SIGIR 2023 Tutorial References
Cross-language Information Retrieval

Eugene Yang[1], Dawn Lawrie[1], James Mayfield[1],
Suraj Nair[2], Douglas W. Oard[2]

[1]HLTCOE, Johns Hopkins University, USA
[2]University of Maryland, College Park, USA
eyang35@jh.edu

INTRODUCTION

CLIR Surveys

1. D. W. Oard and A. R. Diekema. Cross-language information retrieval. *Annual Review of Information Science and Technology (ARIST)*, 33:223–56, 1998
2. J.-Y. Nie. *Cross-Language Information Retrieval.* Morgan and Claypool Publishers, 2010. ISBN 1598298631
3. P. Galuščáková, D. W. Oard, and S. Nair. Cross-language information retrieval. *arXiv preprint arXiv:2111.05988*, 2022

The First Wave: Multilingual Theasuri

1. Organización Internacional de Normalización. Comité Técnico ISO/TC 46, Información y Documentación. *ISO 5964: Documentation: Guidelines for the establishment and development of multilingual thesauri.* International Organization for Standardization, 1985

The Second Wave: Term Translation

1. G. Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21(3):187–194, 1970
2. L. Ballesteros and W. B. Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information retrieval*, pages 84–91, 1997
3. A. Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 55–63. ACM, 1998

4. J. S. McCarley. Should we translate the documents or the queries in cross-language information retrieval? In R. Dale and K. W. Church, editors, *27th Annual Meeting of the Association for Computational Linguistics, University of Maryland, College Park, Maryland, USA, 20-26 June 1999*, pages 208–214. ACL, 1999
5. 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
6. K. Darwish and D. W. Oard. Probabilistic structured query methods. In C. L. A. Clarke, G. V. Cormack, J. Callan, D. Hawking, and A. F. Smeaton, editors, *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada*, pages 338–344. ACM, 2003

The Dawn of the Third Wave: Latent Semantic Indexing

1. T. K. Landauer and M. L. Littman. A statistical method for language-independent representation of the topical content of text segments. In *Proceedings of the 11th International Conference: Expert Systems and Their Applications, 1991*, 1991
2. M. L. Littman, F. Jiang, and G. A. Keim. Learning a language-independent representation for terms from a partially aligned corpus. In *International Conference on Machine Learning*, pages 314–322, 1998

FOUNDATIONAL NEURAL IR METHODS

1. V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, Nov. 2020. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550. URL `https://www.aclweb.org/anthology/2020.emnlp-main.550`
2. O. Khattab and M. Zaharia. 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*, SIGIR '20, page 39–48, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164
3. T. Formal, C. Lassance, B. Piwowarski, and S. Clinchant. Splade v2: Sparse lexical and expansion model for information retrieval, 2021. URL `https://arxiv.org/abs/2109.10086`
4. R. Nogueira and K. Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019

NEURAL CLIR METHODS

Dense Bi-Encoders: One Vector per Query/Doc

1. A. Asai, J. Kasai, J. Clark, K. Lee, E. Choi, and H. Hajishirzi. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online, June 2021. Association for Computational Linguistics
2. X. Zhang, X. Ma, P. Shi, and J. Lin. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics
3. R. Litschko, I. Vulić, S. P. Ponzetto, and G. Glavaš. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, and F. Sebastiani, editors, *Advances in Information Retrieval*, pages 342–358, Cham, 2021. Springer International Publishing. ISBN 978-3-030-72113-8

Dense Bi-Encoders: One Vector per Term

1. S. Nair, E. Yang, D. Lawrie, K. Duh, P. McNamee, K. Murray, J. Mayfield, and D. W. Oard. Transfer learning approaches for building cross-language dense retrieval models. In M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, and V. Setty, editors, *Proceedings of the European Conference on Information Retrieval*, pages 382–396, Cham, 2022. Springer International Publishing. ISBN 978-3-030-99736-6
2. Y. Li, M. Franz, M. A. Sultan, B. Iyer, Y.-S. Lee, and A. Sil. Learning cross-lingual IR from an English retriever. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4428–4436, Seattle, United States, July 2022. Association for Computational Linguistics
3. Z. Huang, P. Yu, and J. Allan. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1048–1056, 2023

Sparse Bi-Encoders: One Vector per Query/Doc

1. S. Nair, E. Yang, D. Lawrie, J. Mayfield, and D. W. Oard. Learning a sparse representation model for neural CLIR. *Design of Experimental Search and Information REtrieval Systems (DESIRES)*, 2022
2. S. Nair, E. Yang, D. Lawrie, J. Mayfield, and D. W. Oard. BLADE: Combining vocabulary pruning and intermediate pretraining for scaleable neural CLIR. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2023

Cross Encoders

1. R. Nogueira, W. Yang, K. Cho, and J. Lin. Multi-stage document ranking with BERT. *arXiv preprint arXiv:1910.14424*, 2019
2. R. Pradeep, R. Nogueira, and J. Lin. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv preprint arXiv:2101.05667*, 2021

CLIR TRAINING STRATEGIES

1. E. Yang, S. Nair, R. Chandradevan, R. Iglesias-Flores, and D. W. Oard. 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*, pages 2507–2512, 2022
2. L. Gao and J. Callan. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540*, 2021
3. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020
4. (Repeated from NEURAL CLIR METHODS) S. Nair, E. Yang, D. Lawrie, K. Duh, P. McNamee, K. Murray, J. Mayfield, and D. W. Oard. Transfer learning approaches for building cross-language dense retrieval models. In M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, and V. Setty, editors, *Proceedings of the European Conference on Information Retrieval*, pages 382–396, Cham, 2022. Springer International Publishing. ISBN 978-3-030-99736-6

CLIR TRAINING DATA

1. P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang. MS MARCO: A human generated machine reading comprehension dataset. arXiv preprint arXiv:1611.09268, 2016
2. L. H. Bonifacio, I. Campiotti, V. Jeronymo, R. Lotufo, and R. Nogueira. mMARCO: A multilingual version of the MS MARCO passage ranking dataset. *arXiv preprint arXiv:2108.13897*, 2021
3. Z. Dai, V. Y. Zhao, J. Ma, Y. Luan, J. Ni, J. Lu, A. Bakalov, K. Guu, K. B. Hall, and M.-W. Chang. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*, 2022
4. J. Mayfield, E. Yang, D. Lawrie, S. Barham, O. Weller, M. Mason, S. Nair, and S. Miller. Synthetic cross-language information retrieval training data. *arXiv preprint arXiv:2305.00331*, 2023
5. (Repeated from CLIR TRAINING STRATEGIES) L. Gao and J. Callan. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540*, 2021

FROM BYTES TO TOKENS

Naming Languages

1. iso.org. ISO 639 language codes. URL `https://www.iso.org/iso-639-language-codes.html`. Accessed: 2023-07-23
2. E. M. Bender. The #BenderRule: on naming the languages we study and why it matters. *The Gradient*, 2019. URL `https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/`

Encodings

1. Unicode Consortium. Unicode, . URL `home.unicode.org`. Accessed: 2023-07-23
2. Unicode Consortium. Unicode® 15.0.0, . URL `https://www.unicode.org/versions/Unicode15.0.0/`. Accessed: 2023-07-23
3. T. Whitlock. Unicode inspector. URL `apps.timwhitlock.info/unicode/inspect`. Accessed: 2023-07-23
4. B. Milde. Shapecatcher.com: Unicode character recognition. URL `https://shapecatcher.com/`. Accessed: 2023-07-23

Normalization

1. Unicode normalization forms (unicode® standard annex #15), 2022. URL `https://www.unicode.org/reports/tr15/`. Version 53, Ken Whistler, ed. Accessed: 2023-07-23
2. Wikipedia contributors. Unicode normalization forms (unicode® standard annex #15), 2023. URL `https://en.wikipedia.org/wiki/Unicode_equivalence`. Accessed: 2023-07-23
3. unicodedata. unicodedata — unicode database, 2023. URL `https://docs.python.org/3/library/unicodedata.html`. Accessed: 2023-07-23

Subwords

1. C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948. URL `http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf`
2. P. Mcnamee and J. Mayfield. Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 7:73–97, 01 2004
3. Y. Shibata, T. Kida, S. Fukamachi, M. Takeda, A. Shinohara, T. Shinohara, and S. Arikawa. Byte pair encoding: A text compression scheme that accelerates pattern matching. Technical Report DOI-TR-161, Kyushu University, Department of Informatics, 1999

4. Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv areprint arXiv:1609.08144, 2016
5. T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. URL `https://aclanthology.org/D18-2012`

Mixed Scripts and Languages

1. I. R. H. Dale. Digraphia. *International Journal of the Sociology of Language*, 1980(26):5–14, 1980. https://doi.org/doi:10.1515/ijsl.1980.26.5. URL `https://doi.org/10.1515/ijsl.1980.26.5`
2. J. Wang and A. Komlodi. Switching languages in online searching: A qualitative study of web users' code-switching search behaviors. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, pages 201–210, 2018

CLIR EVALUATION

Evaluation Introduction

1. E. M. Voorhees. The evolution of Cranfield. *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*, pages 45–69, 2019

Test Collections

1. (Repeated from INTRODUCTION: Surveys) P. Galuščáková, D. W. Oard, and S. Nair. Cross-language information retrieval. *arXiv preprint arXiv:2111.05988*, 2022
2. E. Voorhees and D. Harman. Overview of the sixth text retrieval conference (TREC-6). *Information Processing & Management*, 36:3–35, 01 2000
3. M. Braschler. CLEF 2002 — Overview of results. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Advances in Cross-Language Information Retrieval*, pages 9–27, 2003. ISBN 978-3-540-45237-9
4. S. Lee, S.-H. Myaeng, H. Kim, J. Seo, B. Lee, and S. Cho. Characteristics of the Korean test collection for CLIR in NTCIR-3. In *NTCIR*, 01 2002

5. M. Mitra and P. Majumdar. FIRE: Forum for information retrieval evaluation. In *Proceedings of the 2nd workshop on Cross Lingual Information Access (CLIA) Addressing the Information Need of Multilingual Societies*, 2008
6. B. V. Dobrov, I. Kuralenok, N. V. Loukachevitch, I. S. Nekrestyanov, and I. Segalovich. Russian information retrieval evaluation seminar. In *LREC*, 2004
7. E. M. Voorhees. The evolution of Cranfield. *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*, pages 45–69, 2019
8. D. Lawrie, J. Mayfield, D. W. Oard, and E. Yang. HC4: A new suite of test collections for ad hoc CLIR. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR)*, 2022
9. D. Lawrie, J. Mayfield, D. W. Oard, E. Yang, S. Nair, and P. Galuščáková. HC3: A suite of test collections for CLIR evaluation over informal text. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (Taipei, Taiwan)(SIGIR'23).*, 2023
10. S. Sun and K. Duh. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170, Online, 2020. ACL
11. A. Asai, J. Kasai, J. Clark, K. Lee, E. Choi, and H. Hajishirzi. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online, June 2021. Association for Computational Linguistics

Metrics

1. L. Ballesteros and W. B. Croft. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71, 1998

Building a CLIR Test Collection

1. D. Lawrie, S. MacAvaney, J. Mayfield, P. McNamee, D. W. Oard, L. Soldaini, and E. Yang. Overview of the TREC 2022 NeuCLIR track. In *Proceedings of the Thirty-First Text Retrieval Conference*. 2022
2. D. Graff, C. Cieri, S. Strassel, and N. Martey. The TDT-3 text and speech corpus. In *in Proceedings of DARPA Broadcast News Workshop*, pages 57–60. Morgan Kaufmann, 1999
3. E. M. Voorhees, I. Soboroff, and J. Lin. Can old TREC collections reliably evaluate modern neural retrieval models? *arXiv preprint arXiv:2201.11086*, 2022

4. G. V. Cormack, H. Zhang, N. Ghelani, M. Abualsaud, M. D. Smucker, M. R. Grossman, S. Rahbariasl, and A. Ghenai. Dynamic sampling meets pooling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1217–1220, 2019

MULTILINGUAL IR

1. D. W. Oard and B. J. Dorr. A survey of multilingual text retrieval. Technical report, USA, 1996
2. R. Rahimi, A. Shakery, and I. King. Multilingual information retrieval in the language modeling framework. *Information Retrieval*, 18:246–281, 2015
3. P. Sorg and P. Cimiano. Exploiting Wikipedia for cross-lingual and multilingual information retrieval. *Data & Knowledge Engineering*, 74:26–45, 2012
4. D. Lawrie, E. Yang, D. W. Oard, and J. Mayfield. Neural approaches to multilingual information retrieval. In *European Conference on Information Retrieval*, pages 521–536. Springer, 2023
5. T. Leek, R. Schwartz, and S. Sista. Probabilistic approaches to topic detection and tracking. In *Topic Detection and Eracking: Event-Based Information Organization*, pages 67–83. Springer, 2002
6. (Repeated from CLIR EVALUATION: Test Collections) M. Braschler. CLEF 2002 — Overview of results. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Advances in Cross-Language Information Retrieval*, pages 9–27, 2003. ISBN 978-3-540-45237-9
7. (Repeated from CLIR EVALUATION: Building a CLIR Test Collection) D. Lawrie, S. MacAvaney, J. Mayfield, P. McNamee, D. W. Oard, L. Soldaini, and E. Yang. Overview of the TREC 2022 NeuCLIR track. In *Proceedings of the Thirty-First Text Retrieval Conference*. 2022
8. (Repeated from CLIR EVALUATION: Building a CLIR Test Collection) D. Graff, C. Cieri, S. Strassel, and N. Martey. The TDT-3 text and speech corpus. In *in Proceedings of DARPA Broadcast News Workshop*, pages 57–60. Morgan Kaufmann, 1999
9. M. Braschler, J. Krause, C. Peters, and P. Schäuble. Cross-language information retrieval (CLIR) track overview. In *TREC*, 1999

SOME OPEN RESEARCH QUESTIONS

Language Bias in MLIR

1. (Repeated from MULTILINGUAL IR) D. Lawrie, E. Yang, D. W. Oard, and J. Mayfield. Neural approaches to multilingual information retrieval. In *European Conference on Information Retrieval*, pages 521–536. Springer, 2023

Beyond News Content

1. P. Pecina, P. Hoffmannova, G. J. Jones, Y. Zhang, and D. W. Oard. Overview of the CLEF-2007 cross-language speech retrieval track. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers 8*, pages 674–686. Springer, 2008
2. (Repeated from CLIR EVALUATION: Test Collections) D. Lawrie, J. Mayfield, D. W. Oard, E. Yang, S. Nair, and P. Galuščáková. HC3: A suite of test collections for CLIR evaluation over informal text. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (Taipei, Taiwan)(SIGIR'23).*, 2023

New Sources of Training Data

1. G. Faggioli, L. Dietz, C. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, et al. Perspectives on large language models for relevance judgment. *arXiv preprint arXiv:2304.09161*, 2023
2. J. Mayfield, E. Yang, D. Lawrie, S. Barham, O. Weller, M. Mason, S. Nair, and S. Miller. Synthetic cross-language information retrieval training data. *arXiv preprint arXiv:2305.00331*, 2023
3. F. Daniel, P. Kucherbaev, C. Cappiello, B. Benatallah, and M. Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51 (1):1–40, 2018

Interactive CLIR

1. D. Petrelli, S. Levin, M. Beaulieu, and M. Sanderson. Which user interaction for cross-language information retrieval? Design issues and reflections. *Journal of the American Society for Information Science and Technology*, 57 (5):709–722, 2006

Efficiency

1. (Repeated from NEURAL CLIR METHODS) S. Nair, E. Yang, D. Lawrie, J. Mayfield, and D. W. Oard. BLADE: Combining vocabulary pruning and intermediate pretraining for scaleable neural CLIR. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2023

Other Ways to Cross The Language Barrier

1. Z. Huang, H. Bonab, S. M. Sarwar, R. Rahimi, and J. Allan. Mixed attention transformer for leveraging word-level knowledge to neural cross-lingual information retrieval. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 760–770, 2021

2. C. Lassance, H. Dejean, and S. Clinchant. An experimental study on pretraining transformers from scratch for ir. In *European Conference on Information Retrieval*, pages 504–520. Springer, 2023

CLIR Augmented Large Language Models

1. A. Asai, T. Schick, P. Lewis, X. Chen, G. Izacard, S. Riedel, H. Hajishirzi, and W.-t. Yih. Task-aware retrieval with instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3650–3675, Toronto, Canada, July 2023. Association for Computational Linguistics
2. E. Nie, S. Liang, H. Schmid, and H. Schütze. Cross-lingual retrieval augmented prompt for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8320–8340, Toronto, Canada, July 2023. Association for Computational Linguistics