sense or morphology. A document is relevant if at least one place in the foreign source could be translated to the term(s). In our experiments, the CLIR system simplifies the problem by requiring that each of the terms of the query is a possible translation of at least one foreign word in the document, ignoring any of the semantic or syntactic constraints.

Our ASR systems are trained using the Sage speech processing platform (Hsiao et al., 2016), which integrates multiple machine learning toolkits, and uses Kaldi (Povey et al., 2011) for acoustic model training. Our acoustic models are pre-trained on 1500 hours of data from 11 languages (Keith et al., 2018) and then fine-tuned to the target language. We use a CNN-LSTM acoustic model, which is similar to the TDNN-LSTM (Cheng et al., 2017), but with eight additional convolutional layers prepended to the network.

Language	Baseline	+LM Expansion	+ SST
Somali	60.6	49.4	46.1
Swahili	44.3	33.7	30.1
Tagalog	46.6	33.9	29.6

Table 2: Word error rate (WER) performance on a tuning set (known as Analysis1 in the MATERIAL program). Baseline refers to our multilingual CNN-LSTM acoustic model. LM Expansion expands the LM and lexicon using the automatically collected web data. SST further improves the acoustic model with semi-supervised training.

While word error rate (WER) is not the metric of interest, we show WER results in Table 2 to give a sense of the task difficulty. Our baseline results use our best acoustic model with the given training data, but the WER is still over 40% for each language. A major difficulty for ASR in the IARPA MATERIAL program is the mismatch between the training and test data. All training data is conversational telephone speech (CTS), while the test data is mostly broadcast data. Expanding the language model (LM) with the collected web data partially overcomes this mismatch and gives more than a 10 point absolute improvement in WER. We further reduce the mismatch through semi-supervised training using the evaluation data (approximately 70 hours). Note that this adaptation is unsupervised and is allowed by the MATERIAL program. During decoding we use standard trigram language models. We perform IR on CNets as it significantly improves performance beyond the one-best.

4.5. AQWV/MQWV Results

Table 3(a) contains AQWV results with the various normalization techniques described in the paper (the column "original" is without normalization), for the Train and Test retrieval corpora mentioned in Section 4.1.

Some observations are in order:

1. Compared to the original system scores, almost all normalization methods give gains on the text genre of all datasets. On the Test condition, the average gain (from the supervised normalization) for the text genre is 258%, while the average gain for the audio genre is 96% relative. This shows that, for measures such as AQWV (that rely on hard decisions) score normalization is of crucial importance.

2. In all cases, the supervised, model-based approach, has the best performance on the Test condition among all methods considered. Compared to the best unsupervised method, the supervised approach is 23% bet-

		Train condition				Tune condition				Test condition			
		orig	QST	STO	model	orig	QST	STO	model	orig	QST	STO	model
Somali	text	7.1	16.9	15.7	**22.6**	8.4	19.9	16.2	**22.3**	-2.9	14.0	13.4	**14.6**
	audio	3.9	-1.5	2.2	**9.9**	-2.9	-2.4	-0.9	**4.5**	-0.4	5.2	2.3	**10.3**
Swahili	text	29.4	39.6	34.7	**44.8**	20.9	38.1	30.8	**38.8**	16.5	33.0	33.8	**34.1**
	audio	29.6	21.1	17.0	**33.0**	20.7	19.4	16.1	**31.1**	20.0	17.9	13.8	**28.2**
Tagalog	text	45.7	53.5	48.9	**59.4**	49.8	52.3	47.0	**60.2**	-	-	-	-
	audio	51.1	41.3	39.1	**57.9**	38.8	34.5	31.9	**46.6**	-	-	-	-

(a) AQWV results

		Train condition				Tune condition				Test condition			
		orig	QST	STO	model	orig	QST	STO	model	orig	QST	STO	model
Somali	text	7.1	20.2	15.7	**22.6**	9.3	21.4	16.9	**25.0**	0.2	**16.2**	14.7	15.5
	audio	3.9	8.9	2.2	**9.9**	0.0	3.9	2.9	**4.7**	0.8	**13.1**	**13.1**	11.9
Swahili	text	29.4	40.4	34.7	**44.8**	21.8	38.9	35.0	**39.5**	18.1	33.5	34.1	**35.7**
	audio	29.6	28.3	17.0	**33.0**	21.4	28.6	19.7	**31.8**	21.4	25.6	14.4	**30.0**
Tagalog	text	45.7	54.8	48.9	**59.4**	51.7	55.2	50.6	**60.3**	-	-	-	-
	audio	51.1	55.1	39.1	**57.9**	43.9	43.7	43.2	**48.8**	-	-	-	-

(b) MQWV results

Table 3: (a) AQWV results on two genres of three languages (rows) and three conditions. The best result per dataset is shown in **bold**. (b) Corresponding MQWV results using the oracle decision threshold per condition.

ter (relative) on average over all languages and genres on the Test condition.

3. QST is substantially better than STO in all cases. This is expected, given that QST is designed specifically for AQWV.

Note that, for the Tune and Test conditions, the results of Table 3(a) were obtained with a decision threshold that was optimal on the Train condition. This, of course, can be suboptimal. For example, the AQWV of the original (un-normalized) system for the Somali-text Test condition is negative because the tuned acceptance threshold is too low, which makes the false alarm rate too high (a decision threshold that does not accept anything gives an AQWV of zero). So, to better understand the effect that score normalization has on the performance of a system and remove the error introduced by the imperfect decision threshold, we also computed an oracle AQWV value, the *maximum* AQWV (MQWV), obtained by sweeping over all possible decision thresholds in each one of the conditions presented, which we show in Table 3(b). We see that all MQWV values are now non-negative, and, as expected, greater than the AQWV counterparts of Table 3(a). The supervised method is still the best on average over all languages and conditions (it is worse than QST by 0.95% absolute on Somali Test but better than QST by 3% absolute on Swahili Test).

5. Concluding Remarks

In this paper, we looked at the problem of coming up with producing hard decisions in a CLIR system. One interesting application that we did not have the space to investigate in this paper is where the retrieval is done on-line, in a streaming fashion. Although there is no concept of a "fixed" collection in this case, one can consider a sliding window through the stream for purposes of computing various features, such as the sum of posteriors of Sections 2 and 3. We plan to investigate this problem in a future publication, as well as techniques that integrate score normalization directly into a CLIR engine (e.g., train a neural network CLIR system with the objective to optimize the ultimate measure of interest, instead of an approximate measure such as cross-entropy). Furthermore, with the right architecture, the neural network can come up with the most appropriate features for this task.

6. Acknowledgements

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. Useful discussions with other members of the Analytics and Machine Intelligence Department at Raytheon BBN Technologies are gratefully acknowledged.

7. Bibliographical References

(2011). IARPA Babel program - broad agency announcement (baa). https://www.iarpa.gov/index.php/research-programs/babel.

Cheng, G., Peddinti, V., Povey, D., Manohar, V., Khudanpur, S., and Yan, Y. (2017). An exploration of dropout with LSTMs. In *Proc. Interspeech*.

Fiscus, J. G., Ajot, J., and Garofolo, J. S. (2007). Results of the 2006 spoken term detection evaluation.

Hsiao, R., Meermeier, R., Ng, T., Huang, Z., Jordan, M., Kan, E., Alumäe, T., Silovskỳ, J., Hartmann, W., Keith, F., et al. (2016). Sage: The new bbn speech processing platform. In *Interspeech*, pages 3022–3026.

Karakos, D., Schwartz, R., Tsakalidis, S., Zhang, L., Ranjan, S., Ng, T. T., Hsiao, R., Saikumar, G., Bulyko, I., Nguyen, L., et al. (2013). Score normalization and system combination for improved keyword spotting. In *Au-*

tomatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, pages 210–215. IEEE.

Keith, F., Hartmann, W., Siu, M.-H., Ma, J., and Kimball, O. (2018). Optimizing multilingual knowledge transfer for time-delay neural networks with low-rank factorization. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4924–4928. IEEE.

(2015). DARPA LORELEI Program - broad agency announcement (baa). https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents.

Lv, Z., Cai, M., Zhang, W.-Q., and Liu, J. (2016). A novel discriminative score calibration method for keyword search. In *Interspeech*.

Mamou, J., Cui, J., Cui, X., Gales, M. J. F., Kingsbury, B., Knill, K., Mangu, L., Nolden, D., Pickeny, M., Ramabhadran, B., Schlüter, R., Sethy, A., and Woodland, P. C. (2013). Score combination and score normalization for spoken term detection. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8272–8276. IEEE.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Pham, V. T., Xu, H., Chen, N. F., Sivadas, S., Lim, B. P., Chng, E. S., and H., L. (2014). Discriminative score normalization for keyword search decision. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number EPFL-CONF-192584. IEEE Signal Processing Society.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flanery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press.

Shing, H.-C., Barrow, J., Galuščáková, P., Oard, D. W., and Resnik, P. (2019). Unsupervised system combination for set-based retrieval with expectation maximization. In Fabio Crestani, et al., editors, *CLEF-2019: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 191–197, Cham. Springer International Publishing.

Soto, V., Mangu, L., Rosenberg, A., and Hirschberg, J. (2014). A comparison of multiple methods for rescoring keyword search lists for low resource languages. In *Interspeech*.

Tejedor, J., Toledano, D. T., Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C., Cardenal, A., Echeverry-Correa, J. D., Coucheiro-Limeres, A., Olcoz, J., and Miguel, A. (2015). Spoken term detection ALBAYZIN 2014 evaluation: overview, systems, results, and discussion. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1).

Wang, Y. and Metze, F. (2014). An in-depth comparison of keyword specific thresholding and sum-to-one score normalization. In *Interspeech*.

Wang, D., King, S., Frankel, J., and Bell, P. (2009). Term-dependent confidence for out-of-vocabulary term detection. In *Interspeech*.

Wegmann, S., Faria, A., Janin, A., Riedhammer, K., and Morgan, N. (2013). The tao of atwv: Probing the mysteries of keyword search performance. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 192–197. IEEE.

Wu, S. (2012). *Data Fusion in Information Retrieval*. Springer.

Zbib, R., Zhao, L., Karakos, D., Hartmann, W., DeYoung, J., Huang, Z., Jiang, Z., Rivkin, N., Zhang, L., Schwartz, R. M., and Makhoul, J. (2019). Neural-network lexical translation for cross-lingual IR from text and speech. In Benjamin Piwowarski, et al., editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 645–654. ACM.

Zhang, B., Schwartz, R., Tsakalidis, S., Nguyen, L., and Matsoukas, S. (2012). White listing and score normalization for keyword spotting of noisy speech. In *Interspeech*.

Zhang, L., Karakos, D., Hartmann, W., Hsiao, R., Schwartz, R., and Tsakalidis, S. (2015). Enhancing low resource keyword spotting with automatically retrieved web documents. In *Interspeech*, pages 839–843.