Proceedings of the Cross-Language Search and Summarization of Text and Speech Workshop, pages 14–21
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

MATERIALizing Cross-Language Information Retrieval: A Snapshot

Petra Galuščáková[1], Douglas W. Oard[1], Joseph Barrow[1], Suraj Nair[1], Han-Chin Shing[1], Elena Zotkina[1], Ramy Eskander,[2] and Rui Zhang[3]
[1] University of Maryland, College Park, MD; [2] Columbia University, New York, NY; [3] Yale University, New Haven, CT
petra@umd.edu

Abstract

At about the midpoint of the IARPA MATERIAL program in October 2019, an evaluation was conducted on systems' abilities to find Lithuanian documents based on English queries. Subsequently, both the Lithuanian test collection and results from all three teams were made available for detailed analysis. This paper capitalizes on that opportunity to begin to look at what's working well at this stage of the program, and to identify some promising directions for future work.

Keywords: information, retrieval, evaluation

1. Introduction

To some extent, research on Cross-Language Information Retrieval (CLIR) has repeatedly been a casualty of its own success. Research in the 1970's focused on extending monolingual thesauri to multilingual thesauri. Although there were some issues to address involving the ways conceptual differences were reflected in different cultures (and thus in different languages), the thesaurus-based retrieval systems of the day proved to be relatively easily extended to include entry vocabulary from different languages. Thus, after publication of an ISO multilingual thesaurus standard in 1986 there was little further research left to do along those lines (Oard and Diekema, 1998). The 1990's saw the rapid development of a different paradigm for CLIR, one in which queries were expressed in natural language and the system's goal was to rank, not to select, documents. Much of the initial work focused on dictionary-based techniques and on techniques based on comparable corpora, but it was the introduction of techniques based on parallel text around the turn of the century that essentially solved the cross-language ranking problem (Nie, 2010). Of course, ranking is only useful in interactive applications if the searcher can recognize relevant documents, so success with cross-language ranking led to a continuation of ranked retrieval CLIR research in the first decade of the twenty-first century that focused on the ability of machine translation to support cross-language relevance assessment (Gonzalo and Oard, 2004). Results there were promising as well, even with the limited capabilities of the translation technology of the day, and there the research story largely ends, with attention then shifting to deployment of the technology in applications such as 2lingual[1].

The first two waves of CLIR research were driven by language resources: by thesauri in the first wave, and by CLIR test collections in the second. With the genesis of the IARPA MATERIAL program in 2016 (Rubino, 2016), we now find ourselves at the vanguard of a third wave of CLIR research, one that draws on ideas from the first two, while adding two new twists. Like the first wave, the goal of MATERIAL is not to rank but rather to choose. Like the second wave, the goal is not just to automate the process but to get the human in the loop. Two additional issues for MATERIAL are evident from its name: Machine Translation for English Retrieval of Information in Any Language. One is a broader focus on information rather than text, with both text and speech in the same test collection. The other is a focus on affordable application to any language, even those with limited language resources.

In this paper, we focus most strongly on MATERIAL's focus on choice over ranking. MATERIAL queries are not simply a bag of words, as was typical of second-wave CLIR test collections. Rather, a MATERIAL query is a logical form, specifying what should be found, and the items to be returned (text documents or speech recordings) are all and only those that are logically entailed by the query. If this were a thought experiment, it would be reminiscent of Cooper's pioneering work on logical relevance (Cooper, 1971). But it is not a thought experiment; MATERIAL's focus is on the empirical realization of that vision. Our goal in this paper is to begin to look, at one point in time, at how well that has yet been done, both with an eye towards assessing where we are, and also with an eye towards envisioning possible future directions.

The perspective that we draw on for this paper is based on the exchange of document-level results from all three MATERIAL teams for an evaluation of Lithuanian text and speech retrieval that was conducted in October, 2019[2]. As is common in information retrieval evaluation, aggregate measures for these three runs were reported soon after the runs were submitted. Our focus here, however, is not principally on aggregate measures, but on individual cases:

- What patterns are evident in what was found?
- What patterns are evident in what was not found by any team?
- What happened when there was nothing to be found?
- And, how much better can we do if we have access to different ways of finding things?

[1] https://www.2lingual.com/

[2] This data has not been released publicly.

The paper is organized as follows: first we provide a broad overview of the types of approaches used by the three participating teams, providing individual references for additional details. Sections 3 and 4 provide answers to the four questions raised above. Finally, we conclude the paper with some remarks on next steps.

2. CLIR Systems

All three teams employ complex architectures that generally combine several processing approaches. Each of the teams includes one or more automatic speech recognition (ASR) techniques, and one or more machine translation (MT) approaches, all developed specifically for the MATERIAL task. As each team uses their own data for ASR and MT training, these systems thus not only differ in the approaches used, but also in the training data. Moreover, each team creates different variants of retrieval systems, which not only differ in the applied ASR and MT, but also in their text processing (lemmatization, stemming, character normalization, etc.) and query processing (synonym and hypernym processing, phrase processing, etc.) techniques. Retrieval systems also differ in the ways they transfer the queries and documents into a shared space. Either the English queries can be translated into Lithuanian, the Lithuanian documents can be translated into English, or queries and documents can be transformed into some other shared space (e.g., using embeddings). Evidence from multiple systems can also be combined by a variety of methods. Available data sources can be combined before retrieval, evidence from different systems can be combined during the matching phase, or the documents retrieved by different systems can be combined after the matching phase. Details on the approaches used by the SARAL team are described in (Boschee et al., 2019), the approaches used by the FLAIR team are described in (Zbib et al., 2019; Zhao et al., 2019), and the approaches used by the SCRIPTS team are described in (Oard et al., 2019).

3. Experiments

3.1. Corpus Description

The IARPA MATERIAL corpus currently consists of document collections in six languages: Swahili, Tagalog, Somali, Lithuanian, Bulgarian, and Pashto. Our analysis is based on the Lithuanian collection, for which we have results from all three participating teams. Collection statistics are given in Table 1. Details of the collection and the annotation process can be found in (Zavorin et al., 2020).

Queries There are 1,000 English queries in the collection. The queries are written in the MATERIAL Query Language (MQL), which is specified using a context-free grammar. There are three basic query types: simple, conceptual, and conjunction. Simple queries (also called lexical queries) are queries with either single word or a single phrase. A simple query "requests the system to find documents that contain a translation equivalent of the query string. A translation equivalent should sound natural to a native speaker" (NIST, 2016). Simple queries can have one of three types of semantic constraints: synonym, hypernym, or event frame. Simple queries can also have morphological constraints, where the term must match morphological features of the query string (e.g., past tense on verbs; plural on nouns). One type of conceptual query (indicated by a plus sign) is similar to a TREC query, asking for documents on a topic. Another type of conceptual query is the "example of" operator, which asks for documents which provide specific examples for the query terms. Conjunctive queries require the presence of two query parts. When one of those parts is conceptual, the conjunctive query is referred to as hybrid. We count the number of queries with each feature in Table 2.

Document Genres The corpus contains both text documents and speech recordings, which can be further subdivided by the source. There are a total of 10,203 text documents and 3,297 speech recordings, each modality being broken into 3 different genres. Documents (a term used inclusively in MATERIAL to refer to both text documents and speech recordings) are thus provided in six genres (NIST, 2016):

1. News Text (Text) - newswire or reports. Formal language.
2. Topical Text (Text) - specialty articles or reports. Diverse language formality.
3. Social Media/Blogs (Text) - blogs. Language less formal/edited.
4. News Broadcasts (Speech) - formal spoken language.
5. Topical Broadcasts (Speech) - diverse language formality.
6. Conversational Speech (Speech) - generally informal spoken language.

The amounts of Topical Text and News Text documents are similar, and each is almost three times larger than the amount of Social Media/Blog content. Similarly, the amounts of News Broadcast and Topical Broadcast recordings are similar, and about two times larger than the amount of Conversational Speech.

3.2. Official Results

We refer to the three participating systems as Teams A, B, and C to preserve anonymity. A comparison of scores for each team from the October 2019 evaluation is shown in Table 3. AQWV is the official program measure (NIST, 2016). Although the program objective is set-based retrieval, documents returned by each team also have a confidence score that can be used as a basis for ranking. This enables us to compute Mean Average Precision (MAP) on the returned list of documents, although we note that different systems return different numbers of relevant documents so the MAP values may not be strictly comparable. MQWV is an AQWV variant calculated for an optimal threshold, which is in our case determined by using either an optimal confidence score cutoff (MQWV threshold) or an optimal rank cutoff (MQWV rank) that is tied across all queries. System ordering is the same for each of the four measures.

Modality	Source	Query Type (# of judgements)			Total Documents
		Lexical	Conceptual	Hybrid	
Text	Blogs	1,225	25	181	1,491
	Topical	5,755	106	1,032	4,094
	News	3,922	65	648	4,618
Speech	Conversational	90	1	8	613
	Broadcast	1,008	11	13	1,334
	Topical	1,402	8	189	1,350
Total Queries		691	26	283	

Table 1: Corpus statistics for the IARPA MATERIAL Lithuanian evaluation collection.

Query Feature	# of queries	example
simple	974	”sculpture park”
conjunction	353	cold[hyp:sickness],tea
hybrid	283	”keep balance”,”physical exercise”+
plus sign (conceptual)	249	”copper in food”+
synonym constrint	180	telescope[syn:optical instrument]
morphology constraint	134	”<won> a prize”
hypernym constraint	130	cinnamon[hyp:spice]
example of	60	EXAMPLE_OF(baggage)
event frame constraint	34	conductor[evf:music]

Table 2: Numbers of queries with different features. We consider each of several query features independently. A query such as *lobster,EXAMPLE_OF(shellfish)* would be counted as: a hybrid query, a simple query, a conceptual query, and an example of query.

3.3. Comparison by Query Feature

Because conceptual, simple, and hybrid queries can overlap, we instead look at results by individual *features*, as opposed to query *types*.

The feature with the greatest number of queries is “simple” (974/1,000 queries contain a “simple” component), whereas only 34 queries have an event frame constraint. In Figure 1, we present both MAP and AQWV per query feature. Because of the imbalance in representation of each feature, performing better across a majority of features does not necessarily imply performing the best over all queries. We see this in Figure 1(a), where Team A performs the best across nearly all features, but marginally lower than Team C on queries with “simple” features. In general, conceptual and hybrid queries are difficult for all teams (those with the “plus_sign”, “example_of”, or “hybrid” feature). Results across queries with these features are much lower than for simple queries.

Figure 1 presents both a ranking metric (MAP) and a set-based retrieval metric (AQWV), which give different insights into the systems. AQWV introduces a penalty for returning too many documents, and it thus requires both finding relevant documents and selecting a good cutoff on a per-query basis. Though MAP and AQWV behave similarly for cumulative results (Table 3) and they similarly predict the “hardness” of the query features in Figure 1, the relative ordering of the teams in terms of AQWV and MAP scores often differ (for example Team A outperforms Team C on conceptual queries (“plus_sign”) on text in terms of MAP but Team C actually does a bit better in terms of AQWV). The ordering of the teams in terms of the MAP and AQWV cumulative scores is also in line with the results achieved by both versions of the MQWV measure. Though the score cutoffs cannot be directly compared across the teams as the teams use different score normalization methods, the optimal ranks show that the optimal number of returned documents is the same for teams A and C and it is slightly smaller for team B. Identical optimal ranks for teams A and C also allow us to compare the ranking of these two systems and indicate that team B is doing slightly better in ranking of text documents.

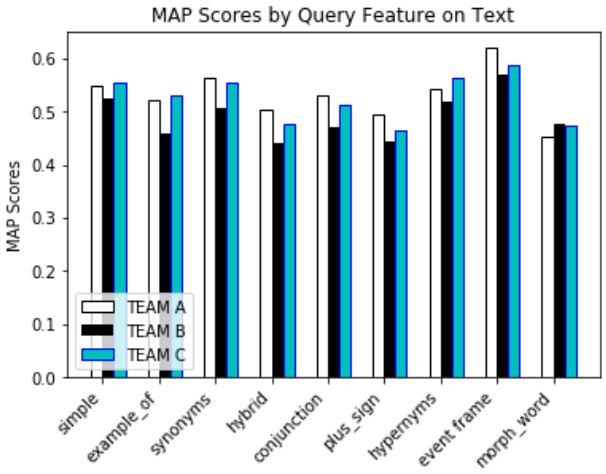
3.4. Document-Level Analysis

Breakdown by Document Types Numbers of retrieved and relevant documents broken down into the document types is in the Table 4. In general, Team A achieves a higher precision and slightly lower recall, while Team B and C achieve a higher recall and a lower precision. Importantly, these result do not translate directly to AQWV, as AQWV is an average across the queries, not across the retrieved documents.

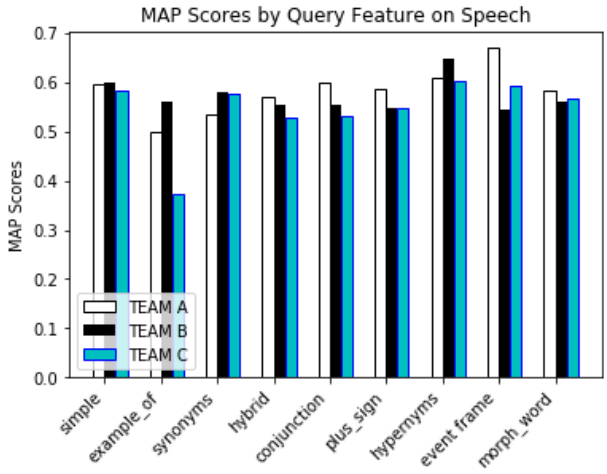
The proportion of the retrieved document types is similar across the three teams, and it differs from the proportion of the collection document types. The ratio of returned blog documents is for each team smaller than the ratio of blog text in the collection (ranging from 9 to 10%, as opposed to 15%), similarly to the ratio of news text (ranging from 33 to 36%, as opposed to 45%), while the ratio of the returned topical text is larger for each team (ranging from

	Text			Speech		
	Team A	Team B	Team C	Team A	Team B	Team C
AQWV	0.617	0.609	**0.650**	**0.609**	0.600	0.605
MQWV threshold	0.619	0.617	**0.650**	**0.616**	0.603	0.605
MQWV rank	0.622 (40)	0.572 (35)	**0.634** (40)	**0.614** (13)	0.550 (10)	0.612 (12)
MAP	0.547	0.513	**0.552**	**0.596**	0.566	0.581

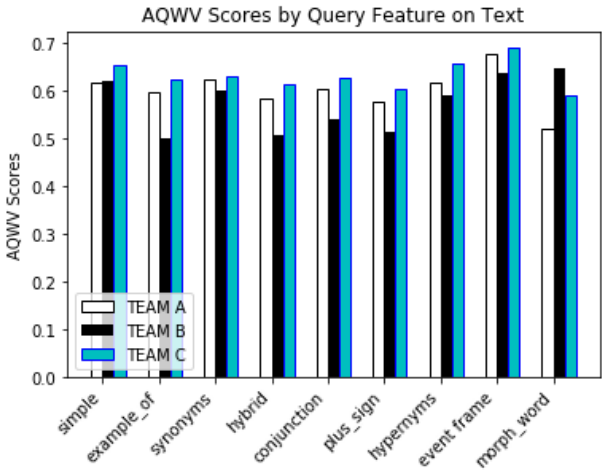
Table 3: Performance of the teams in the evaluation for text and speech. The highest value for each measure for text/speech is in a bold. For MQWV rank we also provide the optimal rank cutoff (in the parentheses).



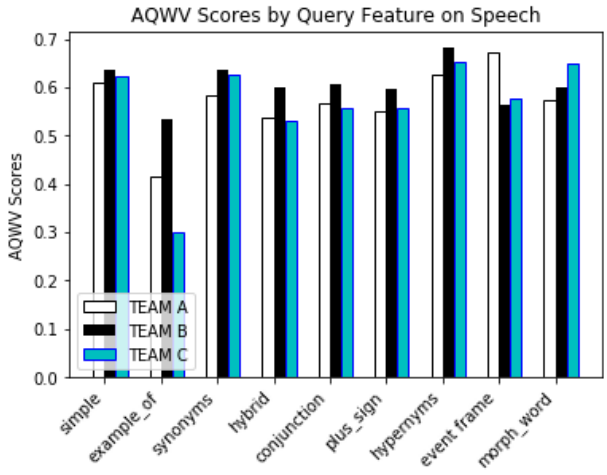
(a) MAP scores on text broken down by query feature.



(b) MAP scores on speech broken down by query feature.



(c) AQWV scores on text broken down by query feature.



(d) AQWV scores on speech broken down by query feature.

Figure 1: Dependence of the MAP and AQWV score on different query features for each team.

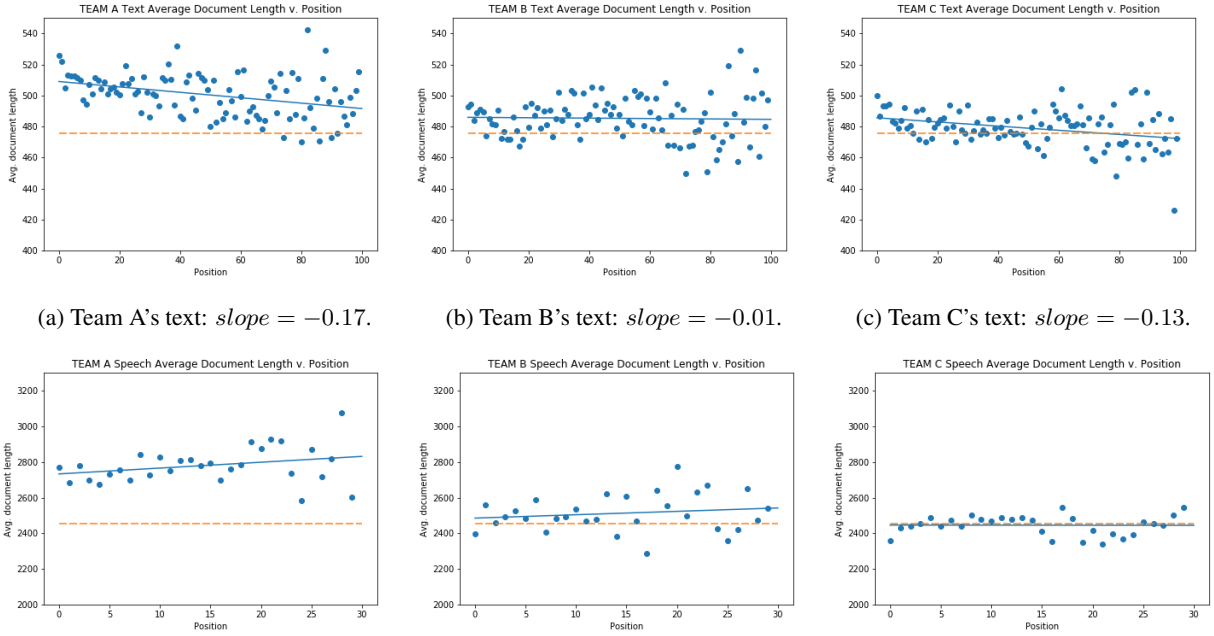
54 to 57%, as opposed to 40% in the collection). The trend is similar in speech, with conversational speech documents forming only 2 to 5% of the returned documents (the ratio of conversational speech in the collection is 19%), and topical broadcast which is for Teams A and B 61% and 57% of the returned documents respectively (compared to 41% of the documents in the collection). However, the proportion of retrieved documents corresponds well with the number of relevant documents of different types (text: 11% of blog, 53% of topical and 36% of news; speech: 4% of conversational, 38% of broadcast and 59% of topical).

Breakdown by Document Length Diverse ranking approaches utilized by different teams might lead to different biases with regard to particular length. The length of the documents retrieved at each position for each team is presented in Figure 2, together with the average length of the relevant documents. These results imply that teams A and B return documents somewhat longer than the average relevant document for both text and speech. Teams A and C show some bias towards returning longer documents first in text.

Missed Documents We investigate the number of relevant documents found and missed by each team, in rela-

		# Relevant / # Retrieved (Precision)		
		Team A	Team B	Team C
Text	Blogs	995 / 3,167 (31%)	1,080 / 4,035 (27%)	1,109 / 4,592 (24%)
	Topical	5,172 / 19,533 (27%)	5,495 / 21,519 (26%)	5,452 / 24,704 (22%)
	News	3,482 / 11,256 (31%)	3,775 / 12,962 (29%)	3,702 / 16,426 (23%)
Speech	Conversational	61 / 216 (28%)	72 / 468 (15%)	53 / 471 (11%)
	Broadcast	863 / 3,329 (26%)	854 / 3,258 (26%)	911 / 6,005 (15%)
	Topical	1,220 / 5,554 (22%)	1,177 / 4,937 (24%)	1,209 / 7,799 (16%)

Table 4: Here we break down the types of documents being returned by each team.



(a) Team A's text: $slope = -0.17$.

(b) Team B's text: $slope = -0.01$.

(c) Team C's text: $slope = -0.13$.

(d) Team A's speech: $slope = 3.28$.

(e) Team B's speech: $slope = 1.90$.

(f) Team C's speech: $slope = -0.05$.

Figure 2: For each team and modality, we plot the average document length in words returned at each position, over all 1,000 queries. For text, we look at the first 100 positions, and for speech the first 30. The solid line is a linear regression over the plot, and the dashed line is the average *relevant* document length for the modality. A negative slope implies that the team is biased towards returning longer documents at higher positions, whereas a flat slope implies an independence between position and length.

tionship to each other (Table 5). All teams found 8,255 relevant text documents (of a possible 12,959) and 1,620 speech documents (of a possible 2,900). The more interesting documents, however, are the ones that all teams missed (1,258 text documents and 332 speech documents). We additionally stratify the analysis by the number of documents that each single team found or missed that both of the other teams missed or found. These results indicate that Team A is the least diverse with respect to teams B and C, as the number of the correctly retrieved exclusively by A is the smallest and the number of relevant documents missed exclusively by A is the largest. Deeper analysis of the missed documents is described next. To complement the one-v-all breakdown in Table 5, we also provide the number of found/missed documents for each team independently (Table 6).

3.5. Failure Analysis

To identify opportunities for improvement, we examined relevant documents that no team retrieved and grouped those documents into categories that seem to us to be potentially useful for explaining those failures. Because there are more such documents than we could examine, we used two sampling strategies. In one approach, we first selected queries that all teams performed relatively poorly on by sorting based on average precision, selecting queries with the lowest values, and examining all documents that were missed by every team for such queries. We augmented this set with some random selection among documents missed by all teams for other queries in order to avoid focusing exclusively on a narrow range of queries. To investigate why a relevant document was missed, we search for the translations of query term(s) using the mapping learned from a parallel corpus. If we are unable to find it, then we manu-

		Teams B + C (Text / Speech)	
		found	missed
Team A	found	8,255 / 1,620	203 / 80
	missed	2,052 / 424	1,258 / 332
		Teams A + C (Text / Speech)	
Team B	found	8,255 / 1,620	650 / 140
	missed	1,351 / 465	1,258 / 332
		Teams A + B (Text / Speech)	
Team C	found	8,255 / 1,620	542 / 116
	missed	1,438 / 395	1,258 / 332

Table 5: Paired comparison of found and missed documents per team.

ally inspect the English translation of the relevant document obtained using our trained machine translation system. As a last resort, we inspect the Google Translate output for the relevant document. The following sections describe some of the systematic error patterns.

3.5.1. Missed translations

There exist several queries for which the relevant documents do not contain the exact query word but rather synonyms of it. For the query *diffidence*, the relevant documents contain translations of *shyness* and *modesty*, which are synonyms of the query word *diffidence*. Similarly for the query *faucet*, an unfound relevant document contains a translation of the word *tap*. Other examples are the queries *futility*, *futility,hope+*, *jello[syn:gelatin dessert]*, *ditch[syn:a trench]*, *prank[syn:a joke]* and *"Christmas ornament"*, where the system does not match them against documents whose translations contain *pointlessness*, *jelly*, *trench*, *joke* and *"Christmas toy"*, respectively. Some of the unfound translations might by found by matching the stem rather than the word. Considering the query *truck[syn:lorry]*, the relevant document contains a translation of *trucks*, which can be stemmed to find the query word *truck*. Another case is the query *EXAMPLE_OF(ground transportation),"commute to work"*, where the relevant document contains a translation of the phrase *"commuting to worker"*. A much harder example is the query *psaltery*, which never occurs in a parallel corpus that we examined, and thus might require some form of expansion to be able to identify the correct translation (kanklės, a Baltic psaltery instrument).

3.5.2. Translation ambiguity

Queries with semantic constraints (synonym, hypernym or event frame) require the system to be able to find the documents that match the correct sense of the query word. For the semantically constrained query, *bachelor[syn:unmarried man]*, the documents returned mention *bachelor's degree* instead of *unmarried man*.

3.5.3. EXAMPLE_OF queries

Systems missed some documents for conceptual queries due to incomplete expansion. The relevant document for the query *"spoiled EXAMPLE_OF(food)"* contains the translation of the phrase *"spoiled shrimp"*. The system needs to correctly expand the query to include *shrimp* as an example of food. Relevant documents for the query *"EXAMPLE_OF(natural resource) mine"* contain the translations of hyponyms of a natural resource; *gold*, *coal*, *uranium*, *lime* and *mint*. These hyponyms might be obtained by expanding the queries using external knowledge sources such as WordNet or by exploiting word embeddings.

3.5.4. Term proximity

For the query *"cause of death",contamination+*, a relevant document contains the translation of the phrase *cause of increasing human mortality*. The challenge here is to recognize *mortality* as the synonym of query word *death* and to be able to match an entire phrase that extends beyond the length of the query phrase. In this case, Sequential Dependence Model (Metzler and Croft, 2005) might be a good choice to capture long-term dependencies.

3.5.5. Morphological constraints

Queries with morphological constraint requires the machine translation systems to correctly translate the document terms preserving the root morphological aspect. For query <*squandered*>, the document is missed since the MT system incorrectly translates the relevant document term to *squandering*. In another example, the document translation produced by MT system contains *shall comfort* which does not entirely match the original query <*will comfort*>, causing the retrieval system to rank it lower.

3.5.6. Incorrect judgements

Each of the systems miss the relevant documents for the query *mistletoe,EXAMPLE_OF(bird)*. On manually inspecting the relevant documents, however, we were not able to find the translation of query word *mistletoe* in them. This might be a case of an erroneous judgement.

3.5.7. Incomplete judgements

For query *volcano*, the documents returned by the systems which are marked as non-relevant contain the word *vulkanas* (translation of volcano). However, it happens to be the name of a football team instead of a volcanic eruption. Technically, these documents should be marked relevant as there are no constraints that require the query term to match the sense of volcanic eruption.

3.6. Number of returned documents

Comparison of the numbers of retrieved documents by different teams is in Table 7. The system from Team C returns the highest number of documents on average, and returns documents for the most queries. The average variance of the number of returned documents is highest for the Team B. We additionally consider the number of queries for which the systems *correctly* returned no documents. For text, that number is low across all three teams; when a system returns no documents, that is the correct choice between 0% and 8% of the time. For speech, all teams tend not to return any documents in more cases, which corresponds well with the smaller number of relevance judgements available for the speech documents (see Table 1). The amount of correctly judged empty queries is in speech notably higher, between 56% and 67%.

		Team A	Team B	Team C	All
Relevant	found	9,649 / 2,144	10,350 / 2,103	10,263 / 2,173	8,255 / 1,620
	missed	3,310 / 756	2,609 / 797	2,696 / 727	4,704 / 1,280

Table 6: Numbers of found and missed relevant documents per team (text/speech modality).

	Text			Speech		
	Team A	Team B	Team C	Team A	Team B	Team C
Avg. # returned docs	34	39	46	9	9	14
Std. Dev. of # returned docs	38	55	31	10	14	11
Total # of queries with no returned docs	39	40	1	90	139	6
Total # of *correctly* empty queries	3	2	0	50	83	4

Table 7: Statistics of the number of documents returned by the submitted systems, broken down into text and speech.

4. System Combination

Post-retrieval combination of multiple systems often leads to improved results on both mono-lingual and cross-lingual information retrieval (Lee, 1997; Shaw and Fox, 1994; Karakos et al., 2013; Shing et al., 2019). We implement MAJORITY VOTE and COMBMNZ (Shaw and Fox, 1994) to combine the results from three teams. See Table 8 for the combination results.

		P_{miss}	P_{FA}	AQWV
Text	Single Best	0.211	0.00348	0.650
	Majority Vote	0.243	0.00183	0.684
	STO CombMNZ	0.194	0.00279	0.695
	MinMax CombMNZ	0.185	0.00277	**0.704**
Speech	Single Best	0.306	0.00211	0.609
	Majority Vote	0.251	0.00129	0.697
	STO CombMNZ	0.210	0.00244	0.693
	MinMax CombMNZ	0.199	0.00243	**0.704**

Table 8: System combination over all three teams. CombMNZ produces the best result for both text and speech. In the case of text, we attained a 5 point absolute increase (from 0.65 to 0.70), and in speech a 9 point absolute increase (from 0.61 to 0.70) over the single best system. P_{miss} and P_{FA} is the probability of misses and false alarm, respectively.

For COMBMNZ, to investigate the effect of normalization before the combination, we implement two normalization approaches: (1) MINMAX: a standard score normalization technique (Lee, 1997): $s'_m = \frac{s_m - \min S_m}{\max S_m - \min S_m}$, where s_m is the retrieved score from a system $m \in M$, set of all systems, and S_m is the set of all scores from the system m, and (2) STO: a sum-to-one normalization technique (Karakos et al., 2013), where s'_m is the original score divided by the sum of the scores for all returned document scores for a particular query (down to some fixed per-system threshold).

After normalization, CombMNZ is applied as followed:

$$CombMNZ = t \cdot \sum_{m=1}^{M} s'_m \quad (1)$$

where t is the number of times the document is retrieved across the $|M|$ systems.

After the CombMNZ combination, we apply a query-specific rank cutoff based on averaging the number of returned documents of the three teams per each query. A cutoff is essential for the system combination if we want to achieve a competitive AQWV: without the cutoff, CombMNZ will have the same AQWV as the union of the result sets over the three teams.

For both speech and text, all combination methods significantly outperform the single systems by a notable margin[3]. For text, the MINMAX COMBMNZ method outperforms all other combination methods significantly. For speech, MINMAX COMBMNZ achieves the best result, though it is not significantly better than the other combination methods.

Comparing COMBMNZ and MAJORITY VOTE, the overall difference on AQWV is relatively small. While COMBMNZ approaches are effective in reducing P_{miss}, the MAJORITY VOTE is effective in reducing P_{false_alarm}. This is in line with our intuition, as MAJORITY VOTE requires at least two teams to agree to retrieve the document, leading to a lower false alarm rate with a price of increased miss rate. COMBMNZ, on the other hand, combines a score-based combination approach ($\sum_{m=1}^{M} s'_m$) with a voting approach (t), which often leads to better ranking. This, together with a reasonable cutoff, helps to reduce the misses without raising the false alarm rate too much.

5. Conclusion and Future Work

One hallmark of the MATERIAL program is a focus on rapid system development, through the so-called "surprise language exercises". The detailed system results that we have started to analyze in this paper were released just over a week before this workshop's submission deadline, so we

[3]Statistically significant at $p < 0.05$, two-sided paired t-test.

might think of these results as having been from something of a "surprise analysis exercise". Despite the short time, we've been able to see four interesting phenomena that may help to guide future work. Perhaps most interestingly, we have identified a document length effect, with systems tending to rank longer text documents earlier, and longer speech recordings later, in a ranked list (i.e., closer to the decision threshold). We also noted that missed relevant documents tended, on average, to be shorter than correctly found relevant documents for two of the three teams. Together, these observations suggest that additional length normalization could pay off. We have also seen that mapping from query terms to document content (at least in text, the condition we were able to analyze) poses a number of systematic challenges, each of which is amenable to further research. Our analysis of system behavior in the zero-relevant case, when there are no relevant documents to be found for some queries, indicates that better modeling that condition could yield useful improvements, at least as measured by the program's target measure (AQWV). This last point is potentially of substantial interest well beyond MATERIAL because zero-relevant cases are common in many applications of search technology, and that is not a condition for which present retrieval systems are typically optimized. Finally, its been said that quantity has a quality all its own, and our results again show that to be true for system combination. Although voting is a straightforward approach to merging results from multiple set-based retrieval systems, we have found that, as would be expected, some additional gain can be achieved when confidence scores are available.

We are nowhere near exhausting the potential of this sort of analysis. For one thing, we have comparably large test collections available in two other languages, Swahili and Somali, and we might thus consider exchanging system results on such collections in the future. Such analysis might be particularly useful for Somali, which has proven to be a particularly challenging language. One limitation of our present approach, relying as it does on submitted result sets, is that it is one-sided—we can analyze confidence scores for items that were returned, but not for those that weren't. In a future study, it might prove productive to look at the other side of the decision boundary as well. There is also surely much to be learned from looking at what each individual team did relatively well at and trying to associate that with specifics of that team's system design, a question that was beyond the scope of this first analysis of ours. So we still do have miles to go before we sleep (Monteiro, 2010), but we believe that these first steps at document-scale analysis of results from multiple systems offer some useful insight into the current state of the art, and that they point the way toward future analyses of this type.

6. Acknowledgments

This research has been supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract FA8650-17-C-9117. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. This work was also supported by a gift of Google Cloud Platform research credits.

7. Bibliographical References

Boschee, E., Barry, J., Billa, J., Freedman, M., et al. (2019). SARAL: A Low-Resource Cross-Lingual Domain-Focused Information Retrieval System for Effective Rapid Document Triage. In *ACL*.

Cooper, W. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1):19 – 37.

Gonzalo, J. and Oard, D. W. (2004). iCLEF 2004 track overview: pilot experiments in interactive cross-language question answering. In *CLEF*.

Karakos, D., Schwartz, R., Tsakalidis, S., Zhang, L., Ranjan, S., et al. (2013). Score normalization and system combination for improved keyword spotting. In *ASRU*.

Lee, J. H. (1997). Analyses of multiple evidence combination. In *SIGIR*.

Metzler, D. and Croft, W. B. (2005). A Markov random field model for term dependencies. In *SIGIR*.

Monteiro, G. (2010). Life of a Poem "Stopping by Woods on a Snowy Evening". *The Robert Frost Review*.

Nie, J.-Y. (2010). Cross-language information retrieval. *Synthesis Lectures on Human Language Technologies*.

NIST. (2016). Evaluation Plan for the IARPA MATERIAL Program. https://www.nist.gov/system/files/documents/2019/10/16/material_op1_eval_plan_v0.0.9.pdf.

Oard, D. W. and Diekema, A. R. (1998). Cross-language information retrieval. *Annual Review of Information Science and Technology (ARIST)*.

Oard, D. W., Carpuat, M., Galuščáková, P., Barrow, J., Nair, S., Niu, X., Shing, H.-C., et al. (2019). Surprise Languages: Rapid-Response Cross-Language IR. In *EVIA*.

Rubino, C. (2016). IARPA MATERIAL program. https://www.iarpa.gov/index.php/research-programs/material/material-baa.

Shaw, J. and Fox, E. (1994). Combination of multiple searches. In *TREC*.

Shing, H.-C., Barrow, J., Galuščáková, P., Oard, D., and Resnik, P. (2019). Unsupervised system combination for set-based retrieval with expectation maximization. In *CLEF*.

Zavorin, I., Bills, A., Corey, C., Morrison, M., Tong, A., and Tong, R. (2020). Corpora for Cross-Language Information Retrieval in Six Less-Resourced Languages. In *LREC 2020 Workshop on Cross-Language Search and Summarization of Text and Speech*.

Zbib, R., Zhao, L., Karakos, D., Hartmann, W., DeYoung, J., , et al. (2019). Neural-network lexical translation for cross-lingual IR from text and speech. In *SIGIR*.

Zhao, L., Zbib, R., Jiang, Z., Karakos, D., and Huang, Z. (2019). Weakly Supervised Attentional Model for Low Resource Ad-hoc Cross-lingual Information Retrieval. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo)*.