Proceedings of the Cross-Language Search and Summarization of Text and Speech Workshop, pages 52–57
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

What Set of Documents to Present to an Analyst?

Richard Schwartz, John Makhoul, Lee Tarlin, Damianos Karakos
Raytheon BBN Technologies
Cambridge MA, USA
{rich.schwartz,john.makhoul,lee.tarlin,damianos.karakos}@raytheon.com

Abstract

We describe the human triage scenario envisioned in the Cross-Lingual Information Retrieval (CLIR) problem of the IARPA MATERIAL Program. The overall goal is to maximize the quality of the set of documents that is given to a bilingual analyst, as measured by the *AQWV* score. The initial set of source documents that are retrieved by the CLIR system is summarized in English and presented to human judges who attempt to remove the irrelevant documents (false alarms); the resulting documents are then presented to the analyst. First, we describe the *AQWV* performance measure and show that, in our experience, if the acceptance threshold of the CLIR component has been optimized to maximize *AQWV*, the loss in *AQWV* due to false alarms is relatively constant across many conditions, which also limits the possible gain that can be achieved by any post filter (such as human judgments) that removes false alarms. Second, we analyze the likely benefits for the triage operation as a function of the initial CLIR *AQWV* score and the ability of the human judges to remove false alarms without removing relevant documents. Third, we demonstrate that we can increase the benefit for human judgments by combining the human judgment scores with the original document scores returned by the automatic CLIR system.

Keywords: cross-lingual information retrieval, average query weighted value, AQWV

1. Introduction

The goal of the IARPA MATERIAL[1] Program is to search a corpus of foreign language documents and to return those documents that are relevant to an English language query in order to give those documents to a bilingual analyst. The program envisions a two-stage procedure. The first stage uses an automatic CLIR system that takes a structured English query and retrieves foreign documents that are likely to be relevant to that query.

However, there is usually a shortage of qualified bilingual analysts. So we would like to do anything we can to reduce the number of false alarms in the returned lists. The solution in the MATERIAL program is a second stage, which is a triage operation in which the system produces a short English summary for each of the returned documents, that provides the evidence for the document being relevant to the query. These summaries are shown to an English-speaking triage analyst whose job is to discard documents that they believe might be irrelevant. In fact, rather than making a binary decision, the analyst is asked to provide a judgment score from 1 to 5 reflecting how likely they think it is that the document is relevant.

In the next section, we will describe the AQWV measure and explain why this measure might be appropriate for this particular task. We compare it with the Mean Average Precision (MAP) measure that is most commonly used for measuring IR performance (Manning et al., 2008).

In section 3, we look at the maximum possible benefit that could be achieved by perfect triage judgments – judgments that discard all of the irrelevant documents without discarding any relevant documents. We show, empirically, that when the acceptance threshold for a system is optimized to maximize AQWV, the loss due to false alarms is relatively constant and fairly small (approximately 10%), across a wide range of conditions. And we also show that this is not true for the MAP measure. Of course, the Triage analysts cannot do this job perfectly, so we look at the theoretical performance that can be achieved, given that the average triage analyst has some probability of correctly rejecting an irrelevant document (*TR*) and another probability of falsely rejecting a relevant document (*FR*). We will show that the triage analyst has a very difficult task, especially if the initial performance of the automatic CLIR system is very good.

In Section 4, we examine the results of actual experiments and we measure the improvement that we get by setting a threshold on the judgment scores produced by the triage analysts. In Section 5, we consider better ways to use the triage analysts' judgments. In particular, we show that it is advantageous to combine the triage judgment score for a document with the original CLIR score before comparing with any threshold. This makes it more likely that the triage judgments can improve the quality of the documents provided to the final bilingual analyst.

2. The AQWV Measure

In some applications (such as web searches), the search engine returns a ranked list of documents and the user may look at as many documents as they need until they find the information they want. So it is particularly important that the most relevant documents are near the beginning of the list. In contrast, in the application here, we assume that the user is not just looking for a "good enough relevant document". Instead, they would like to find *all* relevant documents. But at the same time, they cannot afford to look at too many irrelevant documents. So instead of returning a ranked list of documents, the system will return a truncated list of documents and the analyst will read all of them.

To reflect this different need, the performance measure used is the Average Query Weighted Value (AQWV). For each query, we measure the recall and the false alarm performance. The *recall* = (1 – *pMiss*) is the fraction of all of the relevant documents that were included in the returned list. The false alarm rate, *pFA*, is the fraction of the non-relevant

[1] https://www.iarpa.gov/index.php/research-programs/material

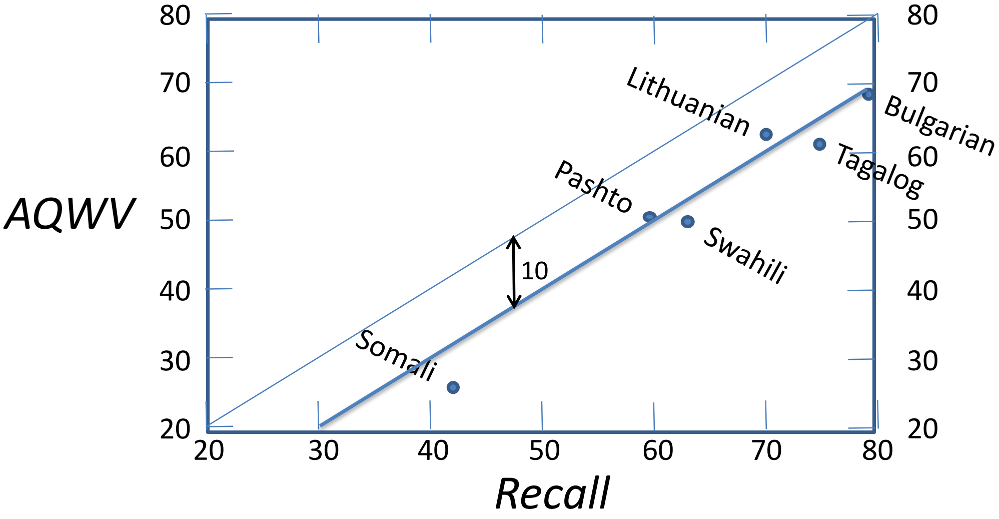

Figure 1: The AQWV vs. Recall values for 6 MATERIAL languages. The upper diagonal line represents *AQWV = Recall*. The lower diagonal line represents *AQWV = Recall – 10.* Most languages fall near the lower line.

documents in the corpus that show up in the returned list. Note that, while *pMiss* might be in the range from 20% to 80%, *pFA* is likely to be a small number, since the number of documents in the corpus is large.

The performance for a single query, or *QWV* is simply a weighted combination of these two measures:

$$QWV_q = 1 - pMiss_q - \beta \times pFA_q \quad (1)$$

$$QWV_q = Recall_q - \beta \times pFA_q \quad (2)$$

where β is a weight that reflects the relative cost of giving false alarms to the analyst and is usually $>>$ 1 because *pFA* is usually much smaller than *pMiss*. In most of our experiments, $\beta = 40$.

The overall score for a set of queries, *AQWV*, is simply the average of the *QWV* for all of the queries.

$$AQWV = Avg_q[QWV_q] \quad (3)$$

However, it is possible that some of the queries might actually have no relevant documents in the corpus being searched, so we cannot compute *Recall* for those queries. At the same time, any irrelevant documents returned (false alarms) in response to those queries are still costly. So we change the computation such that we only compute the average *Recall* on those queries that have relevant documents, while the average *pFA* is computed over all queries.

$$AQWV = Avg_{q-rel}[Recall_q] - \beta \times Avg_{all-q}[pFA_q] \quad (4)$$

The measure that is more commonly used in Information Retrieval (IR) research is the Mean Average Precision (MAP). We assume, here, that the ranked list of documents produced by a system using AQWV and MAP are the same. However, the system does not have the option of changing the number of documents returned for each query. It is a constant number, for example 100. Of course, the goal is to return as many of the relevant documents as possible within that list, but also to rank them such that the relevant documents are as close to the beginning of the list as possible. For each query, we compute the precision at the rank of each relevant document. Any document that is not in the retrieved list is given a precision of zero. Then, we average the precision values over the relevant documents. (Hence the name ”Average Precision”.) So the main difference is that with AQWV, we have the opportunity to vary the length of the list in order to reduce the number of irrelevant documents retrieved for any given query.

3. Possible Benefit for Triage Judgments

We measured the cost of the false alarms ($\beta \times pFA$) over several languages with very different performance. We also measured the benefit for different values of β. One might think that when the cost for false alarms (β) is higher, the possible benefit for triage judgments is larger. In fact, this is not the case.

If the triage judges were perfect, the *AQWV* after the triage would be equal to the Recall for that system. Figure 1 shows the *AQWV* as a function of the *Recall* for six MATERIAL languages with a wide range of *AQWV* and *Recall*. It is worth noting that the value of β was not the same for all of these languages. β was 20 for Swahili and Tagalog, and 40 for the other four languages. But still, we see that the loss for false alarms is roughly the same (actually slightly more for Swahili and Tagalog, even though the cost for each false alarm was smaller). The upper diagonal line shows *AQWV* = *Recall*. The lower diagonal line shows $AQWV = Recall$–10. As can be seen, most of the languages fall very close to the lower line, with losses due to false alarms of 8% to 13% absolute. The loss due to false alarms represents the maximum possible benefit for removing false alarms. We have made similar measurements with different values of β and the results are always the same. When β increases and the system is tuned to choose the optimal threshold, it automatically produces

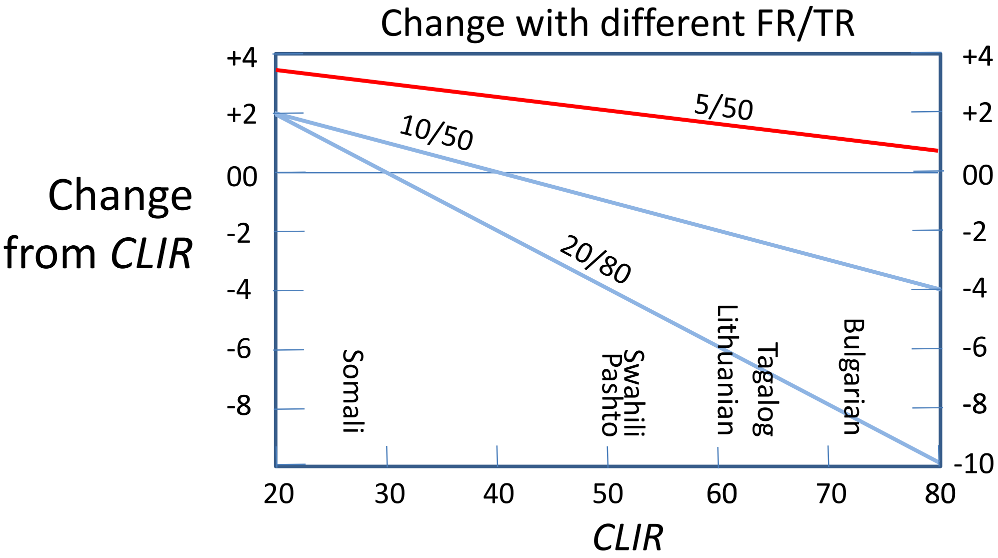

Figure 2: A plot of the expected change in *AQWV* that would accompany a Triage operation with the specified FR/TR (False Rejection / True Rejection) behavior, as a function of the initial *AQWV* produced by the CLIR system. For reference, we show the initial *CLIR* for 6 languages.

fewer false alarms and in doing so, it also decreases recall. Empirically, we find that the resulting loss for false alarms is always about the same. In the Babel program for keyword spotting we used the ATWV measure (Karakos et al., 2013; Alumäe et al., 2017), which is analogous to the AQWV measure. We found this same result for 26 different languages. So it seems to be an empirical property of the measure.

It might seem surprising that maximum possible cost of the false alarms is both relatively constant and also fairly small. This is not typically true with other measures, like MAP. The reason is that, with MAP, the system does not have the opportunity (or any incentive) to reduce the number of false alarms by reducing the number of documents retrieved. If it did reduce the returned documents, the only possible effect would be to replace the precision for some of the retrieved documents with a precision of zero, which is always worse.

Let us consider the case of a representative ranked list. Typically, the ranked list has more relevant documents near the head of the list and the relevant documents are more sparse as we go down the list. Let us consider a query with 10 relevant documents and assume that the relevant documents occur at every power of 2. So the relevant documents are at rank 1, 2, 4, 8, ...512. Only 7 of these 10 documents would appear within the first 100 returned documents. When we compute the average precision at each of these ranks, we get a list of 10 precisions: 1, 1, 3/4, 4/8, 5/16, 6/32, 7/64, 0, 0, 0. The average of these numbers is .3859375 or 38.6%.

Let's say we had a person who could review all of the 100 retrieved documents and correctly remove all of the irrelevant documents. In this case, the precision for the 7 documents within the list would be 1, so the overall precision would be 0.7 or 70%, which is a very large improvement. But the cost for this improvement would be very large because it would require that the person review 93 false documents. The *AQWV* measure is an attempt to include the cost of that review.

But why is it that, when we optimize the threshold or the number of retrieved documents, the cost of the remaining false alarms is always around 10%? There is certainly no proof that this must be the case, because it depends on the distribution of the relevant documents. But let us consider a distribution of relevant documents similar to the one described above. That is, we assume that at any given rank, the number of relevant documents within that rank, R is $\log_2(R) + 1$. So at rank 8, we would have 4 relevant documents, just as in the example above.

In Table 1 below, we show the *AQWV* as a function of the number of documents retained (in the left column) and the value of *Beta*. The second column shows the expected recall for each number of retrieved documents, which is just the number of retrieved documents divided by 10. We assume there are 10,000 documents in the entire corpus. For each number of retrieved documents and value of *Beta*, we give the value of *AQWV*. The optimal *AQWV* (in this quantized table) and any value within 0.004 of this best value is shown in **bold**. For *Beta=10*, the cost of false alarms is very low. So the best result shown is if we retrieve 120 to 140 documents. We see that the recall is between 79% and 81% and the *AQWV* is 68% - about 11% to 13% worse. When *Beta* increases, the best *AQWV* is achieved with fewer retrieved documents, because the cost of false alarms is not worth the sparse relevant documents with larger lists. As can be seen, in each case, the difference between the optimal *AQWV* and the recall at that same list size is between 0.1 and 0.13, or 10% to 13%. We suspect that this will be the case for most functions where the relevant documents become more sparse as we go further down the list. Of course, for any single query, this may not be the case, but when we average over many queries it will always tend to be true.

From our empirical results with different languages and

List-Size	Recall	Beta									
L	$\log_2(L)+1$	10	20	30	40	50	60	70	80	90	100
10	0.432	0.427	0.421	0.415	0.410	0.404	0.398	0.392	0.387	0.381	0.375
15	0.491	0.481	0.471	0.460	0.450	0.440	0.430	0.420	0.410	**0.400**	**0.390**
20	0.532	0.518	0.503	0.488	0.474	0.459	**0.444**	**0.429**	**0.415**	**0.400**	0.385
25	0.564	0.545	0.526	0.506	0.487	**0.468**	**0.448**	**0.429**	0.410	0.390	0.371
30	0.591	0.567	0.543	0.518	**0.494**	**0.470**	0.446	0.422	0.398	0.374	0.350
40	0.632	0.599	0.565	**0.531**	**0.498**	0.464	0.430	0.396	0.363	0.329	0.295
50	0.664	0.621	0.578	**0.534**	0.491	0.448	0.404	0.361	0.318	0.274	0.231
60	0.691	0.638	**0.585**	0.531	0.478	0.425	0.372	0.319	0.266	0.213	0.160
70	0.713	0.650	**0.587**	0.524	0.461	0.399	0.336	0.273	0.210	0.147	0.084
80	0.732	0.660	**0.587**	0.514	0.442	0.369	0.296	0.223	0.151	0.078	0.005
90	0.749	0.667	0.584	0.502	0.419	0.337	0.254	0.172	0.089	0.007	-0.076
100	0.764	0.672	0.580	0.487	0.395	0.303	0.210	0.118	0.026	-0.067	-0.159
110	0.778	0.676	0.574	0.472	0.369	0.267	0.165	0.063	-0.040	-0.142	-0.244
120	0.791	**0.679**	0.567	0.454	0.342	0.230	0.118	0.006	-0.106	-0.218	-0.330
130	0.802	**0.680**	0.558	0.436	0.314	0.192	0.070	-0.052	-0.174	-0.296	-0.418
140	0.813	**0.681**	0.549	0.417	0.285	0.154	0.022	-0.110	-0.242	-0.374	-0.506

Table 1: AQWV scores as a function of list size and Beta value for a corpus of 10,000 documents. The optimal value of AQWV in each column is in bold. The difference between this value and the recall in the second column is usually between 0.1 and 0.13.

conditions, we believe that the maximum we can benefit from removing irrelevant documents is approximately 10% absolute. But of course, real triage judgments will not achieve this benefit because there will be some false rejection of relevant documents and false acceptance of irrelevant documents. Below, we derive the benefit that can be achieved for a system as a function of the initial *AQWV*.
First, we define the cost of false alarms, *cFA*. We denote *CLIR* as a shorthand for the *AQWV* that results from the CLIR system.

$$cFA = \beta \times pFA \tag{5}$$

$$CLIR = Recall - cFA \tag{6}$$

$$Recall = CLIR + cFA \tag{7}$$

Now after rejecting some documents through Triage judgments, we can define the percentage of true rejections, TR, and the percentage of false rejections, FR. Define *Triage* as the *AQWV* that results after removing those documents. So by correctly removing false alarms, *Triage* will go up by $TR \times cFA$. On the other hand, but removing relevant documents, *Triage* will go down by $FR \times Recall$. So the resulting Triage score will be

$$Triage = CLIR + TR \times cFA - FR \times Recall \tag{8}$$

And substituting *Recall* from the preceding equation, the change in *AQWV* from the Triage process will be

$$\begin{aligned} Change &= Triage - CLIR \\ &= TR \times cFA - FR \times (CLIR + cFA) \end{aligned}$$

We can plot $Change$ as a function of the original CLIR score for Triage systems with different FR/TR behavior. In the Figure 2, we assume that *cFA* = 10%, because this is the typical behavior.
For example, a good Triage system (good summaries and good judges) might result in only 10% FR, together with 50% TR. That is, the triage analyst removes half of the false alarms, at a cost of losing only 10% of the relevant documents returned by the CLIR. As can be seen in the figure, as the initial *AQWV* increases, the change in *AQWV* decreases and is usually negative rather than positive. There is only a small predicted gain of about 1% absolute for the lowest initial *AQWV* (on Somali). For the other languages, there are substantial losses rather than the gain hoped for. A different summarization system and set of triage judges might have a different operating point, where they are able to correctly reject 80% of the irrelevant documents, but at a cost of falsely rejecting 20% of the relevant documents. While one might predict that this system might have similar overall performance, the line plotted for this triage system shows that the losses are much larger. This shows that, for this performance measure, the most important feature of the triage performance is that the *FR* rate must be extremely low. Finally, a third line shows what would happen if the triage analysts (together with their summaries) were able to remove 50% of the irrelevant documents, but only falsely discard 5% of the relevant documents. In this case, there is a modest gain for all of the languages. The conclusion is that it is very difficult for a triage analyst to make a significant improvement in *AQWV*.

4. Tuning the Decision Threshold

Next we look at different ways to use the judgments that result from the triage operation. The first thing we look at is the effect of the threshold on the judgment score. We performed a set of experiments using a Lithuanian corpus of text and audio documents within the MATERIAL program. The CLIR system was run on the Analysis set using the Q1 set of 300 queries. Summaries were generated and

Threshold	Text	Audio
1	64.3	53.9
2	**64.3**	**55.0**
3	72.7	53.0
4	62.2	53.2
5	56.9	51.2
Oracle	73.1	64.6

Table 2: AQWV scores on Lithuanian Analysis set using different acceptance thresholds from 1 to 5. The best results are shown in bold. The last row in the table (Oracle) gives the highest possible values for AQWV if the AMT judges made perfect judgments for this data.

were judged using Amazon Mechanical Turk (AMT). Each judgment was on a scale from 1 to 5, with 1 being clearly irrelevant and 5 being clearly relevant.

Table 2 shows the *AQWV* values for each of the five thresholds, for both Text and Audio. For each threshold, we show the result using the judgments. The result with the highest *AQWV* for each condition is shown in bold.

A threshold of 1 means all documents will be accepted, and therefore gives the *AQWV* obtained by the CLIR system. For both text and audio, we see that there is a modest gain for text and a larger gain for audio data. Using thresholds greater than 2 gives worse results than the original *CLIR* (threshold 1).

For reference, we also show in the last row of Table 2 (labeled 'Oracle') the *AQWV* that we would get if the AMT judges made perfect judgments, i.e., if they judged all relevant documents as relevant and all nonrelevant documents as nonrelevant. Note that these Oracle *AQWV* values are 9-11 points higher than the original *CLIR* values. So, this is the maximum possible gain achievable from perfect summaries and judges. By finding the threshold that maximizes *AQWV* in Table 2, we have narrowed that gap a little. Of course, a different system might have a different optimal threshold. So the optimal threshold for a system must be determined empirically.

We shall see below that the gap can be narrowed further by including the CLIR score in our optimization. As can be seen in Table 2, even with the optimal threshold, the gain in *AQWV* for using the judgments is a small fraction of the upper bound. So the question is whether there is any other way to use the scores to get better results.

5. Optimizing End-to-End (E2E) Performance

In the previous section, we discussed the improvement in *AQWV* that we might get if we replace the relevance score for each document, produced by the CLIR system with the judgment score produced by the Triage analyst and used an acceptance threshold. But the CLIR relevance score also contains very useful information. We maintain that, in order to optimize E2E performance, we should make use of both CLIR and Triage scores in making the final decision.

Our proposal is to combine the CLIR relevance and Triage judgment scores (analogous to what we normally do in system combination). A simple weighted linear combination of the two scores for each document is given by:

Interpolation weight w	Text	Audio
0.0 (only AMT score)	64.3	55.0
0.3	**65.6**	57.3
0.7	65.3	**57.9**
1.0 (only CLIR score)	64.3	53.9
Oracle	73.1	64.6

Table 3: Results for combining AMT score with CLIR score (scaled linearly to 1 to 5) as a function of the interpolation weight w. Best results are shown in bold.

$$Combined_{score} = w \times CLIR_{score} + (1-w) \times Triage_{score} \quad (9)$$

where $0 \leq w \leq 1$. We then find the value of w that maximizes *AQWV* for a particular system and condition (text or audio).

Before combining the scores, we first scale all the CLIR scores (for text and audio separately) linearly to occupy the same range as the Triage scores (1-5). In this way, this simple combination mechanism above might be applied to CLIR systems with different types of scores. (One could obviously use a more complex nonlinear combination or learn the optimal combination from a small amount of labeled data. But we wanted to make the point by keeping this really simple.)

In Table 3, we show the results of an E2E experiment using the results of the same CLIR/Triage experiment for Lithuanian reported above. We sweep weight w from 0 (only Triage score) to 1 (only CLIR score). For each value of w, we find the threshold on the combined score that gives the highest value of *AQWV*. The first row in the table (weight 0) are the same values shown in Table 2 for threshold 2, and the row with weight 1.0 are the AQWV values using CLIR scores only. As can be seen from this table, it is possible to improve on overall results by combining Triage and CLIR scores. The improvement for text is 1.3 points and 2.9 points for audio over the best *AQWV* values from using the optimal thresholds for AMT scores.

By comparing the bold numbers in Table 3 with the Oracle numbers in Table 2, we see that the gap has narrowed to about 7 points.

In fairness, we should point out that the weight and the threshold were optimized on the same data on which we measure performance. In a proper procedure, we should estimate these 2 parameters on a held out tuning set. However, since we have 300 queries and 1000 returned documents, we do not believe the results would change much. As we can see in Table 2, the performance does not even change very much between weights of 0.3 and 0.7. So we do not believe these results are unrealistic.

6. Discussion

The simple experiments performed here show that, even though it is very difficult to improve on the CLIR result alone, it is possible to get some improvements if we use the scores in an appropriate way. Undoubtedly, there are better ways of combining the judgment and CLIR scores. These methods were just the simplest reasonable methods.

One reason that the maximum benefit for discarding documents is that we use the same value of β for optimizing the initial CLIR threshold and for scoring the final result after the Triage operation. If we had used a lower value of β for the first stage, thereby returning more documents from the CLIR, there would be more relevant documents and there would be a chance for a higher final AQWV score. Of course, this would come at the cost of having to judge more documents in the Triage stage.

7. Conclusion

We have examined the AQWV measure and the effect it has in a CLIR system with a human Triage component. We have shown that the nature of the measure in our system when optimized system results in a relatively small loss due to false alarms. This in turn, makes it difficult to obtain further gains by using human judgments to remove those false alarms. We showed that if human judgments are used, the scores of the judgments are most powerful if they are combined with all other scores in order to derive the most benefit.

8. Acknowledgements

The system used for all of these measurements was the FLAIR System within the MATERIAL program (Zhang et al., 2020). So we thank all of the people involved in building that system.

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.

9. Bibliographical References

Alumäe, T., Karakos, D., Hartmann, W., Hsiao, R., Zhang, L., Nguyen, L., Tsakalidis, S., and Schwartz, R. M. (2017). The 2016 BBN georgian telephone speech keyword spotting system. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 5755–5759. IEEE.

Karakos, D., Schwartz, R. M., Tsakalidis, S., Zhang, L., Ranjan, S., Ng, T., Hsiao, R., Saikumar, G., Bulyko, I., Nguyen, L., Makhoul, J., Grézl, F., Hannemann, M., Karafiát, M., Szöke, I., Veselý, K., Lamel, L., and Le, V. B. (2013). Score normalization and system combination for improved keyword spotting. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*, pages 210–215. IEEE.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Zhang, L., Karakos, D., Hartmann, W., Srivastava, M., Tarlin, L., Akodes, D., Gouda, S. K., Bathool, N., Zhao, L., Jiang, Z., Schwartz, R., and Makhoul, J. (2020). The 2019 bbn cross-lingual information retrieval system. In *Proceedings of LREC Workshop on Cross-Language Search and Summarization of Text and Speech, Marseille, France, 2020.*