TREC NeuCLIR 2022

Cross-Language IR

跨语言搜索

جستجوی چند زبانه

Межъязыковой поиск

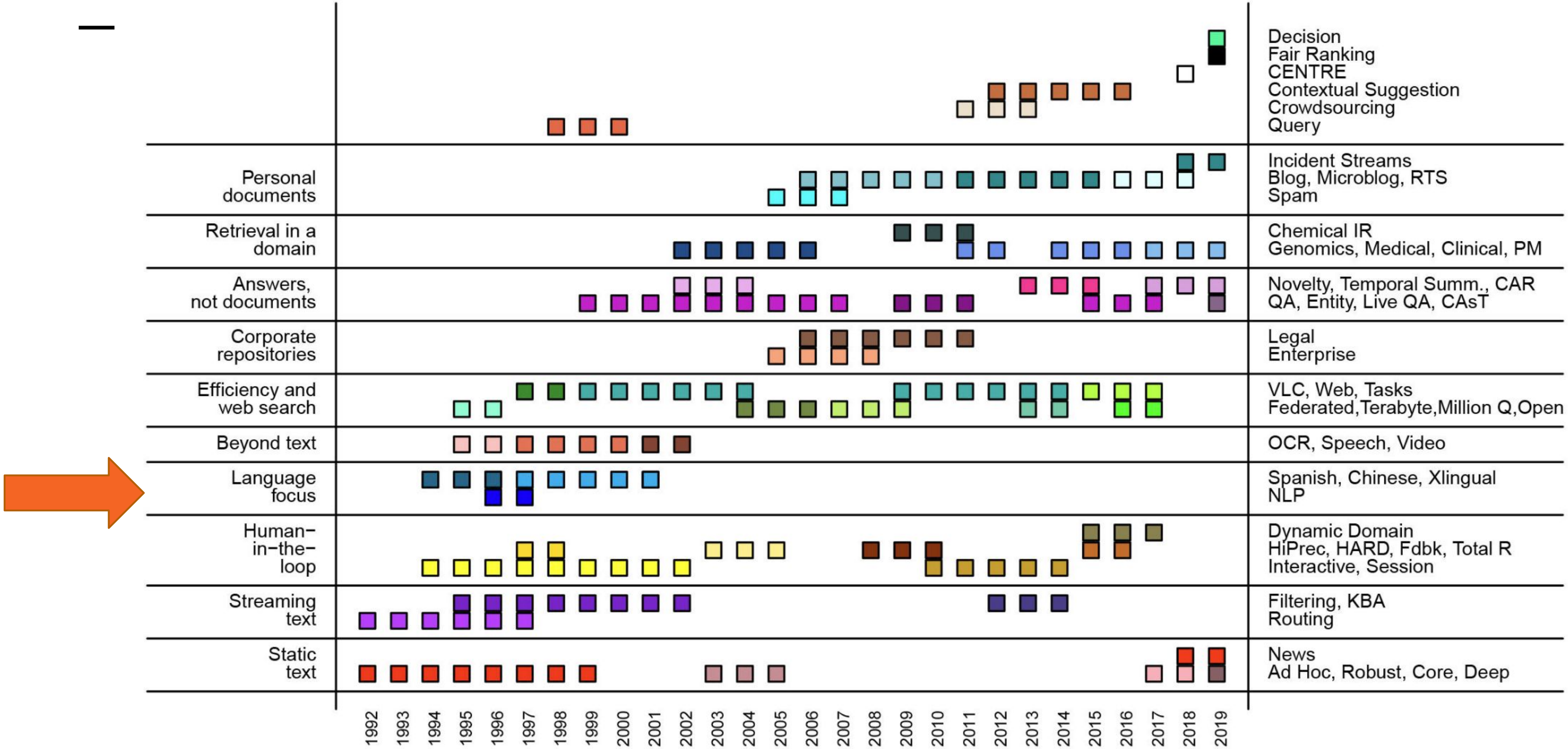
neuclir.github.io · neuclir-organizers@googlegroups.com

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee,
Douglas W. Oard, Luca Soldaini, Eugene Yang

What is TREC?

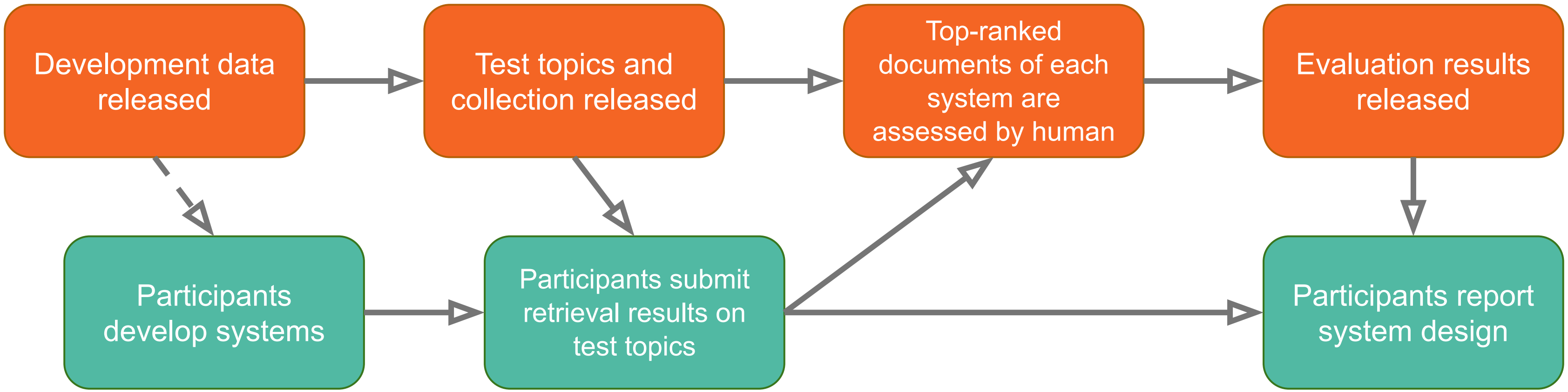
- Text REtrieval Conference organized by NIST since 1991
- Hosts various shared retrieval tasks
  - Domain-specific, e.g. Clinical Trial Track, Incident Stream Track
  - Corporate repositories, e.g. Legal Track, Total Recall Track
  - QA, e.g. CAsT, CAR
- Evaluate/develop state-of-the-art retrieval systems
- Create evaluation collections via pooling

https://trec.nist.gov/



Ellen Voorhees. *Seek and You Will (Probably) Find: Helping Mine COVID-19 Research Data.* July 22, 2020.
https://www.nist.gov/blogs/taking-measure/seek-and-you-will-probably-find-helping-mine-covid-19-research-data

Typical workflow of a TREC Track


Development data released
Test topics and collection released
Top-ranked documents of each system are assessed by human
Evaluation results released
Participants develop systems
Participants submit retrieval results on test topics
Participants report system design

Main Task: Cross-Language Retrieval

🔍 Political reaction to McCain’s death

English Queries

中部3宗山火强风下蔓延 至少8人伤包括消防员｜即时新闻｜亚欧非｜on.cc东网

葡萄牙中部布朗库堡地区发生3宗山火，当局周日（21日）派出逾千名消防员扑救，但强风令火势向三方蔓延。目前至少一名平民在大火中重伤，另有7名消防员受伤。

山火发生在草木丛生的布朗库堡山区，在首都里斯本以北200公里，火势上周六（20日）下午在强风下，分岔向三个方向蔓延。当局派出大批消防员扑救，并出动20架直升机和飞机协助，但入夜后当局难以派出更多飞机。军方则表示，已派出20名士兵及重型机械，为消防员开路。

E.g., Chinese Corpus

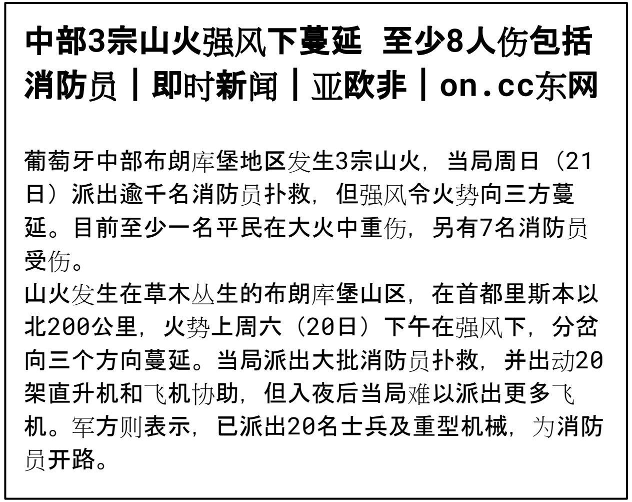
Result list:

1. 麦凯恩之死的时代注脚：中道政治在美国已成往事 - 澎湃新闻
2. 麦凯恩之死的时代注脚：中道政治在美国已成往事 | 早报
3. 麦凯恩对特朗普“反抗到底”，生前“遗嘱”不请总统出席葬礼
4. …

Auxiliary Task: (Non-English) Retrieval

🔎 麦凯恩之死的政治反应

E.g. Chinese Queries



中部3宗山火强风下蔓延　至少8人伤包括消防员 | 即时新闻 | 亚欧非 | on.cc东网

葡萄牙中部布朗库堡地区发生3宗山火，当局周日（21日）派出逾千名消防员扑救，但强风令火势向三方蔓延。目前至少一名平民在大火中重伤，另有7名消防员受伤。

山火发生在草木丛生的布朗库堡山区，在首都里斯本以北200公里，火势上周六（20日）下午在强风下，分岔向三个方向蔓延。当局派出大批消防员扑救，并出动20架直升机和飞机协助，但入夜后当局难以派出更多飞机。军方则表示，已派出20名士兵及重型机械，为消防员开路。

E.g., Chinese Corpus

Result list:

1. 麦凯恩之死的时代注脚：中道政治在美国已成往事 - 澎湃新闻
2. 麦凯恩之死的时代注脚：中道政治在美国已成往事 | 早报
3. 麦凯恩对特朗普“反抗到底”，生前“遗嘱”不请总统出席葬礼
4. …

Why Cross-Language at TREC?

- TREC hosted cross-language tracks in the 90’s and early 00’s
- Yielded to other evaluations (e.g., CLEF, NTCIR, FIRE, ...)
- Since then, these evaluations have largely moved on to other tasks
- Relevance assessments inadequate for testing latest neural models

What could we learn from NeuCLIR?

What are the best neural CLIR approaches?

How does neural CLIR compare to the combination of machine translation and monolingual IR?

How does it compare to the strongest statistical approaches to CLIR?

How do the resource requirements for the various approaches compare?

Can separate ranking and reranking phases improve performance over a single-phase approach?

What resources are most useful for training CLIR systems?

NeuCLIR: Tasks

1. **Cross-language retrieval**
    a. **Full Ranking**
    b. **Reranking**
    Using English topics to retrieve documents from non-English language
2. **Monolingual retrieval**
    a. **Full Ranking**
    b. **Re-ranking**
    Using non-English topics to retrieve documents in the same language

NeuCLIR: Run Types

Automatic

Untouched by human hands after that human has seen one or more test topics

Manual

Any change to topic processing, document processing, retrieval, ranking or reranking after seeing any test topic

NeuCLIR: Corpora

Three News Corpora

- Chinese (~2M docs, avg 650 characters)
    - Chinese includes both traditional & simplified
    - Character mapping script available
- Persian (~1.25M docs, avg 200 tokens)
- Russian (~11.5M docs, avg 400 tokens)

News articles from the Common Crawl

Full corpus provided to task participants

Software to generate corpora to be made available to all

Articles dated from August 2016 - July 2021

Corpora (automatically) de-duplicated

Machine translations of documents into English provided

```
{
  "id": "69a0792a-5b6d-499e-889a-9cc4037ed80d",
  "cc_file":
"crawl-data/CC-NEWS/2019/07/CC-NEWS-20190721080517-00625.war
c.gz",
  "date": "2019-07-21",
  "title": "中部3宗山火强风下蔓延 至少8人伤包括消防员｜即时新闻｜亚欧
非｜on.cc东网",
  "text": "葡萄牙中部布朗库堡地区发生3宗山火，当局周日（21日）派出逾千
名消防员扑救，但强风令火势向三方蔓延。目前至少一名平民在大火中重伤，另有7
名消防员受伤。\n\n\n\n山火发生在草木丛生的布朗库堡山区，在首都里斯本以
北200公里，火势上周六（20日）下午在强风下，分岔向三个方向蔓延。当局派出
大批消防员扑救，并出动20架直升机和飞机协助，但入夜后当局难以派出更多飞
机。军方则表示，已派出20名士兵及重型机械，为消防员开路。\n\n\n\n一名平
民在山火中严重受伤，由直升机送往里斯本救治，另有4名消防员救火期间受伤，3
人则在消防车车祸中受伤。多条公路受山火影响而要关闭。当局表示，疏散了数条
村落以防万一，又向中部及南部6个地区发出最高的火灾警戒。",
  "url":
"https://hk.on.cc/hk/bkn/cnt/aeanews/20190721/bkn-2019072117
0515141-0721_00912_001.html"
}
```

There were three wildfires in the Castelo Branco area in central Portugal. The authorities sent more than 1,000 firefighters to fight the fire on Sunday (21st), but strong winds spread the fire to three parties. At least one civilian was seriously injured in the fire, and another seven firefighters were injured.

Mountain fire occurred in the vegetation of Castelo Branco mountain area, 200 kilometers north of the capital Lisbon. The fire spread in three directions under strong winds last Saturday (20th) afternoon. The authorities sent a large number of firefighters to the rescue, and dispatched 20 helicopters and aircraft to assist, but it was difficult for the authorities to send more aircraft after nightfall. The military said that it has sent 20 soldiers and heavy machinery to clear the way for firefighters.

A civilian was seriously injured in the mountain fire and was sent to Lisbon by helicopter for treatment. Another 4 firefighters were injured during the fire fighting, and 3 people were injured in the fire truck accident. Many highways have to be closed due to the impact of wildfires. The authorities said that several villages were evacuated as a precaution, and the highest fire alert was issued to six areas in the central and southern regions.

NeuCLIR: Topics

Title: Political reaction to McCain’s death
Persian Title: واکنش سیاسی به مرگ مک کین
Description: How did politicians react to the death of John McCain and how did his death impact US politics?
Persian Narrative: Articles which explained reactions to Mccain's death, by Obama, George Bush, Trump, and stated its effect on current politics, such as Trump not responding appropriately or not being invited to the Funeral

Title: Political reaction to McCain’s death
Chinese Title: 对麦凯恩之死的政治反应
Description: How did politicians react to the death of John McCain and how did his death impact US politics?
Chinese Narrative: Describe the political tensions at play at the time of McCain's death and explain how his death impacted that dynamic

```
<title>
<desc>
<narr>
```

English Topics

Typical TREC-style ad hoc title/description/narrative format

Including human translation into target language

- Facilitates a monolingual version of the task
- Helps enrich the judgment pool

Topics focus on events and people

NeuCLIR: Evaluation

Three Relevance Grades. Imagine a report about the topic…

Highly Relevant: Contains facts that would be included in lead paragraph of report

Relevant: Contains facts that would be included elsewhere in report

Not Relevant: Contains no information that would be included

Evaluation Criteria

Precision-oriented measures (e.g., MRR)

Recall-oriented measures (e.g., nDCG)

Efficiency (e.g., Participant-reported MRT, number of model parameters, index size)

Baseline

Machine/Human/PSQ Query Translation with BM25 ranking (enables a re-ranking task)

Resources

Validation Data:
HC4 Dataset, containing similar topics, deep relevance judgments (to be released)

Training Data:
- Machine translations of MS MARCO (to be released)
- CLIRMatrix: https://www.cs.jhu.edu/~shuosun/clirmatrix/
- A variety of large-scale non-English monolingual datasets (ML-WikIR, MMARCO, MR-TyDi, etc.)

Tools:
- **Patapsco:** A tool for running CLIR Experiments: https://github.com/hltcoe/patapsco
- **Other tools/resources:** trec_eval, ir-datasets, PyTerrier, ir-measures, etc.

Recent Survey: Galuščáková, Oard, and Nair. Cross-language Information Retrieval.
https://arxiv.org/abs/2111.05988

Task Website: https://neuclir.github.io/

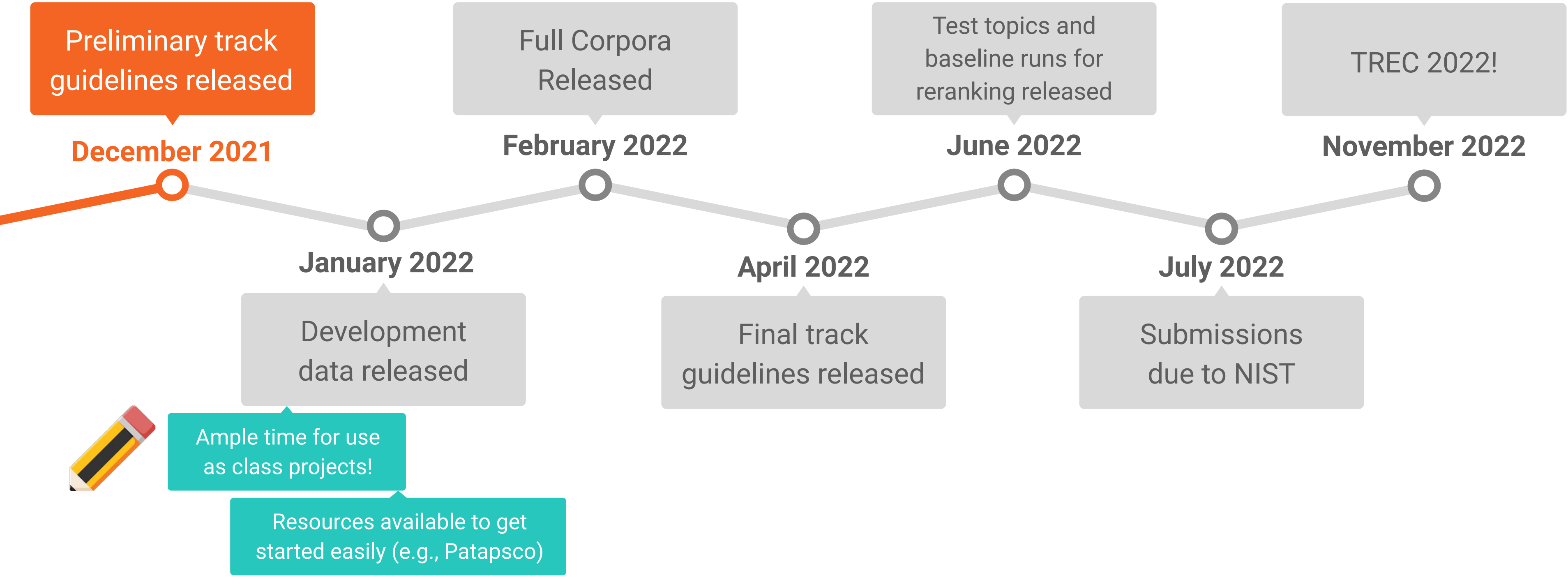
Organizer contact: neuclir-organizers@googlegroups.com

Can NeuCLIR be the first Carbon-Neutral TREC Task?

We plan to:

- Suggest that teams estimate the carbon contribution of their participation.
- Encourage them to find ways to offset their carbon contributions (e.g., with carbon credits).

NOTE: This is not a requirement for participation.



December 2021
Preliminary track guidelines released

January 2022
Development data released

Ample time for use as class projects!

Resources available to get started easily (e.g., Patapsco)

February 2022
Full Corpora Released

April 2022
Final track guidelines released

June 2022
Test topics and baseline runs for reranking released

July 2022
Submissions due to NIST

November 2022
TREC 2022!

Track Organizers

Dawn Lawrie
Johns Hopkins HLTCOE

Sean MacAvaney
University of Glasgow

James Mayfield
Johns Hopkins HLTCOE

Paul McNamee
Johns Hopkins HLTCOE

Douglas W. Oard
University of Maryland

Luca Soldaini
Amazon Alexa AI

Eugene Yang
Johns Hopkins HLTCOE

Feedback/Questions?

- **Will you consider participating?**
- **NeuCLIR as a class project** -- does anybody plan to do this? Are we providing enough resources?
- **Evaluation metrics** -- nDCG, MAP, MRR, others? MRT, # of model parameters, index size?
- **Carbon-neutrality** -- Is this feasible?

TREC NeuCLIR 2022

🌐 neuclir.github.io
✉️ neuclir-organizers@googlegroups.com

Student Projects (for discussion)

- Our efforts to lower the barrier of entry
- Patapsco -- including documentation (e.g., notebooks)
- **How to handle student submissions?**
  - **Need a faculty supervisor/advisor?**