Non-Neural Baseline Experiments for CLIR at TREC 2022

Paul McNamee
Johns Hopkins University
Human Language Technology Center of Excellence

Abstract

Cross-Language Information Retrieval (CLIR) returned to TREC with the advent of the NeuCLIR track in 2022. The track provided document collections in three languages: Chinese, Farsi, and Russian, and the principal task involved ranking documents in response to English language queries. Our goal in participating in the NeuCLIR track was to provide a statistical baseline for retrieval, for which we used the HAIRCUT retrieval engine. Experiments included use of character n-gram indexing, use of pseudo-relevance feedback, and application of collection enrichment.

1 Introduction

All of the experiments we conducted made use of the HAIRCUT retrieval engine described by McNamee and Mayfield (2004). One of HAIRCUT's distinctive traits is support for flexible tokenization, and this has enabled the study of character n-grams as indexing terms, which have proven to be effective for controlling the effects of morphological variation (McNamee et al., 2009). Despite HAIRCUT's multilingual support, we did not undertake any experiments requiring indexing of the native language collections — all of our experiments relied on the track-provided document translations into English.

The main motivation for our submissions was to field a statistical system as a comparison to the neural rankers expected in the track. Our runs were produced over a period of several days right before the submission deadline.

2 Submissions

We submitted the maximum number of permitted runs, four, in each of the three document collections in the Adhoc CLIR task. Each run used the same 'language model' approach to retrieval, but varied in the type of indexing terms, use or non-use of automated relevance feedback, and incorporation of a feedback method known as *collection enrichment* (Kwok and Chan, 1998). All submissions used the title and description topic fields under the supposition that this might work better than use of either title or description alone, or with the addition of the narrative.

A unigram statistical language model for retrieval was employed (Hiemstra, 2001; Miller et al., 1999) and smoothing was accomplished using linear interpolation:

$$P(D|Q) \propto \prod_{t \in Q} \lambda P(t|D) + (1 - \lambda) P(t|C) \quad (1)$$

Relative document term frequency was used to estimate $P(t|D)$, as is standard. $P(t|C)$ was based on the mean relative document term frequency from documents in the collection. The two probabilities were evenly weighted (*i.e.,* a constant of $\lambda = 0.5$ was used) in all conditions. With this model we have generally observed retrieval performance to be insensitive to changes in the smoothing parameter λ, although others have reported differently (Zhai and Lafferty, 2004).

The indexing terms were either unnormalized words, or overlapping, word-spanning, character n-grams using $n = 4$ or $n = 5$. When automated relevance feedback was applied, terms were weighted

Run	Fields	Docs	Terms	RF	CE
*pre.xx*4.td.rf	T,D	MT	4-grams	Yes	No
*pre.xx*5.td.rf	T,D	MT	5-grams	Yes	No
*pre.xx*5.td.ce.rf	T,D	MT	5-grams	Yes	Yes
*pre.xx*words.td.rf	T,D	MT	words	Yes	No

Table 1: Attributes of officially submitted runs. Run names began with 'jhumc', a shortening of the team name `jhu.mcnamee`. *xx* is one of fa, ru, or zh.

based on comparing term frequencies in documents from the top 20 ranks and bottom 75 (of 1000) ranked documents. When unstemmed words were used, queries were expanded (or limited) to 60 words; when n-grams were used, the number of terms in the revised query was 200, whether 4-grams or 5-grams. These settings were based on values that have yielded favorable results in past evaluations.

Key attributes of our submissions are given in Table 1. None of our runs was designated as a "baseline" during run submission.

3 Official Results

NIST provided results by email for each of our official runs, and we report aggregate results over the set of topics in Table 2.

4 Experiments

We downloaded the "3-2-1-0" multi-grade relevance judgments from the NIST website on 11/7/22 and scored several runs that we had prepared during the evaluation window, but did not submit. In the sections below we compare the effects of differing choices of indexing terms, topic fields, and methods of query expansion using forms of automated feedback. Metrics were computed using `trec_eval -M 1000 -l2 -m ndcg -m official` *`qrels runfile`*.

4.1 Tokenization

Based on the official submissions, which all used T+D topic fields, there is a lack of consensus about the best choice of indexing term. On average, character 4-grams appear best for Persian, character 5-grams were best for Chinese, and words outperformed for Russian. These are all machine-translated runs with English text. Based on the results in Table 2 the choice between the best method and second-best method is usually less than a 5% relative difference.

4.2 Relevance Feedback

Automated relevance feedback is one of the most successful techniques for improving query performance in text retrieval. While the technique is not guaranteed to improve any particular query, average query performance benefits, typically with a relative gain in mean average precision of about 25%. One of the Russian topics is showcased in Figure 1, Topic #109, where a word-based run using TDN fields is expanded to 60 query terms. Mean average precision on that particular query rises from 0.1916 to 0.2891 (a 51% relative improvement). On other queries, performance decreases with the use of automated feedback. The method relies on finding some relevant documents or at least documents with term distributions similar to relevant documents in the top ranks.

Table 3 calculates the differences in P@10, MAP, and NDCG between most of our official submissions and a corresponding run that did not use automated relevance feedback. Left out are runs that made use of collection enrichment, as they are discussed below in Section 4.4. All of these runs used the title and description fields.

Differences in P@10 were not especially pronounced. For the Persian collection, several runs did better in P@10 without relevance feedback. In Russian and Chinese use of RF always improved P@10 – the largest case was a 15% gain (Russian, 4-grams)

Average values for MAP and NDCG always improved with use of relevance feedback. In Persian and Chinese changes were generally moderate, yielding about 6 - 10% relative improvements. Gains were largest for the Russian collection, with many relative improvements of about 25%.

The size of the collection could matter – the Russian document collection was over twice the size of the Persian one, though many factors other than the size of the collection could be responsible for the varied effects.

4.3 Topic Fields

In Tables 4, 5, and 6 we compare title-only (T) and title+description+narrative (TDN) runs to runs using title+description (TD).

Lang	Run	relret	P@10	MAP	NDCG
fa	jhumc.fa4.td.rf	**1098**	**0.3196**	**0.2291**	**0.5463**
fa	jhumc.fa5.td.rf	994	0.2587	0.2154	0.5220
fa	jhumc.fa5.td.ce.rf	912	0.2870	0.2089	0.4556
fa	jhumc.fawords.td.rf	1070	0.2435	0.2077	0.5177
ru	jhumc.ru4.td.rf	1159	0.2978	0.1983	0.4704
ru	jhumc.ru5.td.rf	**1294**	0.2800	0.1989	0.4836
ru	jhumc.ru5.td.ce.rf	954	0.2911	**0.2253**	0.4395
ru	jhumc.ruwords.td.rf	1192	**0.3089**	0.2016	**0.4860**
zh	jhumc.zh4.td.rf	1391	0.3122	0.2558	0.5405
zh	jhumc.zh5.td.rf	**1575**	0.3265	**0.2746**	**0.5730**
zh	jhumc.zh5.td.ce.rf	1344	**0.3367**	0.2581	0.4963
zh	jhumc.zhwords.td.rf	1451	0.3224	0.2423	0.5579

Table 2: Performance of officially submitted runs.

Lang	Terms	P@10	Δ	MAP	Δ	NDCG	Δ
fa	4-grams	0.2870	-0.0326	0.2120	-0.0171	0.5091	-0.0372
fa	5-grams	0.2696	+0.0109	0.1902	-0.0252	0.4898	-0.0322
fa	words	0.2543	+0.0108	0.1840	-0.0237	0.4744	-0.0433
ru	4-grams	0.2578	-0.0400	0.1548	-0.0435	0.3786	-0.0921
ru	5-grams	0.2622	-0.0178	0.1554	-0.0435	0.3939	-0.0897
ru	words	0.2467	-0.0622	0.1523	-0.0493	0.4376	-0.0484
zh	4-grams	0.3041	-0.0081	0.2289	-0.0269	0.5158	-0.0247
zh	5-grams	0.3204	-0.0061	0.2299	-0.0447	0.5269	-0.0461
zh	words	0.2918	-0.0306	0.2268	-0.0155	0.5295	-0.0284

Table 3: Corresponding runs for our official TD submissions, *without* automated feedback. Nearly all changes are negative, showing that relevance feedback improved average performance.

Title: Researching dead zones
Исследование мертвых зон

Description: I am looking for articles on researching zones in the Ocean that are devoid of oxygen and life.
Я ищу статьи, исследующие зоны в океане, лишенные кислорода и океанской жизни.

Narrative: Dead zones in the ocean are typically created when too much nitrogen, usually from fertilizer runoff, enters the ocean and displaces the oxygen in the water. These areas are devoid of life: no fish, no animal life at all. I would like to find information on these zones and how they are being managed.
Мертвые зоны в океане обычно образуются, когда слишком много азота, обычно из стоков удобрений, попадает в океан и вытесняет кислород в воде. Эти районы лишены жизни: ни рыбы, ни животного мира. Я хотел бы найти информацию об этих зон

2100 (72), animal (223), arabian (86), areas (325), being (165), bottom(100), carbon (78), climate (102), concentration (80), content (96), created (173), dead (409), death (83), devoid (243), earth (110), emergence (86), evolution (77), fish (316), found (83), giant (78), global (95), growth (72), hydrogen (74), hypoxia (90), layer (78) level (82), life (339), like (206), loss (72), low (94), managed (217), marine (125), much (164), nature (77), nitrogen (320), ocean (639), oceans (180), organisms (173), oxygen (568), pacific (71), past (72), researchers (82), researching (321), rocks (101), runoff (258), science (97), scientists (169), sea (120), seas (73), species (75), study (86), temperatures (74), too (145), underwater (79), university (84), usually (132), warming (120), water (362), waters (122), zones (675)

Figure 1: Russian topic 109 with the provided human-produced translation in Russian, and an expanded query after relevance feedback, with relative weights in parenthesis.

	Farsi collection				Russian collection				Chinese collection			
Fields	**relret**	**P@10**	**MAP**	**NDCG**	**relret**	**P@10**	**MAP**	**NDCG**	**relret**	**P@10**	**MAP**	**NDCG**
T	1046	0.2717	0.1935	0.4882	1148	0.3067	**0.2262**	**0.4809**	1365	0.3245	0.2637	0.5381
TD	1098	**0.3196**	0.2291	0.5463	**1159**	0.2978	0.1983	0.4704	1391	0.3122	0.2558	0.5405
TDN	**1141**	0.3087	**0.2309**	**0.5525**	1074	**0.3356**	0.2086	0.4661	**1425**	**0.3367**	**0.2775**	**0.5574**

Table 4: Comparing topic field usage on the three collections. Each run used character 4-grams with RF.

	Farsi collection				Russian collection				Chinese collection			
Fields	**relret**	**P@10**	**MAP**	**NDCG**	**relret**	**P@10**	**MAP**	**NDCG**	**relret**	**P@10**	**MAP**	**NDCG**
T	1009	0.2413	0.1937	0.4977	1238	0.2467	**0.2308**	0.4630	1535	0.3286	0.2627	0.5493
TD	994	0.2587	0.2154	0.5220	1294	0.2800	0.1989	0.4836	**1575**	0.3265	**0.2746**	**0.5730**
TDN	**1120**	**0.2783**	**0.2322**	**0.5554**	**1309**	**0.3667**	0.2268	**0.5107**	1550	**0.3653**	0.2680	0.5701

Table 5: Comparing topic field usage on the three collections. Each run used character 5-grams with RF.

	Farsi collection				Russian collection				Chinese collection			
Fields	**relret**	**P@10**	**MAP**	**NDCG**	**relret**	**P@10**	**MAP**	**NDCG**	**relret**	**P@10**	**MAP**	**NDCG**
T	927	0.1783	0.1532	0.4659	1173	0.2956	0.2006	0.4560	1407	0.3184	0.2295	0.5297
TD	1070	0.2435	0.2077	**0.5177**	1192	0.3089	**0.2016**	**0.4860**	**1451**	**0.3224**	**0.2423**	**0.5579**
TDN	**1083**	**0.2674**	**0.2125**	0.5173	**1296**	**0.3267**	0.1871	0.4741	1399	0.3204	0.2143	0.5121

Table 6: Comparing topic field usage on the three collections. Each run used unstemmed words with RF.

On the Farsi collection it is clear that TDN runs are best, and that both TD and TDN runs substantially outperform title-only runs. These trends hold across all three types of indexing terms.

The picture is less clear on the Russian collection. Differences appear to be less pronounced, although the best combination was character 5-gram indexing with TDN queries.

On the Chinese collection TDN was better with 4-gram indexing, but TD runs were better with 5-grams and words. Title-only runs were clearly less effective.

4.4 Collection Enrichment

The idea of collection enrichment is that the documents used for expanding queries and reweighting query terms do not need to be the same documents which are being ranked. It is possible that a different collection, a larger collection, or a targeted sample of a collection may yield better expansion terms, and make relevance feedback more effective on the document collection that is being ranked. For the NeuCLIR 2022 track, the documents in the three different languages are contemporaneous, and translations of each document collection into English were provided by the track. Thus it is straightforward to build a combined "English" collection by combining the translations of the Persian, Russian, and Chinese collections. This collection can be searched and used for expansion, and the revised query can then be applied to just the English translations of interest (*i.e.,* say the Russian subset).

In practice what we did was just rank the aggregate collection of machine translated documents to rank 5,000 and then filtered out the docids from the non-targeted collections. In some cases this left us with less than 1,000 documents per topic in our submissions.

Results are shown in Table 7 using character 5-gram indexing with TD queries. Performance using collection enrichment with the larger collection generally appears to be worse in MAP and NDCG compared to use of traditional relevance feedback using only the specific target language collection, however, P@10 is slightly improved.

5 Conclusions

Differences were fairly modest between the type of tokenization used across the three collections, however, our best performing runs used character n-grams. Automated relevance feedback was performance enhancing, but gains were significantly larger for the Russian dataset, than for the Persian and Chinese collections. We found that using TD or TDN topic fields was better than title alone, and that differences between TD and TDN were fairly small, with TDN probably being best, especially in Persian. When unstemmed words were used we saw that TD tended to outperform TDN in Russian and Chinese, but generally using the TDN topic fields was best.

	Farsi collection				Russian collection				Chinese collection			
Feedback	**relret**	**P@10**	**MAP**	**NDCG**	**relret**	**P@10**	**MAP**	**NDCG**	**relret**	**P@10**	**MAP**	**NDCG**
None	907	0.2696	0.1902	0.4898	988	0.2622	0.1554	0.3939	1267	0.3204	0.2299	0.5269
RF	**994**	0.2587	**0.2154**	**0.5220**	**1294**	0.2800	0.1989	**0.4836**	**1575**	0.3265	**0.2746**	**0.5730**
CE	912	**0.2870**	0.2089	0.4556	954	**0.2911**	**0.2253**	0.4395	1344	**0.3367**	0.2581	0.4963

Table 7: Collection enrichment runs using character 5-grams and TD topic fields.

References

Djoerd Hiemstra. 2001. *Using Language Models for Information Retrieval*. Ph.D. thesis, University of Twente.

K. L. Kwok and M. Chan. 1998. Improving two-stage adhoc retrieval for short queries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 250–256, New York, NY, USA. ACM.

Paul McNamee and James Mayfield. 2004. Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 7(1-2).

Paul McNamee, Charles Nicholas, and James Mayfield. 2009. Addressing morphological variation in alphabetic languages. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 75–82. ACM.

David R. H. Miller, Tim Leek, and Richard M. Schwartz. 1999. A hidden Markov model information retrieval system. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 214–221, New York, NY, USA. ACM.

Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, April.