Yale

Query-Specific Thresholding and Sum-to-One Normalization

Neha Verma, Rui Zhang, and Dragomir Radev

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 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 how classifying query types and changing score normalization schemes affects the maximum retrieval scores attainable across several systems.

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 first examined the effect of classifying queries has the maximum AQWV (MQWV) attainable. We considered splitting simple vs. complex queries evenly, based on features such as the presence of phrases, hypernymy, or synonymy. We also 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.

Sum-to-One Normalization and System Combination

$$scr^{\text{Adaptive-STO}}(D_i) = \frac{scr_S(D_i)^{\gamma}}{\sum_{i'=1}^{n} scr_S(D_{i'})^{\gamma}}$$

$$scr^{\text{MQWVCombMNZ}}(D) = m_D \times \sum_{j=1}^{m} w_{S_j}^{\text{MQWV}} \times scr_{S_j}^{\text{Adaptive-STO}}(D)$$

The Adaptive-STO score normalization, with parameter gamma, normalizes a document across other document scores. The MQWVCombMNZ system combination equation first penalizes zero-scores returned by systems, and then weights the STO scores by a weight proportional to the MQWV attained by the system.

Query Specific Thresholding Equations

$$\rho_q = \frac{\beta N_q}{N + (\beta - 1) N_q}$$

$$\hat{N}_q = \delta \sum_{d \in D} s_{q,d}^{\gamma}$$

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

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 $\beta = 20$ in this setting.

System (SW)	Prev MQWV	MQWV	Gamma	Delta
bbn_text_fast	0.3429	0.3947	0.5	1
customindri	0.3333	0.3700	0.6	0.9
indri_words-fj05q_tN	0.2268	0.3168	1.2	0.7
indri_words-fj05_tEDINMTN	0.21	0.2965	1.2	0.8
indri_words-fj05_tSMTN	0.1878	0.3045	1.2	0.8
indri_words-fj05_tUMDNMTN	0.2157	0.2908	1.2	0.7
indri_words-f_tEDINMTsN	0.2136	0.3575	1.5	1
indri_words-f_tSMTsN	0.1827	0.3495	1.4	1
indri_words-f_tUMDNMTsN	0.245	0.3347	1.5	1
nbest_words_UMD	0.2038	0.2737	3.2	1.8
nbest_words_EDIN	0.1692	0.2609	0.5	3.8

Table 1. Table of previous MQWVs attained from Sum-to-One normalization, the Maximum AQWV (MQWV) from Query-Specific Thresholding (QST), and the delta and gamma values for which the MQWV was attained for EN->SW Q2Q3/EVAL123 retrieval.

System (SO)	Prev MQWV	MQWV	Gamma	Delta
customindri	0.0737	0.1460	0.9	1.3
indri_words-fj05q_tN	0.0912	0.1272	0.9	2
indri_words-fj05_tEDINMTN	0.0889	0.1824	1.2	1.4
indri_words-fj05_tSMTN	0.1029	0.1771	1	1.4
indri_words-fj05_tUMDNMTN	0.102	0.1830	1.1	1.3
indri_words-f_tEDINMTsN	0.097	0.1832	1.3	1.4
indri_words-f_tSMTsN	0.1089	0.1783	1.9	1.2
indri_words-f_tUMDNMTsN	0.0819	0.1888	1.7	1.3
nbest_words_UMD	0.1061	0.1905	1.8	2.2
nbest_words_EDIN	0.1108	0.1681	0.5	3.2

Table 2. Table of previous MQWVs attained from Sum-to-One normalization, the Maximum AQWV (MQWV) from Query-Specific Thresholding (QST), and the delta and gamma values for which the MQWV was attained for EN->SO Q2Q3/EVAL123 retrieval.

Results

Firstly, we observed that while classifying queries into simple and complex categories does not improve the MQWVs attainable for complex query sets, as it does for simple query sets, we do see a possibility for overall MQWV improvement by classifying queries. We note that for each of the systems in which MQWV decreases for complex queries, the decrease in MQWV is smaller than the increase in MQWV for simple queries. Since there are an equal number of complex and simple queries, we predict the overall MQWV of the system would improve after predicting query difficulty.

Additionally, as seen in Tables 1 and 2, all MQWVs that are a result of query-specific thresholding outperform previous MQWVs attained by Adaptive-STO score normalization. In previous work, we have observed that the MQWV attained using system combination outperformed each of the MQWVs attained by individual systems, while using STO normalization. Therefore, since this new form of normalization, QST, outperforms each of the MQWVs given by STO normalization, we should expect to see an increased MQWV when QST is combined with system combination.

Conclusions and Future Work

From our results, we show that QST combined with system combination has potential to outperform our current pipeline of STO normalization plus system combination. We also show that classifying query types during pre-retrieval may aid in improving the MQWV of a system.

Future work will involve further improving score normalization for system combination, like feature-based supervised learning of normalized document scores.

Acknowledgement

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