HLTCOE at TREC 2022 NeuCLIR Track

Eugene Yang, Dawn Lawrie, James Mayfield
Human Language Technology Center of Excellence, Johns Hopkins University, USA
{eugene.yang,lawrie,mayfield}@jhu.edu

ABSTRACT

The HLTCOE team applied ColBERT-X to the TREC 2022 NeuCLIR track with two training techniques – translate-train (TT) and multilingual translate-train (MTT). TT trains ColBERT-X with English queries and passages automatically translated into the document language from the MS-MARCO v1 collection. This results in three cross-language models for the track, one per language. MTT creates a single model for all three document languages by combining the translations of MS-MARCO passages in all three languages into mixed language batches. Thus the model learns about matching queries to passages simultaneously in all languages. While TT is more effective than MTT in each individual language due to its specificity, MTT still outperforms a strong baseline of BM25 with document translation. On average, MTT and TT perform 34% and 48% higher than the median in MAP with title queries, respectively.

KEYWORDS

cross language information retrieval, neural methods, translated training data

1 INTRODUCTION

ColBERT-X [18] generalizes the state-of-the-art monolingual dense retrieval model, ColBERT [14], to cross-language retrieval and achieves state-of-the-art effectiveness on various benchmark collections, such as HC4 [16] and CLEF [5–7]. The key to such successful generalization is not only replacing the underlying language model with a multilingual one but also the translate-train (TT) training procedure. In the NeuCLIR track, the HLTCOE team applied ColBERT-X with TT and performed, on average, 48% and 54% above the submission median on MAP with title and title+description queries, respectively. In terms of nDCG@20, which focuses at top of the ranking, TT outperforms the median by 20% and 35% with title and title+description queries, respectively.

With an eye toward multilingual retrieval (MLIR) in which the collection consists of documents in multiple languages, the HLTCOE team also experimented with multilingual translate-train (MTT), an extension of TT for multilingual retrieval, on the NeuCLIR CLIR tasks. Instead of training a specialized model for each query-document language pair, MTT results in a single model capable of retrieving documents in any language it has been trained on (Chinese, Persian, and Russian, in this case). Although MTT is designed for retrieving documents from a mixed-language collection, we evaluated its CLIR capability with the NeuCLIR collection as unofficial runs.

Beyond automatic runs, we also recruited human assessors to perform manual searches in the language of the collection using BM25. The assessors developed multiple manual queries by interacting with the search engine for each topic after reading the topic's title, description, and narrative. Interactive search is as effective as searching with the official human query translation using BM25. However, we believe the manual effort provides a more diverse set of relevant documents that improve the reusability of the NeuCLIR-1 collection.

In this notebook paper, we document the model architecture and the training procedure of our automatic runs. For the manual interactive search run, we document how the interactive search was performed and how we created a track submission from each interactive multiple-query session. Note that the members of the HLTCOE team are also co-organizers of the NeuCLIR tracks. Our runs are all marked as manual for fair comparison with other teams, since it is possible that prior exposure to choices made for the track affected our system choices, which then improved our scores. That said, we froze our systems prior to the release of the NeuCLIR topics and made no changes to the systems after topic release.

2 AUTOMATIC RUNS WITH COLBERT-X

This section discusses automatic CLIR runs using the official English title and perhaps description to formulate the queries. We present a more detailed description of ColBERT-X and then present the results with comparisons to our BM25 baselines that we submitted as baseline runs.

2.1 Model Design

ColBERT-X [18] generalizes the ColBERT-v1 [14] retrieval model for CLIR. This retrieval architecture consists of an encoder that embeds the documents as token representations, and a late interaction mechanism that scores each document given a query by summing over each query token the maximum similarity score with any document token. Late interaction enables the classic separation between indexing and query serving found in sparse retrieval systems such as BM25, since it allows document representations to be generated offline.

The ColBERT-X model exploits the late-interaction architecture and instantiates the cross-language ability by using a multilingual pretrained language model, and training with the translate-train (TT) [19] technique. Translate-train uses existing monolingual training resources, such as MS-MARCO [3], by translating the training documents to match the desired query/document language pair.

Table 1: ColBERT-X Index Statistics. The index sizes are identical between ColBERT-X trained with TT and MTT because of the identical indexing setting.

	Chinese	Persian	Russian
# Passages (Millions)	19.8	14.0	25.1
Index Size (TB)	0.9	0.6	1.1
TT Indexing Time (Hours)	9.13	6.48	10.04
MTT Indexing Time (Hours)	7.81	6.34	12.18

Table 2: Effectiveness Summary of ColBERT-X Automatic Runs. The highest scores using the same query type are bold. MTT is marked with * since they are unofficial runs.

Query Type	System	Query Lang.	Doc Lang.	Chinese nDCG@20	Chinese MAP	Chinese R@1000	Persian nDCG@20	Persian MAP	Persian R@1000	Russian nDCG@20	Russian MAP	Russian R@1000
Mean of Submission Medians				0.281	0.185	0.727	0.320	0.198	0.820	0.373	0.258	0.759
T	BM25	MT	Native	0.328	0.261	0.781	0.334	0.221	0.786	0.365	0.287	0.757
		Native	MT	0.356	0.281	**0.805**	0.341	0.232	**0.797**	0.327	0.238	**0.771**
	MTT*	Native	Native	0.388	0.289	0.725	0.372	0.269	0.736	0.412	0.292	0.735
	TT	Native	Native	**0.439**	**0.332**	0.781	**0.395**	**0.273**	0.773	**0.456**	**0.335**	**0.771**
T+D	BM25	MT	Native	0.331	0.258	0.768	0.326	0.234	0.785	0.360	0.281	0.770
		Native	MT	0.340	0.264	0.781	0.355	0.253	**0.829**	0.292	0.216	0.774
	MTT*	Native	Native	0.379	0.282	0.755	0.403	0.285	0.772	0.402	0.277	0.744
	TT	Native	Native	**0.446**	**0.350**	**0.811**	**0.404**	**0.291**	0.808	**0.451**	**0.328**	**0.784**

The model learns retrieval from the translated training collection, providing state-of-the-art effectiveness in CLIR benchmarks.

While TT is effective in CLIR, it requires a document collection in a single language. With an eye toward multilingual retrieval (MLIR), we would like the model to be capable of retrieving documents in a set of languages. The HLTCOE team evaluated a ColBERT-X MLIR model trained with multilingual translate-train (MTT) in the NeuCLIR tasks as an unofficial run. MTT [17] generalizes TT by translating the documents into each target language to equip the model with the ability to retrieve content expressed in these languages.

Unlike training one ColBERT-X CLIR model with TT for each language pair (resulting in three models), we apply the same ColBERT-X MLIR model trained with MTT to all three language pairs simultaneously. This produces a single model capable of participation in each of the NeuCLIR tasks.

2.2 System Pipeline

Both TT and MTT models are trained with translations of MS-MARCO, produced in-house by the HLTCOE. Our translation model is built on top of a transformer base architecture (six-layer encoder/decoder) using Sockeye [11]. We split the original MS-MARCO passages using *ersatz* [21] to produce sentences. We then translated the passages into target languages sentence by sentence.

The ColBERT-X models use XLM-RoBERTa Large [8] for its effectiveness in IR and multilingual tasks [18, 22]. The models were fine-tuned on MSMARCO with translated passages for 200,000 steps with a batch of 64 using four NVIDIA V100 GPUs and a learning rate of 5×10^{-6}.

Table 1 summarizes the statistics of the indexes. The documents in the NeuCLIR collection were tokenized by the XLM-RoBERTa tokenizer and separated into overlapping passages of length 180 tokens with a stride of 90. Indexing uses four GPUs in parallel. The ColBERT-X models retrieve passages, not full documents; the document score is the maximum score of its component passages [4, 10].

2.3 Results

Table 2 presents the aggregated results of the ColBERT-X automatic runs. For both title (T) and title concatenated with description (T+D) as queries, end-to-end ColBERT-X trained with TT provides better effectiveness at the top of the ranking than MTT and BM25 in all three languages as shown by nDCG@20. Interestingly, BM25 with query translation is above the median in Chinese and Persian when using title queries. However, adding the description to the title has mixed results for BM25, which is contrary to what is seen in many collections. This may be due to BM25 failing to utilize the additional information in the description, likely due to query drift from the stop structure in the descriptions. This is also the case for ColBERT-X with MTT but not TT. TT with T+D queries provides even stronger effectiveness compared to its title query variant.

Since both TT and MTT use machine translated MS MARCO during training, differences in translation quality appear to have a significant impact on model effectiveness [18]. Among the three NeuCLIR languages, the quality of translated MS MARCO is worse for Persian, for which the MT model was trained on at most 2/3 of the parallel text of the other models and yields a BLEU score of 20.2 compared to 35.9 and 38.6 for Chinese and Russian, respectively.

ColBERT-X models have lower recall than our statistical model using BM25, especially for title queries. These discrepancies indicate room for improvement on the first phase FAISS [12] nearest neighbor retrieval in ColBERT-X. Simply retrieving documents that contain tokens similar to the query tokens might not be sufficient for retrieving a wide range of relevant documents, resulting in worse recall at 1000 than BM25.

The MLIR variant of ColBERT-X that was trained with MTT performs slightly worse than its TT counterpart. However, it is still more effective than half of the track submissions when evaluated using nDCG@20 or MAP. Note that the median scores presented in Table 2 average the median score for each topic, which is a different and potentially a much higher baseline than the median system score, which is an average of that system over all topics.

Although TT and MTT seem equally good at ranking given that their scores on precision-oriented metrics are similar, TT produces higher Recall@1000. Thus the approximate nearest neighbor search

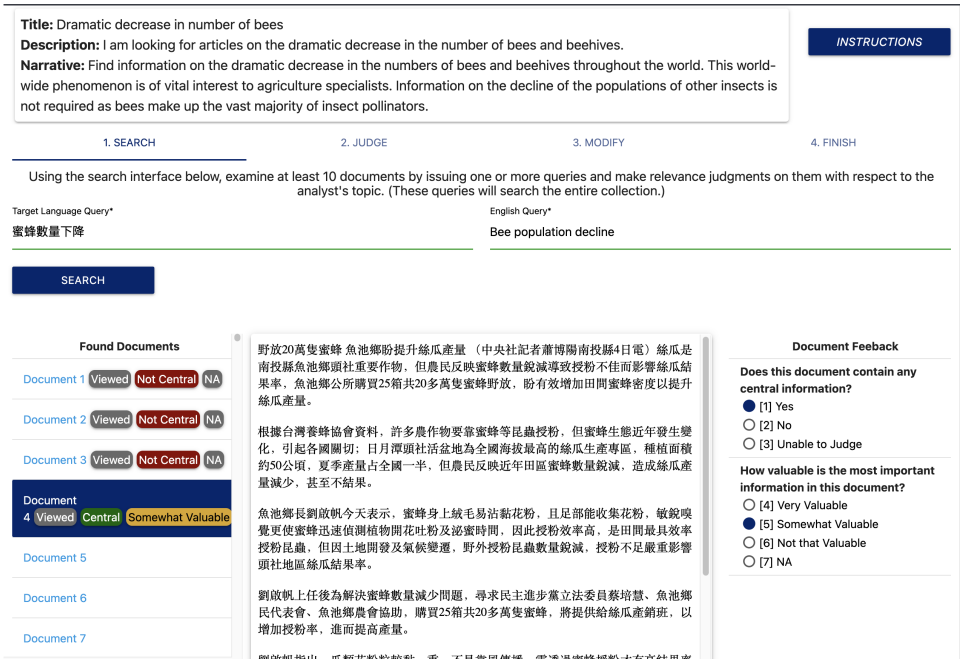

Figure 1: Interface used by assessors for interactive search.

used for the first retrieval stage appears to be working better using the TT model than it does using the MTT model.

3 MANUAL MONOLINGUAL RUNS WITH HUMAN QUERIES

We had access to two sources of human-generated queries. The first source was the human translations of the title and description that were provided to all NeuCLIR participants. We ran these as queries against a BM25 retrieval model using the Patapsco framework [9]. The second source was bilingual speakers that used the English title, description, and narrative to write their own queries in the document language to find relevant documents using interactive search. This section describes this second source in more detail and then presents the results.

3.1 Interactive Runs

It has been shown during the development of prior TREC collections, e.g., TREC Robust, that manual runs where a person issues queries to find relevant documents not discovered by automatic runs helps to make collections more robust [13, 20]. This may be in part because automatic runs are limited in their expression of the query to variants of the title and description. As Alaofi et al. [2] discuss, query variance can be larger than system variance. To perform manual runs in NeuCLIR, one needs access to people with the necessary language skills. We used an assessor pool of bilingual speakers. A majority of these assessors were native English speakers, although a few were native speakers of the language of the documents. While some were proficient in more than two languages, no one was proficient in more than one of Chinese, Persian and Russian.

Using a Mechanical Turk-style platform, Turkle[1], assessors were shown the topic title, description, and narrative. For search the interface provided a monolingual BM25 search engine using the Patapsco framework. Assessors created queries for the search engine and then viewed and judged documents using the interface shown in Figure 1. The left panel allows assessors to select any of the top 100 documents. Labels associated with their judgments are added. This helps assessors to quickly see whether the same judged documents are being returned from different query variants.

HiCAL [1] is an active learning system that helps an interactive user identify relevant documents. Once assessors identified at least one relevant document, they could optionally use HiCAL to recommend additional documents to judge by switching to the “JUDGE" tab in the interface shown in Figure 2. The “next” button was used to ask for more documents from HiCAL. The Turkle platform captured the queries issued to the search engine as well as translations provided by the assessors, the top 1000 document returned by each search, relevance judgements, and whether each document was identified by interactive search or HiCAL. Assessors were given the same graded relevance scale and instructions as NIST assessors for judging documents.

From these assessments, we assembled a single baseline run for each annotated topic. Each run had as its top-rated documents those judged *very valuable* by the assessor; these documents were given a score of 2000. They were followed by documents judged as *somewhat valuable*, which were given a score of 1500. The remainder of the required 1000 documents were selected round-robin over the documents for the issued queries that were not examined by the assessor. The order of the queries for the round-robin assembly

[1] https://github.com/hltcoe/turkle

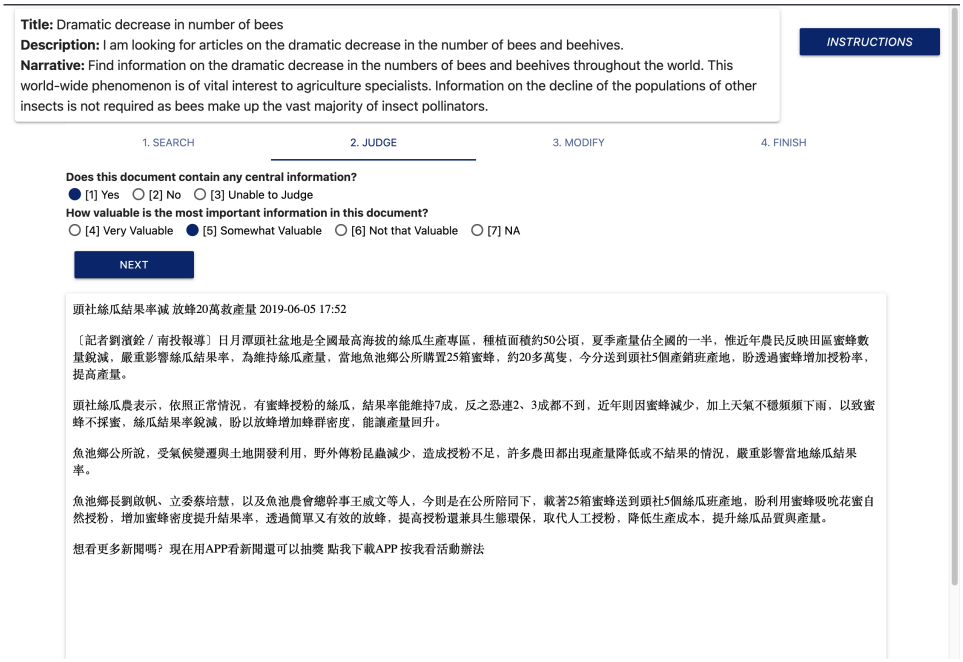

Figure 2: Interface used by assessors for HiCAL.

Table 3: Effectiveness Summary of Monolingual Baselines and Manual Runs.

System	Query	Chinese			Persian			Russian		
		nDCG@20	AP	R@1000	nDCG@20	AP	R@1000	nDCG@20	AP	R@1000
BM25	Title	0.416	0.320	0.778	0.326	0.222	0.788	0.363	0.288	0.759
with official	Description	0.332	0.266	0.735	0.281	0.197	0.742	0.273	0.196	0.701
human translation	Title+Description	0.368	0.296	0.794	0.309	0.212	0.773	0.363	0.279	0.766
BM25	Manual Queries	0.398	0.273	0.687	0.337	0.231	0.795	0.347	0.254	0.680
ColBERT-X(TT)	Manual Queries	0.036	0.025	0.571	0.032	0.022	0.705	0.049	0.036	0.663

was determined by the number of documents the assessor said were very valuable or somewhat valuable in the retrieval results. The top document from this process received a score of 1000 and each subsequent document received a score of one less than its predecessor. No duplicates were allowed.

An additional baseline run was constructed using the query translations. Each English version of the query was issued to ColBERT-X. Although not necessary, we only ran queries issued by an annotator for a particular language against the corresponding document collection; queries created by the Chinese assessor were used for the Chinese collection and so on. Because we were using a language-agnostic CLIR system, all queries for a topic no matter the language of the assessor could have been used in this baseline run. A single manual run was submitted by using a round-robin assembly of results from multiple queries.

3.2 Results

Table 3 shows the results of our monolingual runs. For official translations, the title queries were more effective than title+description queries, although title+description queries led to the highest recall at 1000. This is likely an indication that the stop structure in the descriptions depressed performance, and the descriptions rarely added helpful vocabulary to the queries.

The assessors' interactive searches were more effective than the description queries, but only in Persian were they more effective than the title queries. Part of this lack of effectiveness came from disagreement about relevance between the interactive searchers and NIST assessors. There was more agreement between Chinese assessors (65%) than between Persian (45%) and Russian (52%) assessors, which is also in alignment with the inner-annotator assessment performed by NIST [15]. Including documents that were retrieved during the interactive search but not viewed by the assessor had a small impact; of such documents, only 14% of Chinese, 14% of Persian, and 18% of Russian documents were judged relevant.

While interactive search demonstrated effectiveness competitive with automatic runs, the interactive search queries created for BM25 and run with ColBERT-X TT models were not as effective. One cause may have been that keyword-style queries were not encoded

as effectively by ColBERT-X. The importance of longer queries for ColBERT-X can be observed in Table 2 where title+description queries generally perform better than title queries. Another hypothesis is that the round-robin assembly of multiple queries into a single run could have prevented documents for better expressions of the topic from being judged. More analysis is needed to understand these effects.

4 CONCLUSION

The HLTCOE team participated in both automatic and manual runs (officially all of our runs are manual). The ColBERT-X models outperform the BM25 baselines by a large margin. Although the manual runs performed only on par with title queries, we believe that it is due to the disagreement on document relevance between our searchers and NIST's assessors, which of course is expect since relevance is an opinion, not a fact. There are several directions of future research including more investigation into the query variants provided by interactive search as well as better understanding of first stage retrieval by ColBERT-X MTT.

REFERENCES

[1] Mustafa Abualsaud, Nimesh Ghelani, Haotian Zhang, Mark D Smucker, Gordon V Cormack, and Maura R Grossman. 2018. A System for Efficient High-Recall Retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 1317–1320.

[2] Marwah Alaofi, Luke Gallagher, Dana Mckay, Lauren L. Saling, Mark Sanderson, Falk Scholer, Damiano Spina, and Ryen W. White. 2022. Where Do Queries Come From?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2850–2862. https://doi.org/10.1145/3477495.3531711

[3] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268* (2016).

[4] Michael Bendersky and Oren Kurland. 2008. Utilizing passage-based language models for document retrieval. In *European Conference on Information Retrieval*. Springer, 162–174.

[5] Martin Braschler. 2001. CLEF 2001—Overview of Results. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 9–26.

[6] Martin Braschler. 2002. CLEF 2002—Overview of results. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 9–27.

[7] Martin Braschler. 2003. CLEF 2003–Overview of results. In *Workshop of the cross-language evaluation forum for european languages*. Springer, 44–63.

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[9] Cash Costello, Eugene Yang, Dawn Lawrie, and James Mayfield. 2022. Patapasco: A Python Framework for Cross-Language Information Retrieval Experiments. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022* (Stavanger, Norway). Springer-Verlag, Berlin, Heidelberg, 276–280. https://doi.org/10.1007/978-3-030-99739-7_33

[10] Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 985–988.

[11] Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. 2020. The Sockeye 2 Neural Machine Translation Toolkit at AMTA 2020. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*. Association for Machine Translation in the Americas, Virtual, 110–115.

[12] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[13] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. 1999. Overview of IR tasks at the first NTCIR workshop. In *Proceedings of the first NTCIR workshop on research in Japanese text retrieval and term recognition*. 11–44.

[14] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 39–48.

[15] Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, DouglasW. Oard, Luca Soldaini, and Eugene Yang. 2022. Overview of the TREC 2022 NeuCLIR Track. In *The 31st Text REtrieval Conference*. NIST.

[16] Dawn Lawrie, James Mayfield, Douglas W. Oard, and Eugene Yang. 2022. HC4: A New Suite of Test Collections for Ad Hoc CLIR. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022*.

[17] Dawn Lawrie, Eugene Yang, Douglas W Oard, and James Mayfield. 2023. Neural Approaches to Multilingual Information Retrieval. In *Advances in Information Retrieval: 45th European Conference on IR Research, ECIR 2023*.

[18] Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W Oard. 2022. Transfer learning approaches for building cross-language dense retrieval models. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022*. Springer, 382–396.

[19] P Shi and J Lin. 2019. Cross-lingual relevance transfer for document retrieval. *arXiv preprint arXiv:1911.02989* (2019).

[20] Ellen M Voorhees, Ian Soboroff, and Jimmy Lin. 2022. Can Old TREC Collections Reliably Evaluate Modern Neural Retrieval Models? *arXiv preprint arXiv:2201.11086* (2022).

[21] Rachel Wicks and Matt Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 3995–4007.

[22] Eugene Yang, Suraj Nair, Ramraj Chandradevan, Rebecca Iglesias-Flores, and Douglas W. Oard. 2022. C3: Continued Pretraining with Contrastive Weak Supervision for Cross Language Ad-Hoc Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. ACM, New York, NY, USA, 2507–2512. https://doi.org/10.1145/3477495.3531886