Yale

Query-Specific Thresholding with System Combination

Neha Verma[1], Rui Zhang[2], and Dragomir Radev, PhD[2]

[1]Yale College and [2]Department of Computer Science, Yale University

LILY Lab

Introduction

Cross-Lingual Information Retrieval (CLIR) involves returning documents relevant to a given query where the language of the documents and query differ. We are specifically interested in the task of CLIR for low-resource languages.

In returning a ranked retrieval list or set of documents to a user, we are interested in determining how many documents to return, as to avoid returning an excess of irrelevant information. We wish to make this "cutoff" point specific to individual queries, as queries may differ in difficulty of retrieval. In addition to across queries, we also wish to compare retrieval scores across different systems in order to be able evaluate their outputs.

We examine the use of query-specific thresholding as a score normalization technique in comparison with sum-to-one normalization

Materials and Methods

All the data used in this project was provided by IARPA for the MATERIAL (Machine Translation for English Retrieval in Any Language) program. We consider two settings for cross-lingual information retrieval: (1) English (EN) queries with Swahili (SW) documents and (2) English queries with Somali (SO) documents. For each setting, we use three query sets, Q1, Q2, and Q3. We used the DEV and EVAL text and speech transcribed document collections for SW and SO. We are given a set of relevance judgements for each of the queries from each of their retrieval settings.

We use the Actual Query Weighted Value (AQWV) as our evaluation metric, as defined by IARPA. We examined the effect of splitting the document types between text and speech transcriptions on the MQWV. Finally, we compare the performance of query-specific thresholding for score normalization against sum-to-one (STO) normalization.

SW-text	MQWV from STO	MQWV from QST
bbn_text_fast	0.3066	0.3006
customindri	0.277	0.2898
indri_words-fj05q_tN	0.2398	0.1739
indri_words-fj05_tEDINMTN	0.2306	0.1541
indri_words-fj05_tSMTN	0.2374	0.1529
indri_words-fj05_tUMDNMTN	0.2178	0.1443
indri_words-f_tEDINMTsN	0.268	0.2051
indri_words-f_tSMTsN	0.2678	0.2004
indri_words-f_tUMDNMTsN	0.2478	0.1679
nbest_words_UMD	0.1869	0.1563
nbest_words_EDIN	0.1934	0.1825

Table 1. A comparison of MQWV results for EN-SW text retrieval systems.

SO-speech	MQWV from STO	MQWV from QST
aEDINMTN	0.0508	0.0638
aSMTN	0.0431	0.0558
aUMDNMTN	0.0467	0.0412
custom_indri_PSQ-3ef-CDF097-N	0.0566	0.0608
custom_indri_PSQ-3f-CDF097-N	0.0566	0.0608
words-f_aEDINMTsN	0.0752	0.0765
words-f_aSMTsN	0.0548	0.0613
words-f_aUMDNMTsN	0.0696	0.0678
words-fq-N_best_aN	0.036	0.0571
words-fs_aSMTdN	0.0535	0.0591
nbest_words-f_tEDINMTN+aEDINMTN_EdiNMT	0	0
nbest_words-f_tUMDNMTN+aUMDNMTN_UMDNMT	0	0

Table 3. A comparison of MQWV results for EN-SO speech retrieval systems.

SO-text	MQWV from STO	MQWV from QST
customindri	0.0842	0.0774
indri_words-fj05q_tN	0.0904	0.082
indri_words-fj05_tEDINMTN	0.1346	0.1123
indri_words-fj05_tSMTN	0.1274	0.1188
indri_words-fj05_tUMDNMTN	0.1304	0.1165
indri_words-f_tEDINMTsN	0.133	0.1161
indri_words-f_tSMTsN	0.1248	0.1087
indri_words-f_tUMDNMTsN	0.1294	0.1148
nbest_words_UMD	0.1228	0.0765
nbest_words_EDIN	0.1364	0.1005

Table 2. A comparison of MQWV results for EN-SO text retrieval systems.

SW-speech	MQWV from STO	MQWV from QST
aEDINMTN	0.1276	0.1145
aSMTN	0.1309	0.1559
aUMDNMTN	0.1356	0.1519
bbn_text_fast_PSQ-3f-CDF097-N	0.2218	0.2338
bbn_text_fast_PSQ-4f-CDF097-N	0.2387	0.2475
custom_indri_PSQ-3ef-CDF097-N	0.0714	0.0734
custom_indri_PSQ-3f-CDF097-N	0.1946	0.1938
words-f_aEDINMTsN	0.1496	0.158
words-f_aSMTsN	0.1605	0.1867
words-f_aUMDNMTsN	0.1496	0.1447
words-fq-N_best_aN	0.0818	0.1118
words-fs_aSMTdN	0.1585	0.1799
nbest_words-f_tEDINMTN+aEDINMTN_EdiNMT	0.0556	0.093
nbest_words-f_tUMDNMTN+aUMDNMTN_UMDNMT	0.0618	0.1081

Table 4. A comparison of MQWV results for EN-SW speech retrieval systems.

QST

$$\rho_q = \frac{\beta N_q}{N + (\beta - 1) N_q} \qquad \hat{N}_q = \delta \sum_{d \in D} s_{q,d}^{\gamma}$$

$$\bar{s}_{q,d} = s_{q,d}^{-1/\log(\rho_q)}$$

Query Specific Thresholding determines a parameter rho, which is calculated using an estimate of the number of documents a query appears in, N_q. N_q is estimated using tunable parameters delta and gamma. Finally, the normalized score of the document is calculated using the final equation. We use $\beta = 40$ in this setting.

Results

Results for individual systems are shown in tables 1-4. MQWV scores for individual systems are reported for retrieval on Swahili-text, Somali-text, Swahili-speech, and Somali-speech. As compared to previous results regarding sum-to-one normalization, SW-text sees improvement in 1/11 system, SO-text none, SW-speech, 11/14 systems, and SO-speech 8/13 systems.

Conclusion and Future Work

QST score normalization improves MQWV scores for a number of retrieval systems for speech documents, as compared to STO normalization. Separating document classes by speech vs. text proved useful in this exploration of QST as a normalization technique. QST+system combination on speech sets will likely outperform the previous STO+system combination AQWV score. We will explore the possibility of building a system that has different score normalization techniques based on the identity of the documents, either speech or text. Finally, we wish to investigate a supervised score normalization technique in conjunction with QST in order to potentially further improve the AQWV for these CLIR tasks.

Acknowledgement

I would like to thank Professor Dragomir Radev and Rui Zhang for their assistance and guidance in this project