Probabilistic Structured Queries:
The University of Maryland at the TREC 2022 NeuCLIR Track

Suraj Nair[1,2] and Douglas W. Oard[1,3]

[1]UMIACS
[2]Computer Science Department
[3]College of Information Studies
University of Maryland, College Park
srnair@cs.umd.edu

Abstract

The University of Maryland submitted three baseline runs to the Ad Hoc CLIR Task of the TREC 2022 NeuCLIR track. This paper describes three baseline systems that cross the language barrier using a well-known translation-based CLIR technique, Probabilistic Structured Queries.

1 Introduction

The TREC NeuCLIR track focuses on Cross-Language Information Retrieval (CLIR), where the goal is to match English queries with documents that are written in different languages (Chinese, Persian and Russian). At the organizers' request, we submitted three baseline runs (one for each language) in order to support comparison of a statistical baseline with the neural methods that we expected would be tried by others.

2 PSQ

A key challenge in building CLIR systems is to match the queries with documents that use a different vocabulary. Traditional lexical-matching systems such as BM25 [6], which relies on term frequencies and inverse document frequencies, would not work straight off the shelf due to the vocabulary mismatch problem. To address this, we use Probabilistic Structured Queries (PSQ) [3], which leverages translation techniques from Statistical Machine Translation (SMT). PSQ maps term frequency vectors from the document language to the query language using a matrix of translation probabilities (normalized to sum to one in the document language to query language direction) as a simple matrix-vector product. These translation probabilities are learned through word-alignment tools (e.g., GIZA++ [5], BerkeleyAligner [4]) that are traditionally part of SMT systems. Any traditional term weighting function that can accept partial term counts (e.g., BM25 or query likelihood) can then be computed on the resulting term frequency vector.

In our submission, we rely on a HMM-based implementation of PSQ, which in practice operates similarly to query likelihood techniques. We use a 2-state hidden Markov model (HMM) [7] to estimate the relevance of the document given an input query. In the first state θ_e, generates English terms, while in the second state, θ_d, generates document-language terms. Each English query q may consist of N terms $t_1, ..., t_N$.

The generation of query q can then be expressed as shown in equation (1), where f is a document-language term, and ϵ can implement a simple approach to document length normalization.

$$p(q|doc) = \prod_{n=1}^{N} \left[\alpha P(t_n^{(e)}|\theta_e) + (1-\alpha) \sum_{f \in t_n^{(f)}} P(t_n^{(e)}|f) P(f|\theta_d)^{\epsilon} \right] \tag{1}$$

The probability of generating document-language word f from state θ_d is estimated from the counts shown in equation (2).

$$P(f|\theta_d) = \frac{c(f, doc)}{\sum_f c(f, doc)} \tag{2}$$

The probability of generating English word $t_n^{(e)}$ from state θ_e is similarly estimated from counts in a large corpus of English (specifically, the Google one billion word collection [1]). We chose α as 0.1, thereby assigning higher weights to the second state (θ_d) as compared to the first state (θ_e). We set ϵ to 1.

To estimate translation probabilities $P(t_n^{(e)}|f)$, we used IBM Model-1 within the GIZA++ word alignment toolkit. We use sentence-aligned bitext containing ∼36M, ∼6M, and ∼51M sentence pairs for Chinese, Persian and Russian, respectively. For query/document preprocessing as well as the bitext pairs, we use the Patapsco [2] toolkit with lowercasing and Spacy language normalization.

3 Results

Table 1 shows the results of the baseline PSQ runs on the three NeuCLIR languages. All runs complied with the NeuCLIR track's definition of an automatic run using title queries.

Run	Language	nDCG@20	MAP	Recall@100	Recall@1000
umcp_hmm_zh	Chinese	0.3029	0.1471	0.3429	0.6323
umcp_hmm_fa	Persian	0.2716	0.1206	0.4172	0.6498
umcp_hmm_ru	Russian	0.3192	0.1809	0.3482	0.5966

Table 1: Official evaluation measures for the baseline PSQ runs using title queries.

4 Conclusion

In this paper we describe our baseline PSQ runs that were submitted to the TREC 2022 NeuCLIR track. We build a probabilistic matching system using translation probabilities estimated with sentence-aligned bitext pairs.

References

[1] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

[2] C. Costello, E. Yang, D. Lawrie, and J. Mayfield. Patapasco: A python framework for cross-language information retrieval experiments. In M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, and V. Setty, editors, *Advances in Information Retrieval*, pages 276–280, Cham, 2022. Springer International Publishing.

[3] K. Darwish and D. W. Oard. Probabilistic structured query methods. In *Proceedings of the 26th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 338–344, 2003.

[4] A. Haghighi, J. Blitzer, J. DeNero, and D. Klein. Better word alignments with supervised ITG models. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 923–931, Suntec, Singapore, Aug. 2009. Association for Computational Linguistics.

[5] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[6] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al. Okapi at TREC-3. *NIST Special Publication Sp 109*, 1995.

[7] J. Xu and R. Weischedel. Cross-lingual information retrieval using hidden Markov models. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 95–103, Hong Kong, China, Oct. 2000. Association for Computational Linguistics.