Proceedings of the Cross-Language Search and Summarization of Text and Speech Workshop, pages 1–6
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

The Effect of Linguistic Parameters in Cross-Language Information Retrieval Performance
Evidence from IARPA's MATERIAL Program

Carl Rubino
Intelligence Advanced Research Projects Activity (IARPA)
Washington, DC 20511 USA
Carl.Rubino@iarpa.gov

Abstract

In IARPA's MATERIAL program, choosing languages and acquiring corpora to develop an effective End-to-End Cross-Language Information Retrieval (CLIR) system for speech and text, and component technologies thereof, was strategically planned to enable language-independent methods for CLIR development and evaluation. It was believed that a typologically diverse set of languages, coupled with a heterogeneous evaluation condition would stimulate participating research teams to construct engines that would be usable in diverse environments and responsive to changing data conditions. This paper will detail how the MATERIAL program investigated certain linguistic parameters to guide the language choice, data collection and partitioning, and better understand evaluation results.

Keywords: linguistic evaluation, cross-language information retrieval, linguistic parameters, language typology, NLP program design

1. Introduction

IARPA's Machine Translation for English Retrieval of Information in Any Language (MATERIAL) program was launched in 2017 to stimulate research on a wide array of human language technologies optimized to support cross-language information retrieval and summarization. Four multinational teams (led by Columbia University, Johns Hopkins University, Raytheon BBN and USC-ISI), chosen via competitive selection, were tasked to build End-to-End CLIR systems capable of retrieving in a fully automated way, foreign language speech and text documents responsive to a new typology of English queries, and provide evidence or relevance, in English, of the retrieved documents for human consumption (Rubino, 2017).

Prior to the 2017 kickoff of the program, nearly two years were devoted to negotiating the data collection, guided by the program's strategic evaluation methodology. This included separate training and testing conditions for both speech and text, a diverse set of languages to explore, and challenging development time frames that decreased as the program progressed.

IARPA collaborated with its Test and Evaluation (T&E) partners at the University of Maryland's Center for Advanced Study of Language (CASL), NSA's Center for Applied Machine Translation (CAMT), and MIT-Lincoln Laboratories to choose an optimal mix of diverse languages which would be incrementally released to the performing teams to stimulate and measure progress across three program periods. Two factors were most critical in initially determining the language choice: typological diversity, measured by divergent phonological, morphological and syntactic properties, and resource availability. To allow for the program's mismatch between the training and testing conditions and the requirement to identify domains without additional source language training, the languages eventually collected and annotated also had to have a substantial presence on the web. This would enable the performing teams to harvest relevant data to complement the small training sets provided by IARPA to seed the CLIR system development. Web harvesting was crucial to the program to improve the performance of applications against genres not represented in the training data, e.g. for speech, all the training data were conversational telephony, but the evaluation condition included broadcast news (Rubino, 2019). IARPA followed a strict language release schedule, not divulging the language identities until the start of each relevant development phase. This ensured that progress could be measured temporally and consistently between teams. As of May 2020, six languages were provided. Listed in order of release, these were: Tagalog (TGL), Swahili (SWA), Somali (SOM), Bulgarian (BUL), Lithuanian (LIT) and Pashto (PUS).

2. The Metrics

It was important to IARPA to evaluate the systems on a meaningful task-based measure. The primary performance measure used to assess the CLIR aspect of performer systems was a novel detection metric, related to the keyword spotting metric Actual Term Weight Value ($ATWV$) used in the IARPA Babel program (Fiscus et al., 2007). The MATERIAL metric, Actual Query Weighted Value ($AQWV$), expresses an average of all Query Values for a system operating under its actual decision threshold. This allowed for all queries to be equally treated regardless of the number of documents annotated as relevant to them in the ground truth. Query Value (QV) is defined as:

$$QV = 1 - P_{Miss} - \beta \times P_{FA} \qquad (1)$$

where P_{Miss} is the probability that a relevant document for the query will not be detected (a miss against the ground truth), and P_{FA} is the probability that a non-relevant document will be incorrectly detected (a false alarm against the ground truth). The parameter β allowed for the relative weighting of misses and false alarms. It was derived from the following formula:

$$\beta = \frac{C}{V} \times (\frac{1}{P_{Rel}} - 1) \quad (2)$$

where C is the cost of an incorrect detection, V is the value of a correct detection, and P_{Rel} is the prior probability that a document is relevant to the query. This value changed under different conditions but will remain constant for all data described herein. A perfect system that returned all relevant documents without false alarms would receive a score of 1. A system that did not return anything would receive a score of 0. If all the documents a system detected were false alarms, the score would be -β.

IARPA also provided roughly six hundred translated and transcribed documents, released as an Analysis Set, to allow the teams to measure component progress in speech recognition and machine translation (MT) using traditional metrics Word Error Rate (WER) and BLEU, respectively.

3. Linguistic Parameters Measured

Building CLIR systems capable of addressing both speech and text entails creating multiple component technologies, then learning how to optimally integrate them for information retrieval. Since a primary purpose of the MATERIAL program was to inspire novel research in both speech and translation, presumed challenges stemming from linguistic complexities and language anomalies were actively sought out by the T&E team as a means to advance research appropriately.

From a linguistic perspective, a number of parameters that could potentially affect system performance may immediately come to mind, to include both typological features of the languages such as phonetic inventory, morphological complexity, and word order, to sociolinguistic features to include dialectology, script standardization, literacy and diglossia. MATERIAL's T&E Team collected linguistic statistics on the candidate languages, focusing on features that were assumed to have a higher chance of correlation with Natural Language Processing (NLP) performance. For a sample of these kinds of linguistic variables, selected parameter values from the World Atlas of Language Structures (WALS) for the MATERIAL languages released so far are given in Table 1 with their numeric WALS Feature value (Dryer and Haspelmath, 2013). Parameters considered to be more challenging for NLP applications in the table are shown in bold.

For some linguistic features, typological resources do exist that enable us to quantify differences between or across languages. The URIEL knowledge base and its lang2vec utility, for example, provide vector identifications of languages measured from a variety of parameters taken from typological, geographical and phylogenetic databases to aid in NLP correlational analysis (Littell et al., 2017). Using lang2vec, vectors representing multiple syntactic features (often binary), manually drawn from WALS, and the Syntactic Structures of the World Languages (Collins and Kayne, 2011) can be compared across languages to compute a relative distance between any set of languages for an available amalgamation of categories. While such vector values may appear to be helpful in differentiating languages by their features, some caveats should be noted. First, no weighting mechanism is introduced to calculate the vector; all categories, regardless of their potential effect on NLP applications are treated equally. Furthermore, not all languages in the collection are represented equally for all the typological dimensions measured. Some features, in fact, were predicted from typological inference and genetic relationships. Nevertheless, we felt a conglomerate distance measure derived from a wide variety of linguistic categories was worth investigating. Table 2 exemplifies the lang2vec tool's distance calculations between English and the MATERIAL languages for four dimensions: phonological features, syntactic features, a compound value of the product of phonological and syntactic distance, and phonetic inventory.

Because Automatic Speech Recognition (ASR) was an integral part of the program, the T&E Team paid considerable attention to phonological features and phonetic inventories of the languages they chose to roll out. Multiple resources were available to capture phonetic and phonological properties, then relay them to the performing teams with each language via a document entitled "Language Specific Design Document", jointly authored by CASL and the data collector Appen Butler Hill. To contrast the specific MATERIAL languages for this paper, we counted three inventories as shown in Table 3: the number of consonants, number of vowels, and the number of segments (composed of the number of consonants, vowels and tones). These measures were extracted from the Phoible database which provides online search through an intuitive interface (Moran and McCloy, 2019). Because no single database provides complete coverage of the languages for which phonetic inventories have been documented, Phoible contains multiple databases that often conflict with each other in their counts. Where differing counts in the Phoible database were encountered, the values cited in the UCLA Phonological Segment Inventory Database took precedence, followed by the Stanford Phonology Archive.

4. The Baseline Systems

To relate the linguistic features to current program progress, we will introduce results for several baseline systems contributing to the CLIR pipeline, as well as the CLIR system itself. These rudimentary systems were produced with minimal training data, often just the program build pack and other noted, publicly available low-hanging-fruit resources. Development for the program parameters was also minimal. Table 4 reports the component technology baselines in terms of BLEU (for machine translation) and WER (for speech recognition) calculated for the MATERIAL Analysis Set. For machine translation the following baselines were reported: a phrase based statistical (PBMT) system trained on the MATERIAL Build Pack augmented

WALS Feature, #	Tagalog	Swahili	Somali	Lithuanian	Bulgarian	Pashto
Consonants, 1A	Mod Small	Mod Large	Avg	**Large**	Avg	Mod Large
Vowel Quality, 2A	Avg (5-6)	Avg (5-6)	**Large** (7-14)	Avg (5-6)	Avg (5-6)	Avg (5-6)
Syllable Structure, 12A	Mod Complex	Simple	Mod Complex	**Complex**	**Complex**	**Complex**
Uncommon Consonants, 19A	None	*th* sounds	**Pharyngeals**	None	None	None
Case, 49A	None	None	3	6-7	No	3
Word Order, 81A	VSO	SVO	SOV	SVO	SVO	SOV

Table 1: WALS Parameters for the MATERIAL languages released so far.

Language	Distance Calculations from English			
	Phon.	Syn.	Phon * Syn	Inventory
TGL	0.3433	0.66	0.226578	0.461
SWA	0.2736	0.71	0.194256	0.484
SOM	0.4816	0.62	0.298592	0.465
LIT	0.3498	0.68	0.237864	0.469
BUL	0.2804	0.48	0.134592	0.521
PUS	0.5687	0.57	0.324159	0.598

Table 2: Lang2Vec values for chosen linguistic attributes (phonological, syntactic).

Lang.	Segments	Consonants	Vowels	Syllable Structure
TGL	23	18	5	Moderately Complex
SWA	36	31	5	Simple
SOM	32	22	10	Moderately Complex
LIT	52	36	16	Complex
BUL	42	36	6	Complex
PUS	38	31	7	Complex

Table 3: Phonetic Inventories from Phoible.

with the Long Now Foundation's PanLex lexicon available at panlex.org, and three neural MT (NMT) systems trained on the MATERIAL Build Pack with PanLex (NMT), with additional engines trained on additional in-language data available from a web harvest (NMT-Mono), and a third NMT engine that also includes training data from additional, often related, languages (NMT-Multi).

Model	TGL	SWA	SOM	LIT	BUL	PUS
	MT Baselines (BLEU)					
PBMT	33.0	22.8	17.3	17.6	32.3	13.3
NMT	27.9	23.6	14.7	19.5	33.3	N/A
NMT-Mono	N/A	N/A	N/A	29.8	43.1	12.6
NMT-Multi	38.7	35.4	22.3	30.2	43.2	17.5
	Speech Recognition Baselines (WER)					
CNN-LSTM	46.6	44.3	60.6	47.9	40.0	42.8
CNN-LSTM+	33.9	33.7	49.4	23.4	21.3	39.9

Table 4: MT and ASR Baselines.

The ASR baselines reported involve a CNN Long Short-Term Memory Network (CNN-LSTM) system trained on MATERIAL Audio Build data and 1500 hours from several languages, including languages released in the Babel program, English and Arabic. The CNN-LSTM+ model cited also includes an expanded model and lexicon generated from a web text harvest and lexicon which significantly decreased the Out-of-Vocabulary (OOV) rate and improved WER scores.

The CLIR baselines detailed in Table 5 reflect the AQWV results from the MATERIAL Analysis Set, with separate numbers provided for retrieval on text vs. speech, presented as Text / Speech. For the first three languages of the program, Tagalog, Swahili and Somali, the low resource conditions were augmented with a web harvest that include Panlex and data from DARPA's LORELEI program. These additional resources were incrementally included in the CLIR systems for Lithuanian, Bulgarian, and were not present in Pashto.

5. Correlates of Performance

Because ASR systems for the MATERIAL languages were trained with multilingual features without regard to English, we initially only investigated what we considered to be potential correlations between the syntactic vectors with two program tasks that would require English language transfer: machine translation (via BLEU) and CLIR (via AQWV). We found no strong correlation between the English syntactic distance vectors and the MT task measured by BLEU (NMT $r(4) = -.09$, PBMT $r(4) = -.22$), see Figure 1, or the CLIR Task measured by AQWV (Text $r(4) = .03$, Speech $r(4) = .20$). A number of reasons can be postulated for why no correlation would exist between CLIR scores and English distance scores, such as highly diverse datasets measured for information retrieval per language, non-uniform averaged relevance probabilities for the query sets built for each language, and varying degrees of complexity between the query sets used to evaluate each language. While the number of queries released per language was relatively uniform, the composition of query types was not. More detailed descriptions of the query typology and datasets can be found in the MATERIAL Evaluation plan here: https://bit.ly/39cNGoo.

Surprisingly, when we compared MT performance to phonological distance, we found a strong negative correlation with NMT BLEU $r(4) = -.93$, $p = .008$; but not

Model	Tagalog	Swahili	Somali	Lithuanian	Bulgarian	Pashto
Baseline	-	-	-	32.0 / 14.5	41.3 / 19.4	47.3 / 38.7
+Paracrawl	-	-	-	60.5 / 22.9	64.6 / 29.9	-
+Paracrawl+Web	59.4 / 57.9	44.8 / 33.0	22.6 / 9.9	66.3 / 63.3	72.9 / 68.8	-

Table 5: CLIR Baselines in terms of AQWV (Text/Speech).

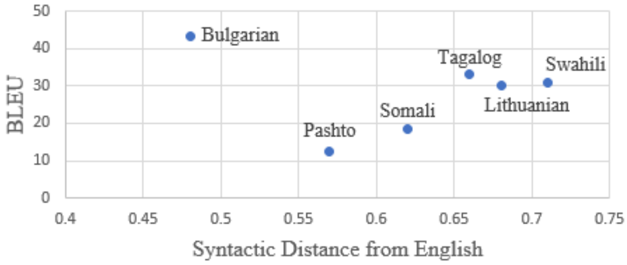

Figure 1: Syntactic Distance from English vs. BLEU

against PBMT performance where $r(4) = -.72, p = .106$. To compare MT performance with a more intuitive measure, we calculated a new compound linguistic measure, the product of syntactic and phonological distance, where the negative correlation with NMT and PBMT is more apparent and significant, $r(4) = -.95, p = .004$. See Table 2.

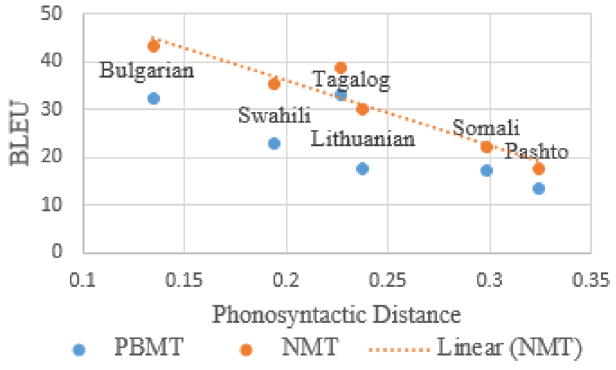

Figure 2: Phono-syntactic distance with NMT BLEU.

Not surprisingly, exploring the segment counts detailed in Table 3 to compare with a baseline CNN-LSTM monolingually trained engine yielded no evidence of correlation, $r(4) - .24$, $p = .642$ (Figure 3). Even less surprising was the observation that the Inventory Distance vector from English and ASR performance on the CNN-LSTM system were also not correlated, $r(4) = -.53$, $p = .281$. Much diversity was present in the program's speech data. The audio used for evaluation was somewhat consistent for genre distribution and sampling rates between languages but not for recording quality, or other critical factors such as the amount of data with music, dialect diversity in the collection or the number of speakers recorded.

Categorizing languages with absolute features can be intriguing theoretically, but most advantageous to the performers and the T&E team were quantifiable measures derived from program corpora. One way, for instance, of projecting possible lexical coverage problems would be to calculate OOV rates existing between development and test partitions of the IARPA released training data. Languages with higher OOV rates may presumably have lexical gaps in text and possibly, transcription anomalies in speech. Table 6 shows OOV counts calculated from the BBN team, to include both IARPA-provided corpora and their harvest.

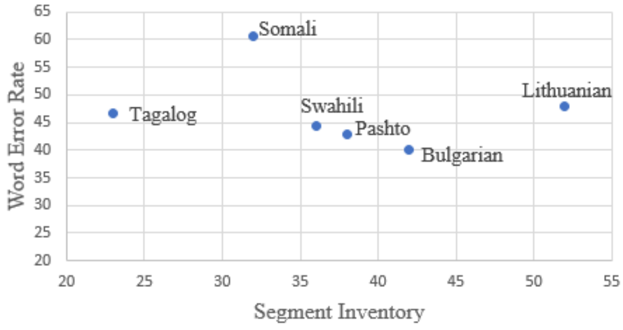

Figure 3: Segment Inventory vs. CNN-LSTM WER.

Lang.	Text		Speech	
	Parallel training (words)	% OOV	ASR Training (hours)	% OOV
TGL	1,950k	4.3	128	5.5
SWA	1,738k	5.0	68	12.7
SOM	2,278k	13.7	48	18.0
LIT	18,939k	3.7	66	2.6
BUL	25,984k	1.5	41	1.4

Table 6: OOV rates calculated by training partition.

The text OOV rates did indeed correlate with the performance of the NMT engine trained with multilingual data, perhaps as a function of the effectiveness of each language's data harvest of differing sizes to lower the OOV rates, $r(3) = -.87$, $p = .005$. Likewise, the LSTM+ ASR engine performance correlates to the OOV rates observed in speech, $r(3) = .93$, $p = .022$. See Figures 4 and 5.

For seeding machine translation development, IARPA provided training data for each language consisting of sentence-aligned bitexts from multiple news sources. To maximize diversity of the rather homogeneous collection, no more than five sentences were taken from the same article. Table 7 provides the word counts for these training corpora, along with translation ratios (foreign

Language	# Words	# Unique Words	# Translated Words	# Unique Translated Words	Unique Word Ratio	Translation Ratio
SWA	718562	55814	807766	31455	0.07767	0.88957
TGL	782525	50903	809547	30114	0.06505	0.96662
SOM	734132	73941	758337	21935	0.10072	0.96808
BUL	723042	71404	817910	35025	0.09875	0.88401
LIT	607274	91809	834541	30821	0.15118	0.72767
PUS	975595	59815	809597	28026	0.06131	1.20504

Table 7: MATERIAL MT Training Data Statistics.

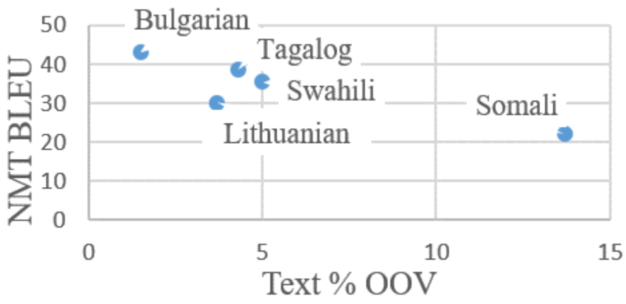

Figure 4: ASR performance as correlated to text OOV.

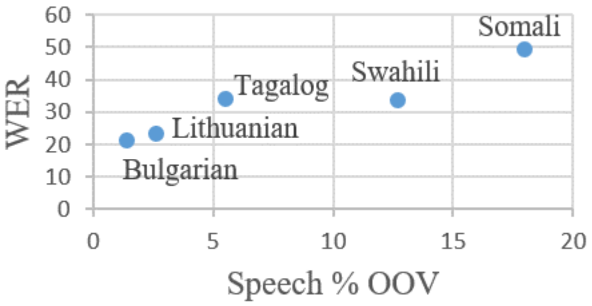

Figure 5: ASR performance as correlated to speech OOV.

words/English words) and unique word ratios (unique source words/all source words). We investigated the unique word ratio as a potential correlate for vocabulary growth. Higher ratios indicating larger vocabulary expansion may derive from a variety of factors, such as lack of orthographic standards, segmentation anomalies, or increased morphological complexity. There was a weak negative correlation between the NMT Multilingual BLEU result and the unique word ratio, $r(4) = .73$, $p = .101$.

Comparing baseline BLEU scores against the unique word ratios at the bitext size of 800K foreign language words offered slight evidence of correlation for NMT $r(4) = -.73$, $p = .101$ but not for PBMT performance, BLEU $r(4) = -.48$, $p = .339$. Likewise, no correlation was found between BLEU scores and vocabulary size in a smaller speech dataset of 80K words shown in Table 8, PBMT $r(4) = .06$, $t = .911$, NMT $r(4) = .36$, $t = .489$.

Lang.	Vocabulary size at 80K words (K words)	OOV(%) with Acoustic Build Data
TGL	11.3	13.5
SWA	13.1	14.1
SOM	12.2	15.7
BUL	13.1	13.3
LIT	19.4	21.3
PUS	7.2	6.0

Table 8: Vocabulary statistics from the Speech Build packs.

6. Conclusion

From the IARPA MATERIAL experience, choosing languages by linguistic parameters helps to ensure parametric diversity, critical to our ability to develop language-independent CLIR solutions in low resource conditions, a fundamental question posed by the program. Certain typological parameters we may assume to be tightly linked to CLIR results often have no correlation with the actual performance of the NLP applications to which the parameters would seem intuitively relevant. Discerning which linguistic parameters correlated with overall performance enabled IARPA to evaluate CLIR progress when different languages were measured. Some parameters were also a significant factor for Performing Teams to determine the most effective CLIR pipeline design, customized to handle language-specific properties deemed necessary to address. These pipelines, as well as data collection and use strategies, differed between teams and languages, the details of which are beyond the scope of this paper.

We have shown, albeit with a relatively small sample of diverse languages and only using immature baseline systems, that amalgamate typological distance vectors between the MATERIAL languages and English quite unexpectedly and counter-intuitively did correlate with MT BLEU scores, but not AQWV or WER measures.

We suggest that when choosing languages to design or evaluate an NLP research program, ample attention is paid to the language dimension as measured by the properties of the data used for both training, development and evaluation, as their correlation with performance is likely to exceed that of typological parameters presumed to be critical from a linguistic perspective.

7. Acknowledgments

Special thanks to Timothy McKinnon for introducing me to lang2vec and fruitful discussions on linguistics and evaluation. I am also grateful to Audrey Tong, Catherine Cotell, and the anonymous editors of this volume for their comments on a previous version of this paper. This work is supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA). The views and conclusions contained herein are those of the author 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.

8. Bibliographical References

Collins, C. and Kayne, R. (2011). *Syntactic Structures of the World's Languages*. New York University, New York.

Dryer, M. and Haspelmath, M. (2013). *The World Atlas of Language Structures Online*. Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology.

Fiscus, J. G., Ajot, J., Garofolo, J. S., and Doddingtion, G. (2007). Results of the 2006 spoken term detection evaluation. In *Proceedings of the ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, pages 51–55. ACM SIGIR.

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., and Levin, L. (2017). Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Steven Moran et al., editors. (2019). *PHOIBLE 2.0.* Jena: Max Planck Institute for the Science of Human History.

Rubino, C. (2017). *MATERIAL Broad Agency Announcement*. https://bit.ly/37gKhV9.

Rubino, C. (2019). IARPA's Contribution to Human Language Technology Development for Low Resource Languages. In *Language Technologies for All Conference*. UNESCO. https://bit.ly/39e2mD4.